Lectures on Financial Economics

by Antonio Mele

Swiss Finance Institute, University of Lugano

Centre for Economic Policy Research

swiss:finance:institute

December 2017

© 2003-2017, Antonio Mele







Front cover explanations

Top: Illustration of the increased efficiency in maritime routing allowed by the Suez Canal (right panel) opened in 1869, and the Panama Canal (left panel) opened in 1913, two amongst the most enduring technological marvels with global economic and political implications.

Bottom: A 75 year 3% coupon bearing bond issued by the Panama Canal Company ("Compagnie Universelle du Canal Interocéanique de Panama") in October 1884. The company defaulted in 1889 under the leadership of the Count Ferdinand de Lesseps, who during 1858 had also founded the Suez Canal Company ("Compagnie Universelle du Canal Maritime de Suez").

Information about the author

Antonio Mele holds a Senior Chair at the Swiss Finance Institute and is a Professor of Finance at the University of Lugano, after a decade spent as a tenured faculty at the London School of Economics. He is also a Research Fellow for the Financial Economics program at the Centre for Economic Policy Research (CEPR) in London. He holds a PhD in Economics from the University of Paris. His work spans a variety of fields in financial economics, pertaining to financial market volatility, interest rates and credit markets, macro-finance, financial markets and business cycles, information in securities markets in contexts with information networks and Knightian uncertainty, and has appeared in journals such as the *Journal of Financial Economics*, the *Review of Economic Studies,* the *Review of Financial Studies*, and the *Journal of Monetary Economics*. He authored or co-authored three books on themes regarding financial market volatility.

His work outside academia includes developing fixed income volatility indexes and related tradable instruments for Chicago Board Options Exchange and S&P Dow Jones Indices. He is currently a member of the *Securities and Markets Stakeholder Group* of the European Securities Markets Authority (ESMA), the supra-national supervisor of European financial markets, where he previously served as a member of its Group of Economic Advisers.

"Antonio Mele does not accept any liability for any losses related to the use of the models, data, and methods described or developed in these lectures."

Contents

Introduction **14**

I Foundations 28

1 The classic capital asset pricing model **29**
1.1 Introduction 29
1.2 Portfolio selection 31
1.2.1 Wealth constraints 31
1.2.2 Portfolio choice: the "Capital Market Line" 31
1.2.3 Without the safe asset: the "Efficient Portfolio Frontier" 32
1.2.4 Risk-return trade-offs in the two asset case 33
1.2.5 The global minimum variance portfolio 34
1.2.6 The market portfolio 35
1.3 The CAPM 37
1.3.1 Restrictions on securities expected returns 38
1.3.2 Restrictions on alternative investment returns 39
1.3.3 Zero-beta CAPM 40
1.3.4 An excursion into risk-premiums and certainty equivalents 41
1.3.5 Back to CAPM: Equilibrium with expected utility 44
1.4 The APT 48
1.4.1 Exact APT 48
1.4.2 Risk-neutral tilts, or the fundamental theorem of asset pricing 49
1.4.3 Uncertainty and asset evaluation 52
1.4.4 The APT with idiosyncratic risk and a large number of assets 53
1.4.5 Systematic risk 54
1.5 Empirical evidence 54

1.5.1 Fama & Mac Beth . . . 54
1.5.2 Macroeconomic forces . . . 55
1.5.3 Fama & French . . . 55
1.5.4 The high-beta stocks anomaly . . . 56
1.5.5 "Lucky factors" . . . 56
1.6 Market practice . . . 56
1.6.1 Risk-parity . . . 56
1.6.2 Black-Litterman . . . 56
1.6.3 Smart beta . . . 56
1.6.4 Factor investing . . . 56
1.7 Appendix 1: Details regarding portfolio choice . . . 57
1.7.1 The primal program . . . 57
1.7.2 The dual program . . . 58
1.7.3 Efficient portfolios generate efficient portfolios . . . 59
1.7.4 Covariance of global minimum variance and efficient portfolio returns . . . 59
1.8 Appendix 2: The market portfolio and the security market line . . . 60
1.8.1 The tangent portfolio is the market portfolio . . . 60
1.8.2 Tangency conditions . . . 60
1.8.3 An alternative derivation of the SML . . . 61
1.9 Appendix 3: Risk and risk aversion . . . 62
1.9.1 Modeling and approximating risk-premiums . . . 62
1.9.2 Stochastic dominance and mean-preserving spreads . . . 64
1.10 Appendix 4: Money demand and liquidity traps . . . 67
References . . . 69

2 Arbitrage, equilibrium and pricing 71
2.1 Introduction . . . 71
2.2 The static general equilibrium in a nutshell . . . 73
2.2.1 Walras' Law . . . 74
2.2.2 Competitive equilibrium . . . 74
2.2.3 Optimality . . . 75
2.3 The role of financial securities in markets with uncertainty . . . 79
2.3.1 Commodity markets . . . 79
2.3.2 Financial securities and rational expectations . . . 80
2.3.3 Securities prices as shadow values . . . 81
2.3.4 Gambles and securities . . . 81
2.3.5 Arrow-Debreu securities . . . 84
2.4 Arbitrage and replication: Introduction . . . 86
2.4.1 Rain and sunshine . . . 87
2.4.2 Replication and pricing: the role of complete markets . . . 89
2.5 No-arbitrage: theory . . . 90
2.5.1 Lands of Cockaigne . . . 90
2.5.2 Enforced asset prices . . . 91

2.6 Equivalent martingales, and equilibrium . . . 93
2.6.1 Rational expectations . . . 94
2.6.2 Stochastic discount factors . . . 95
2.6.3 Equilibrium . . . 96
2.7 Consumption-CAPM . . . 102
2.7.1 Risk-neutral pricing and macroeconomic risks . . . 102
2.7.2 The beta relation . . . 104
2.7.3 CCAPM & CAPM . . . 104
2.8 Infinite horizon . . . 105
2.9 Further topics on incomplete markets . . . 105
2.9.1 Nominal assets and real indeterminacy of the equilibrium . . . 105
2.9.2 Nonneutrality of money . . . 106
2.10 Appendix 1 . . . 108
2.11 Appendix 2: Proofs of selected results . . . 109
2.12 Appendix 3: The multicommodity case . . . 112
References . . . 114

3 Infinite horizon economies 115
3.1 Introduction . . . 115
3.2 Consumption-based asset evaluation . . . 115
3.2.1 Recursive plans: introduction . . . 115
3.2.2 Asset pricing: the marginalist argument . . . 117
3.2.3 Intertemporal elasticity of substitution . . . 117
3.2.4 Lucas' model . . . 118
3.3 Production: foundational issues . . . 122
3.3.1 Decentralized economy . . . 122
3.3.2 The social planner solution . . . 123
3.3.3 Dynamics . . . 124
3.3.4 Stochastic economies . . . 126
3.4 Production-based asset pricing . . . 131
3.4.1 Firms . . . 131
3.4.2 Consumers . . . 134
3.4.3 Equilibrium . . . 135
3.5 Money, production and asset prices in overlapping generations models . . . 135
3.5.1 Introduction: endowment economies . . . 135
3.5.2 Diamond's model . . . 138
3.5.3 Money . . . 138
3.5.4 Money in a model with real shocks . . . 143
3.5.5 Sunspots . . . 143
3.6 Optimality . . . 143
3.6.1 Production economies . . . 143
3.6.2 Over-accumulation of capital . . . 144
3.6.3 Bubbles . . . 145

3.6.4 Money . . . 145
3.7 Appendix 1: Finite difference equations, with economic applications . . . 146
3.8 Appendix 2: Neoclassic growth in continuous-time . . . 150
3.8.1 Convergence from discrete-time . . . 150
3.8.2 The model . . . 151
3.9 Appendix 3: Notes on optimization of continuous time systems . . . 153
References . . . 156

4 Information, security design and financial contracting 157
4.1 Introduction . . . 157
4.2 Theoretical challenges to frictionless markets: information problems . . . 157
4.2.1 The economics of information . . . 158
4.2.2 Adverse selection . . . 160
4.2.3 Moral hazard . . . 161
4.2.4 Signaling . . . 168
4.3 The classics: capital structure and Modigliani-Miller propositions . . . 168
4.3.1 The irrelevance of debt . . . 168
4.3.2 Irrelevance through Arrow-Debreu security pricing . . . 169
4.3.3 The cost of capital and financial frictions . . . 169
4.4 Symmetric information again: full insurance . . . 170
4.5 Debt and moral hazard . . . 171
4.6 Debt and adverse selection with costly state verification . . . 173
4.7 Capital structure and incomplete contracting . . . 173
4.8 Appendix 1: Details regarding debt and moral hazard . . . 174
References . . . 175

5 Continuous time models 177
5.1 Introduction . . . 177
5.2 An introduction to no-arbitrage and equilibrium . . . 178
5.2.1 Time . . . 178
5.2.2 The origins: Black & Scholes . . . 179
5.2.3 Asset prices as Feynman-Kac representations . . . 183
5.2.4 The Girsanov theorem . . . 185
5.2.5 The APT in continuous time . . . 188
5.2.6 Example: no-arbitrage in Lucas tree . . . 191
5.3 Distorsions and numéraires . . . 197
5.3.1 Leading example: consumption-based probabilities . . . 197
5.3.2 Numéraire pricing . . . 198
5.4 Martingales and arbitrage . . . 202
5.4.1 The information framework . . . 202
5.4.2 Viability . . . 203
5.4.3 Market completeness . . . 204
5.5 Equilibrium with a representative agent . . . 206

5.5.1 Constant investment opportunity sets . . . 206
5.5.2 Stochastic opportunity sets . . . 206
5.5.3 Arrow-Debreu densities and restrictions on expected returns . . . 211
5.5.4 Interest rates . . . 214
5.6 Martingale methods . . . 216
5.6.1 Intertemporal budget constraints . . . 216
5.6.2 The optimization problem . . . 218
5.6.3 Marginal utility of income . . . 219
5.6.4 Example: log-utility . . . 220
5.6.5 Equilibrium . . . 220
5.6.6 Continuous time Consumption-CAPM . . . 221
5.6.7 Partial hedging in incomplete markets: introduction . . . 221
5.7 Inaction: the economics of American options . . . 221
5.7.1 Early exercise premiums: an introductory example . . . 221
5.7.2 Gambles and securities again . . . 223
5.7.3 Real options theory . . . 224
5.7.4 Perpetual puts . . . 225
5.7.5 Perpetual calls . . . 226
5.8 Further topics on real options and controlled Brownian motions . . . 229
5.8.1 Irreversible investments and the decision to invest . . . 229
5.8.2 A model of determination of exchange rates in target zones . . . 230
5.8.3 Liquidity constraints and optimal dividend policy . . . 231
5.9 Portfolio constraints . . . 235
5.9.1 Technical background . . . 235
5.9.2 Artificial markets . . . 238
5.10 Jumps . . . 239
5.10.1 Poisson jumps . . . 239
5.10.2 Interpretation . . . 240
5.10.3 Properties and related distributions . . . 241
5.10.4 Asset pricing implications . . . 242
5.10.5 An option pricing formula . . . 243
5.11 Continuous time Markov chains . . . 243
5.12 Appendix 1: An introduction to stochastic calculus for finance . . . 244
5.12.1 Stochastic integrals . . . 244
5.12.2 Stochastic differential equations . . . 254
5.13 Appendix 2: Self-financed strategies, from discrete to continuous time . . . 259
5.13.1 With intermediate consumption . . . 259
5.13.2 With final consumption only . . . 259
5.14 Appendix 3: Proof of selected results . . . 262
5.14.1 Proof of Theorem 5.3 . . . 262
5.14.2 Details regarding multibeta CAPM . . . 262
5.14.3 Proof of Eq. (5.109). . . . 263

5.14.4 Walras consistency tests . . . 264
5.15 Appendix 4: The Green's function . . . 265
5.15.1 Setup . . . 265
5.15.2 The PDE connection . . . 266
5.16 Appendix 5: Portfolio constraints . . . 267
5.17 Appendix 6: Topics on jumps . . . 269
5.17.1 The Radon-Nikodym derivative . . . 269
5.17.2 Arbitrage restrictions . . . 270
5.17.3 State price density: introduction . . . 270
5.17.4 State price density: general case . . . 271
References . . . 273

6 Taking models to data **275**
6.1 Introduction . . . 275
6.2 Data generating processes . . . 275
6.2.1 Basics . . . 275
6.2.2 Restrictions on the DGP . . . 276
6.2.3 Parameter estimators . . . 277
6.2.4 Basic properties of density functions . . . 277
6.2.5 The Cramer-Rao lower bound . . . 278
6.3 Maximum likelihood estimation . . . 278
6.3.1 Basics . . . 278
6.3.2 Factorizations . . . 278
6.3.3 Asymptotic properties . . . 279
6.4 M-estimators . . . 281
6.5 Pseudo, or quasi, maximum likelihood . . . 282
6.6 GMM . . . 283
6.7 Simulation-based estimators . . . 286
6.7.1 Three simulation-based estimators . . . 287
6.7.2 Asymptotic normality . . . 289
6.7.3 A fourth simulation-based estimator: Simulated maximum likelihood . . . 292
6.7.4 Advances . . . 293
6.7.5 In practice? Latent factors and identification . . . 293
6.8 Asset pricing, prediction functions, and statistical inference . . . 294
6.9 Appendix 1: Proof of selected results . . . 299
6.10 Appendix 2: Collected notions and results . . . 300
6.11 Appendix 3: Theory for maximum likelihood estimation . . . 303
6.12 Appendix 4: Dependent processes . . . 304
6.12.1 Weak dependence . . . 304
6.12.2 The central limit theorem for martingale differences . . . 304
6.12.3 Applications to maximum likelihood . . . 304
6.13 Appendix 5: Proof of Theorem 6.4 . . . 306
References . . . 307

II Empirical lessons and market inefficiencies 310

7 Neo-classical kernels and puzzles 311
7.1 Introduction . . . 311
7.2 The equity premium puzzle . . . 312
7.2.1 A single-factor model . . . 312
7.2.2 Extensions . . . 315
7.2.3 Equity premium and interest rate puzzles . . . 315
7.3 Hansen-Jagannathan cup . . . 317
7.4 Multifactor extensions . . . 319
7.4.1 Exponential affine pricing kernels . . . 320
7.4.2 Lognormal returns . . . 321
7.5 Conditional CAPM . . . 323
7.6 Pricing kernels and Sharpe ratios . . . 323
7.6.1 Market portfolios and pricing kernels . . . 323
7.6.2 Pricing kernel bounds . . . 325
7.7 Conditioning bounds . . . 327
7.8 Survivorship bias? . . . 327
7.9 Appendix . . . 328
References . . . 331

8 Aggregate fluctuations in equity markets 332
8.1 Introduction . . . 332
8.2 Empirical evidence: bird's eye view . . . 333
8.2.1 Equity markets and the business cycle . . . 333
8.2.2 Predictability . . . 338
8.2.3 Risk-return trade-offs . . . 340
8.3 Volatility: a business cycle perspective . . . 341
8.3.1 Volatility cycles . . . 341
8.3.2 Understanding the empirical evidence . . . 343
8.3.3 What to do with stock market volatility? . . . 348
8.3.4 What did we learn? . . . 352
8.4 Rational market fluctuations . . . 353
8.4.1 The dynamics of asset returns . . . 353
8.4.2 Stocks as options . . . 355
8.5 Time-varying discount rates or uncertain growth? . . . 358
8.5.1 Two types of economies . . . 358
8.5.2 Markov pricing kernels, asset returns and volatility . . . 359
8.5.3 External habit formation . . . 361
8.5.4 Large price swings as a learning induced phenomenon . . . 365
8.5.5 Linearity-generating processes . . . 372
8.6 Modeling market-to-book ratios . . . 375
8.7 Appendix 1: Estimation of the market expected return . . . 376

8.8 Appendix 2: Calibration of the tree in Section 8.3 . . . 377
8.9 Appendix 3: Asset prices in a multifactor model . . . 379
8.10 Appendix 4: Arrow-Debreu PDEs . . . 380
8.11 Appendix 5: Increasing risk and convexity . . . 381
8.11.1 The maximum principle . . . 381
8.11.2 Increasing risk . . . 383
8.11.3 Proof of Theorem 8.1 . . . 383
8.11.4 A digression on a "macro-asset" derivative . . . 384
8.12 Appendix 6: Dynamics of habit in Campbell and Cochrane (1999) . . . 386
8.13 Appendix 7: An algorithm to simulate discrete-time pricing models . . . 388
8.14 Appendix 8: Heuristic details of learning in continuous time . . . 389
8.15 Appendix 9: Linear regime-switching economies . . . 390
8.16 Appendix 10: Bond price convexity revisited . . . 391
References . . . 392

9 Macrofinance **396**
9.1 Introduction . . . 396
9.2 Non-expected utility . . . 398
9.2.1 Recursive formulations . . . 398
9.2.2 Testable restrictions . . . 399
9.2.3 Risk premiums and interest rates . . . 400
9.2.4 Campbell-Shiller approximation . . . 402
9.2.5 Risks for the long-run . . . 402
9.3 Heterogeneous agents and "catching up with the Joneses" . . . 404
9.4 Idiosyncratic risk . . . 406
9.4.1 A static model . . . 406
9.4.2 Self-insurance and persistence of idiosyncratic shocks . . . 408
9.4.3 A model with countercyclical income inequality . . . 408
9.5 Incomplete markets with homogeneous and heterogenous agents . . . 410
9.5.1 Idiosyncratic shocks unrelated to aggregate risk . . . 410
9.5.2 A two-agents economy . . . 411
9.6 Disagreement and learning . . . 414
9.6.1 Learning with multiple signals . . . 414
9.6.2 Overconfidence and bubbles . . . 415
9.6.3 General equilibrium without frictions . . . 419
9.7 Coping with Knigthian uncertainty . . . 427
9.7.1 Prelude . . . 427
9.7.2 Uncertainty aversion and Ellsberg paradox . . . 429
9.7.3 Portfolio selection and market participation . . . 431
9.7.4 A model of multiple likelihoods . . . 434
9.8 Production . . . 439
9.9 Government spending and asset prices . . . 441
9.10 Leverage and volatility . . . 441

9.10.1 Primitives . . . 441
9.10.2 Equity volatility: a decomposition formula . . . 442
9.10.3 Bankruptcy . . . 443
9.11 Multiple trees and the cross-section of asset returns . . . 443
9.12 The term-structure of interest rates . . . 443
9.13 Prices, quantities and the separation hypothesis . . . 445
9.13.1 A closed-form expression for non-expected utility . . . 445
9.13.2 Preferences for robustness . . . 446
9.13.3 Irrelevance . . . 446
9.14 Endogenous risk and the financial accelerator doctrine . . . 447
9.14.1 Procyclicality . . . 447
9.14.2 Premise: Asymmetric information, asset prices and business cycles . . . 448
9.14.3 Credit cycles . . . 448
9.14.4 Amplification . . . 453
9.14.5 Amplification with a banking sector . . . 453
9.14.6 Additional literature . . . 454
9.15 Appendix 1: Non-expected utility . . . 455
9.15.1 Detailed derivation of optimality conditions and selected relations . . . 455
9.15.2 Details regarding models of long-run risks . . . 458
9.15.3 Continuous time . . . 458
9.16 Appendix 2: Economies with heterogenous agents . . . 459
9.17 Appendix 3: Knightian uncertainty . . . 464
9.18 Appendix 4: Credit rationing . . . 466
References . . . 469

10 Information and other market frictions **475**
10.1 Introduction . . . 475
10.2 Prelude: imperfect information in macroeconomics . . . 477
10.3 Informational efficiency: roadmap . . . 480
10.4 Walrasian equilibria as informationally inefficient outcomes . . . 481
10.5 Rational Expectations Equilibrium . . . 483
10.6 Noisy Rational Expectations Equilibrium . . . 484
10.6.1 Asymmetric information: information transmission . . . 486
10.6.2 Differential information: information aggregation . . . 490
10.7 Dealers markets: Introduction . . . 493
10.7.1 Markets with symmetric information . . . 494
10.7.2 With asymmetric information: bid-ask spreads . . . 495
10.7.3 Inventory risk and bid-ask spreads . . . 498
10.8 Markets with strategic players . . . 499
10.8.1 The Kyle baseline model . . . 500
10.8.2 Markets with multiple traders and dealers . . . 501
10.8.3 Dynamic markets . . . 507
10.8.4 Mandatory disclosure . . . 510

10.9 Limits to arbitrage and further frictions . . . 513
10.9.1 Liquidity trading . . . 513
10.9.2 Price impacts and derivatives . . . 513
10.9.3 Arbitrage imperfections . . . 513
10.10 Over-the-counter markets . . . 514
10.10.1 Background . . . 514
10.10.2 Search . . . 515
10.10.3 A model with symmetric information . . . 515
10.11 Coordination failures in financial markets . . . 519
10.11.1 Banking crises . . . 519
10.11.2 Higher order beliefs and beauty contests . . . 519
10.12 Appendix 1: The projection theorem . . . 520
10.13 Appendix 2: Details regarding solutions of selected models . . . 521
10.14 Appendix 4: Some foundations to pricing behavior in macroeconomics . . . 526
References . . . 529

III Asset pricing and reality 532

11 Options and volatility 533
11.1 Introduction . . . 533
11.2 Forwards and futures . . . 534
11.2.1 Forwards: definition and pricing in frictionless markets . . . 534
11.2.2 Forwards as a means to borrow money . . . 535
11.2.3 Marking to market . . . 535
11.2.4 Futures . . . 535
11.2.5 Backwardation and Contango . . . 536
11.3 Optionality and no-arb bounds . . . 539
11.3.1 Model-free properties . . . 539
11.3.2 Hedging . . . 543
11.3.3 A case study: accumulators, decumulators . . . 544
11.4 Evaluation . . . 545
11.4.1 A pricing formula . . . 545
11.4.2 Black & Scholes . . . 546
11.4.3 Surprising cancellations and "preference-free" formulae . . . 548
11.4.4 Future options and Black's formula . . . 548
11.4.5 Hedging . . . 549
11.4.6 Endogenous volatility . . . 550
11.4.7 Properties of options in diffusive models . . . 551
11.5 Stochastic volatility . . . 554
11.5.1 Statistical models of changing volatility . . . 554
11.5.2 Implied volatility, smiles and skews . . . 556
11.5.3 Option pricing with stochastic volatility . . . 561

11.6 Trading volatility with options . . . 569
11.6.1 Payoffs . . . 569
11.6.2 P&Ls of Δ-hedged strategies . . . 573
11.7 Local volatility . . . 576
11.7.1 Issues . . . 576
11.7.2 Implied binomial trees . . . 576
11.7.3 The perfect fit, in continuous time . . . 579
11.7.4 Relations with implied volatility . . . 581
11.8 The price of (equity) volatility . . . 583
11.8.1 One introductory example: range-based volatility . . . 583
11.8.2 "Fear gauge" contracts . . . 584
11.8.3 Forward volatility trading . . . 589
11.8.4 Marking to market . . . 590
11.8.5 Stochastic interest rates . . . 590
11.8.6 Hedging . . . 591
11.9 A digression on skewness . . . 592
11.10Dealing with market imperfections . . . 593
11.11Appendix 1: The original arguments of Black & Scholes . . . 594
11.12Appendix 2: Black (1976) . . . 595
11.13Appendix 3: Stochastic volatility . . . 596
11.13.1 Hull & White equation . . . 596
11.13.2 Extensions . . . 596
11.13.3 Smile analytics . . . 597
11.14Appendix 4: Local volatility . . . 599
11.15Appendix 5: Variance contracts . . . 601
11.16Appendix 6: Skewness contracts . . . 604
References . . . 605

12 Engineering of fixed income securities 608
12.1 Introduction . . . 608
12.1.1 Relative pricing in fixed income markets . . . 609
12.1.2 Many evaluation paradigms . . . 609
12.1.3 Plan of the chapter . . . 610
12.2 Markets and interest rate conventions . . . 610
12.2.1 Markets for interest rates . . . 610
12.2.2 Mathematical definitions of interest rates . . . 613
12.2.3 Yields to maturity on coupon bearing bonds . . . 615
12.2.4 Accruals, invoice, and clean prices on coupon bearing bonds . . . 615
12.3 Duration and convexity hedging and trading . . . 617
12.3.1 Duration . . . 618
12.3.2 Convexity . . . 620
12.3.3 Asset-liability management . . . 620
12.4 Foundational issues in interest rate modeling . . . 628

12.4.1 Tree representation of the short-term rate . . . 629
12.4.2 Tree pricing . . . 632
12.4.3 Introduction to calibration . . . 633
12.4.4 Calibrating probabilities throught derivative data . . . 648
12.4.5 Extensions to trinomial trees . . . 656
12.5 The Ho and Lee model . . . 656
12.5.1 The tree . . . 657
12.5.2 Price movements and the martingale restriction . . . 658
12.5.3 The recombining condition and interest rate volatility . . . 658
12.5.4 Model's solution . . . 660
12.5.5 Calibration of the model . . . 662
12.5.6 An example . . . 662
12.5.7 Continuous-time approximations, with an application to barbell trading . 666
12.6 Beyond Ho and Lee: Calibration through Arrow-Debreu securities . . . 670
12.6.1 Extracting Arrow-Debreu securities from the yield curve . . . 671
12.6.2 Two model examples . . . 674
12.7 Callables, puttable and convertibles with trees . . . 683
12.7.1 Definitions and rationale . . . 683
12.7.2 Callable bonds . . . 686
12.7.3 Convertible bonds . . . 690
12.8 Appendix 1: Botstrapping and no-arbitrage restrictions . . . 694
12.9 Appendix 2: Proof of Eq. (12.17) . . . 698
12.10Appendix 2: The Ho and Lee price representation . . . 700
References . . . 702

13 Interest rates **703**
13.1 Introduction . . . 703
13.2 Bond prices and interest rates . . . 704
13.2.1 A first representation of bond prices . . . 704
13.2.2 Forward rates . . . 706
13.2.3 A second representation of bond prices . . . 706
13.3 Stylized facts . . . 707
13.3.1 The expectation hypothesis . . . 707
13.3.2 Bond returns predictability . . . 708
13.3.3 The yield curve and the business cycle . . . 710
13.3.4 Additional stylized facts about the US yield curve . . . 712
13.3.5 Common factors affecting the yield curve . . . 713
13.4 Models of the short-term rate: Introduction . . . 716
13.4.1 Models versus representations . . . 717
13.4.2 The bond pricing equation . . . 717
13.4.3 Stochastic duration . . . 721
13.4.4 Some famous models . . . 722
13.4.5 The Monetary Experiment and interest rate volatility . . . 728

13.4.6 Jumps . . . 731
13.5 Multifactor models of the short-term rate . . . 732
13.5.1 Stochastic volatility . . . 732
13.5.2 Three-factor models . . . 737
13.5.3 Affine and quadratic term-structure models . . . 738
13.5.4 Unspanned stochastic volatility . . . 739
13.5.5 Topics regarding estimation and trading strategies . . . 740
13.6 No-arbitrage models: early formulations . . . 742
13.6.1 Fitting the yield-curve, perfectly . . . 743
13.6.2 Ho & Lee . . . 744
13.6.3 Hull & White . . . 745
13.7 The Heath-Jarrow-Morton framework . . . 746
13.7.1 Framework . . . 746
13.7.2 The model . . . 747
13.7.3 The dynamics of the short-term rate . . . 748
13.7.4 Embedding . . . 748
13.7.5 Stochastic string shocks models . . . 749
13.8 Interest rate derivatives . . . 752
13.8.1 Fixed income market volatility and the persistence of the short-term rate 752
13.8.2 Hypothetical continuous payoffs . . . 755
13.8.3 Forward martingale probabilities . . . 756
13.8.4 European options on bonds . . . 757
13.8.5 Callable and puttable bonds . . . 760
13.8.6 Options on fixed coupon bonds . . . 762
13.8.7 Interest rate swaps . . . 763
13.8.8 Caps & floors . . . 766
13.8.9 Swaptions . . . 767
13.9 Market models . . . 768
13.9.1 Models and market practice . . . 768
13.9.2 Simply-compounded forward rate dynamics, and no-arb restrictions . . . 769
13.9.3 Applications to derivative evaluation . . . 770
13.10 Volatility surfaces . . . 774
13.10.1 Implied volatilities . . . 774
13.10.2 Local volatilities and SABR models . . . 775
13.11 Appendix 1: The FTAP for bond prices . . . 779
13.12 Appendix 2: Certainty equivalent interpretation of forward prices . . . 781
13.13 Appendix 3: Additional results on forward probabilities . . . 782
13.14 Appendix 4: Principal components analysis . . . 783
13.15 Appendix 5: Comparison results applying to defaultable bonds . . . 784
13.16 Appendix 6: Details regarding exponential-affine models . . . 785
13.16.1 Derivation of the closed-form solution . . . 785
13.16.2 Application to Fong and Vasicek model . . . 785

13.17Appendix 7: A few analytical details regarding the Hull and White model 786
13.18Appendix 8: Expectation theory and embedding in selected models 787
13.19Appendix 9: Additional results on string models 789
13.20Appendix 9: Changes of numéraire and Jamshidian's (1989) formula 790
References 791

14 Risky debt and credit derivatives **796**
14.1 Introduction 796
14.1.1 A brief history of credit risk and financial innovation 796
14.1.2 Plan of the chapter 799
14.2 Conceptual approaches to valuation of defaultable securities 799
14.2.1 Firm value, or structural, approaches 799
14.2.2 The structural approach in practice: the pricing of convertible bonds . . 814
14.2.3 Reduced form approaches: rare events, or intensity, models 817
14.2.4 Ratings 821
14.3 Credit derivatives and structured products based thereon 825
14.3.1 Options and spreads 825
14.3.2 Credit Default Swaps 826
14.3.3 Collateralized Debt Obligations (CDOs) 843
14.4 Managing loan losses 856
14.4.1 Regulatory framework 856
14.4.2 Foundations of risk-management 859
14.4.3 Credit risk and VaR 864
14.5 Procyclicality, credit crunches and quantitative easing 866
14.5.1 The 2007 subprime crisis 867
14.5.2 Top tier capital ratio targets and endogenous volatility 870
14.5.3 Credit crunches and quantitative easing 876
14.5.4 Where did QE go? 879
14.6 Appendix 1: Details on strategic defaulting 882
14.7 Appendix 2: Proof of selected results 883
14.8 Appendix 3: Transition probability matrices and pricing 884
14.9 Appendix 4: Bond spreads in markets with stochastic default intensity 886
14.10Appendix 6: Conditional probabilities of survival 888
14.11Appendix 7: Details regarding CDS index swaps and swaptions 889
14.12Appendix 8: Modeling correlation with copulae functions 891
14.13Appendix 9: Details on CDO pricing with imperfect correlation 893
References 894

Introduction

A. A brief description of the book

Lectures on Financial Economics is a graduate textbook that originates from a set of initial notes I wrote in support of graduate and advanced undergraduate lectures in financial economics, macroeconomic dynamics, financial econometrics and financial engineering. Unifying these notes is tantamount to engage into a long and patient journey into historical intellectual developments as well as the interactions of ideas and theories with actual markets behaviors. The book attempts at a "synthesis" of the state of knowledge accumulated during 65 years of initially intermittent but, later, incessant contributions to this very important field of economics.[1]

Finance has the potential to oil the wheels of the real economy. While economists still debate about the benefits of finance for our society, more than a dozen of scholars researching into this field may be counted as Nobel Memorial Prize laureates in Economic Sciences. Progress was sometimes faster than any attempts at organizing our thoughts. Initially, efforts at synthesis focussed on the mathematical structures underlying the pioneering work leading to the foundations of finance. Later, synthesis became more problematic as research work proliferated through such disparate domains including, among others, the evaluation of derivatives instruments, the behavior of markets over the business cycle, information problems in corporate finance and asset markets and, last but not least, the then nascent econometrics of financial markets. The initial "classics" would often cover non-overlapping spaces, as reviewed in Section C of this Introduction. We are still struggling with the creation of a comprehensive treatment of financial economics. This book is an attempt at such a treatment, an attempt at linking various theories and ideas to empirical puzzles and, sometimes, established market practice.

Financial economics relies on sophisticated methods that have already received a comprehensive textbook treatment, since at least the early classics. While this book still aims at providing foundations and methodology, it is intended as a narrative of the historic milestones in the progress of thought. Empirical puzzles have motivated the emergence of new explanations of

[1] In fact, nearly one century of contributions, once we account for those portions of this book dealing with markets plagued by Knightian uncertainty (Keynes, 1921; Knight, 1921) and those market failures identified by Keynes (1936). In Section B of this Introduction, I review the early contribution of Bachelier (1900).

financial market behaviors and, then, new foundations; likewise, new theories have prompted for additional testable predictions. This virtuous interaction has led to immense knowledge that I try to account for while trying to connect various areas in a single piece. I would have liked to write a "history of financial economics," but, as noted, I only attempt at synthesizing as much as I can in this book. Section B of this Introduction provides motivation and historical perspective regarding the progress of knowledge that occurred during the last many decades, and a broad outline of the book.

Writing an account on the state of knowledge in financial economics is a significant challenge. I may count more than 200 models underlying the explanations in this book. Does model uncertainty disqualify financial economics from being a science? I am writing this Introduction with a humble but decisively optimistic view, even though underlying this view is the acknowledgment that we are dealing with such a large and sometimes fragmented field, and still far from being unified. But while we cannot rely on controlled experiments as in other fields, our models lead to predictions that are typically testable through the availability of myriads of data. I hope that this book will make the reader comfortable with the idea that financial economics is progressing on a well-defined path, that the two-hundred models I discuss belong to a common paradigm and, finally, that each of these models is very important, by shedding light into specific angles of the varied and complex structure of financial markets. This project has the potential to produce a durable impact once these learning objectives are met.

B. Overview and coverage

This book aims to track the milestones achieved in the history of thought in financial economics. Its objective is to provide a comprehensive reference while attempting at organizing almost one century of work, while relying on a rigorous analytical framework and, finally, while providing methodological tools that make it self-contained. At the same time, it endeavors to help explain real phenomena and how these phenomena and, sometimes, market practice, have helped economists reformulate previous theories. Furthermore, the book includes many examples and solved problems that illustrate the main lessons conveyed by the models analyzed in the book. I don't provide supplementary material such as solutions, answers or other material to accompany the book. The book tries to be self-contained.

While our field is very large, the present work tries to cover as much as I can, while maintaining a balance between theoretical explanations and empirical evidence and identifying the practical relevance of our knowledge. The outcome is still patchy—again, a reflection of the nature of our field. However, I hope that I am managing to provide the reader with a coherent treatment of many disparate aspects of financial markets, those arising in idealized explanations (be they based on abstract or empirical methodologies), those that are most relevant to the market practice and, finally, those that may be of interest to scholars working in related fields.

The book is organized in three parts: (I) Foundations, (II) Empirical Lessons and Market Inefficiencies, and (III) Asset Pricing and Reality.

"Part I: Foundations" develops primordial tools for analysis while striving to keep track of historical backgrounds. For example, Chapter 1 deals with portfolio selection problems arising in the early 1950s and the initial theories of asset prices of the 1960s, works that are understood to be at the origins of financial economics. The next chapters in this Part provide refinements

of the initial theories, based on subsequent breakthroughs made to understand the role of asset prices in the general equilibrium of dynamic economies subject to uncertainty, in both finite (Chapter 2) and infinite horizon (Chapter 3). Historically, the apex of these developments was reached with the advent of 'continuous time finance' and its methods occurring in the 1970s. Continuous time models would elegantly address difficult problems including no-arbitrage pricing of redundant securities (derivatives, in some cases), or portfolio choices through dynamic programming. The 'martingale methods' of the 1980s-1990s would seal this toolbox with additional instruments, but they also paved the way to the analysis of incomplete markets and other market imperfections. It is the 'golden age' of financial theory, a famous expression proposed by Darrell Duffie in his classic *Dynamic Asset Pricing Theory* (2001, p. xiii).

This Part covers details of this progress but its perspective regards the economic significance of the problems we are dealing with. The goal of these lectures is to understand capital market fluctuations and, sometimes, the behavior of firms subject to financial constraints, or the role that these fluctuations and behaviors play in conveying information and resources back to society. Information asymmetries and market imperfections are actually a recurrent theme in these lectures. Thus, Chapters 1 through 3 deal with the idealized markets leading to the initial explanations of asset prices; for example, Chapter 3 deals with financial markets while attempting at taking a broad perspective, one in which finance is part of a general ecosystem—however, our discussions in this chapter (often, not always) rely on models without frictions. In these markets, asset prices relate to consumption, production, money, and the links arise through the behavior of fully rational individuals.

However, Chapter 4 points to serious conceptual difficulties in defining a general equilibrium, difficulties arising due to asymmetric information, and regarding the very existence of markets or the process of securities creation. Information problems are indeed powerful sources of inspiration for economists. More in detail, Chapter 4 does rely on information and provides the reader with an overview of theories of financial contracting and theories of debt. Why do corporations issue debt or equity, i.e., financial assets that may subsequently trade on secondary markets? Or, suppose that a firm receives funds to undertake a project; shouldn't then the same firm lose some of its initial motivation in undertaking the necessary care while handling the project? How to incentivize this firm to exert the entrepreneurial efforts that would make its interests aligned to those of the investors? What is the optimal capital structure in an economy in which we may not be able to foresee all future contingencies? Next, Chapter 5 shows the beauties of continuous time finance applied to classical problems such as no-arbitrage pricing or portfolio selection (relying on dynamic programming), but also the theory of irreversible investments, the firms' dividends distribution policy in the presence of liquidity constraints, or theories of financial contracting in continuous time.

Some theories are important because of their main qualitative conclusions, and it may not be needed to test for their 'functional form'. But financial economics is also a field that lends itself so naturally to vast empirical investigations, where models may predict different outcomes according to parameter values. We need to assess the statistical relevance of certain theories and, even more fundamentally, we need to estimate a model to be used by decision makers, be they policy makers or private investors. Chapter 6 deals with theory and methods of statistical inference needed to deal with models arising in financial economics, relying on classical econometric tools such as maximum likelihood, methods of moments, and the relatively more modern simulation-based inference methods.

"Part II: Empirical lessons and market inefficiencies" is about explaining the main empirical facts and the challenges that these facts pose to financial economists. The first puzzles regard excess price volatility, that is, the difficulty of the early dynamic models to explain the aggregate market behavior. According to the early models, market volatility (and the premium required by investors to invest in a volatile environment) is one order of magnitude less than that we observe in the markets. Chapter 7 is an introduction to these critical problems and in particular to their measurement methods. These methods were developed mainly during the 1980s-1990s and, in part, the 2000s, and are obviously statistical in nature. At the same time, they rely on the principle of no-arbitrage: there are many ways we can price assets while only requiring absence of arbitrage; however, there is a benchmark amongst these ways, solely relying on security market data, which can be used to assess whether a model under scrutiny implies implausible parameter values (such as, say, the investors' risk appetite).

To address the volatility puzzles, financial economists added explanations of capital market behavior based on a variety of assumptions: investors' attitude to risk-taking (e.g., non-expected utility, or habit formation), idiosyncratic risk, incomplete markets or restricted market participation, heterogeneous beliefs, learning in contexts with incomplete information, a fully specified production sector, or Knightian uncertainty. This progress relied on foundational work made during the 'golden age' of financial theory described in Part I. It occurred in the 2000s-2010s and is described in Chapters 8 and 9. Interestingly, these new models addressed relatively older issues; for example, they predict that, under conditions, the aggregate equity premium and stock market volatility are both countercyclical, a fact known from earlier empirical research. But while these models were developed, financial economists also realized that they could explain additional 'cross-sectional anomalies' such as the value premium (the tendency for firms with low multiples to perform better in the future than those with high), or the hoary issue of predictability (the tendency of the market to reverse its trends after a while, maybe in tandem with the business cycle).

A common trait of these models is their adherence to the 'separation hypothesis', that is, the assumption that the real economy is not affected by capital market developments. Perhaps due to the dramatic counterfactual evidence brought about by the Great Recession of the late 2000s, a new research agenda relaxed this assumption, aiming to revitalize previous work made by macroeconomists on 'financial accelerator' mechanisms—the power of capital markets to exacerbate the business cycles. Integrating financial markets into the real sphere of the economy is the explicit intellectual acknowledgment of the crucial role that capital markets (frictions) play in society. Chapter 9 contains many links to this literature, and these links form the basis for additional discussions in various junctures of Part III.

Chapter 10 concludes this Part while pointing to other famous puzzles and frictions. Do capital markets provide useful information for society? How deep 'price discovery' is, that is, how well asset prices reflect the fundamentals in a world with heterogenous information? This chapter begins with the classical analysis of markets plagued with information problems. Investors obviously have different pieces of information, and some of them are even known to possess superior information. A 'lemons problem' arises: what are the incentives to trade with better informed investors? One answer is that trading may only arise when markets are somehow (informationally) inefficient: when information is not available symmetrically in the marketplace, we can only trade once we know the reason someone is also trading with us does not arise because our counterparty is better informed than we are. For example, the price we observe may not allow us to reverse-engineer the information of better informed traders;

this inefficiency (a price that is only partially revealing) may actually be what makes markets function.

For longtime, this price inefficiency has been modeled as arising, endogenously, from the presence of exogenous liquidity shocks. In fact, liquidity and information problems have long been understood as the two sides of the same coin. But liquidity is not only information, and this chapter describes alternative explanations for it. Remarkably, these information and related problems were tackled while, at the same time, economists were in the process of developing market microstructure theory, i.e., the theory of price formation in trading venues relatively more realistic than hypothesized during the golden age. However, capital markets and trading venues can be even much more complex than the literature had initially hypothesized: Chapter 10 explains that the presence of irrational traders, herding behavior and information cascades, information networks, information percolations, agents engaged in coordination games, or decentralized trading systems (e.g., over-the-counter markets), were all exciting topics of research from the 1990s through the 2010s, which still promise to improve our understanding of such a complex phenomenon as the price formation process. Chapter 10 concludes with a survey of topics regarding coordination failures in financial markets. Why do financial crises arise? What are the determinants of a bank-run? What makes agents coordinate to equilibrium outcomes where prices can deviate from fundamentals, as in the famous beauty contests introduced by John Maynard Keynes in his *General Theory of Employment, Interest and Money* (1936, chapter 12)? Remember, in these contests, the winners are those who pick up the most popular faces from many photographs, and are thus incentivized to forecast the forecasts of others, where everyone is doing the same thing. How do higher order beliefs affect price dynamics or the emergence of the risk of a bank run?

"Part III: Asset pricing and reality" aims just to this: to rely on the lessons drawn from Part II and cope (through the main analytical tools in Part I) with the main challenges posed by actual capital markets, arising from option pricing and trading, interest rate modeling, or credit risk and the associated derivatives. In a sense, Part II is about the big puzzles we face in fundamental research, while Part III is about how to live within our current and certainly unsatisfactory paradigms, so as to cope with demand for intellectual expertise.

The importance of these topics can never be emphasized enough. Investments or business cycles are clouded with uncertainty. While investing, decision makers put their jobs and the security of their families at risk, thereby affecting human capital accumulation and, hence, the life of future generations. Sometimes, the effects of poorly informed choices can be devastating. The infamous 2007-subprime crisis and the subsequent Great Recession certainly illustrate these mechanisms. In general, capital market volatility is huge, for one reason or another, illustrated by the theories and facts in Part II. It is, thus, a naturally human response to try and find a solution to cope with these risks.

Derivative securities are instruments to insure against risks related to certain investment decisions. As is well known, they are called 'derivatives' because their value is drawn from that of other securities. For example, if we are long a number of shares, we may wish to purchase put options on these shares (or on a dedicated index of them), which pay off when the average shares value is down; intuitively, then, the price of these options decreases with the price of the underlying securities. Louis Bachelier's *Théorie de la spéculation* in 1900 is the first attempt at tackling these evaluation problems—problems that were tackled again during the golden age, based on no-arbitrage principles.

At the heart of this principle lie different assumptions, and one of them dictates that the underlying securities (or, in general, risks) should be well understood by all market participants. For example, the shares underlying the previous options should be traded in reasonably liquid markets, a condition for price discovery. It is most likely the case with many derivative securities such as the equity index options or the U.S. Treasury derivatives that are traded in well functioning Exchanges, but also with a variety of derivatives traded in over-the-counter markets (e.g., interest rate swaps or credit related products). However, if risks are poorly understood and price discovery is scarce, derivatives may be mispriced. A case of 'toxicity' may then arise: investors may inadvertently add too high doses of complex derivatives in their portfolios than justified by their risk-return trade-off profile. Unfortunately, financial history shows many cases of toxicity. The last chapter of Part III examines some details of one of them, related to the process of securitization of very risky mortgages.

Engineering can be defined as a set of processes and methods that attempt to use established scientific knowledge to solve practical problems, as with the case of steam engines utilized during the first Industrial Revolution. In fact, if it wasn't for the previous mishaps, it would have been very tempting to title this Part *"Financial engineering."* Instead, "Asset pricing and reality" reminds us that while our engines do certainly rely on scientific and rigorous knowledge, this knowledge seems to be more limited than in the domains of the physical sciences. It does not mean that financial economists are not in the process of building up financial engineering. Nor does it mean that financial innovation is unnecessary or toxic. We shall learn throughout the whole lectures that financial innovation may allow for risk-sharing (the transfer of some risks from those who are not willing to bear them to those who are) when the pre-existing markets are not diverse enough (i.e., incomplete). In fact, and interestingly, financial economists not only are inspired by the events they see (as with the previous revival of the financial accelerator hypothesis motivated by the Great Recession); sometimes, they lead to institutional changes: option markets would most probably not exist today without the golden age revolution of the 1970s.

Yet our most successful inventions regard risks that can at least be identified. Some of these inventions attract liquidity, which, in turn, generates price discovery and, then, liquidity again, over a virtuous circle. Liquidity begets liquidity: a product is more likely to trade if a trader knows he could easily trade it when, in the future, he will decide to get out from his current trade. Financial products are a little bit like a fax machine was at the time of its introduction: they are worth because others are willing to use them. Potential market makers and financial economists alike (see Chapter 10) are well aware of this chicken-and-egg problem: coordination may fail even when risks are well identified. The exposition in this Part is affected by a sort of survivorship bias: it regards products, trading methods and processes that have been successful. The exception is the description of credit related products at the epicenter of the 2007-2009 crisis.

Chapter 11 illustrates well the scope of Part III: while Part II describes theories and facts regarding asset market volatility, this chapter analyzes ways to trade it in the equity space. The technicalities can be actually complex: we have realized volatility, stochastic volatility, implied volatility and volatility surfaces, implied binomial trees and local volatility. Some models can be useful to buy-side institutions; others to sell-side firms engaged into pure intermediation activities. Furthermore, some unexpected developments occurred in financial theory during the 1990s-2000s, which gave rise to financial innovations regarding the way volatility is traded. It is another episode of financial history when theory had preceded market practice.

Chapter 12 and Chapter 13 repeat these analyses in the much more complex field that is fixed income. Fixed income securities are complex due to theoretical reasons (they track the time value of money), technical reasons (they have multiple dimensions, such as expiration or tenors of the various contracts) and, last but not least, because price discovery in these markets may be somehow hindered by their trading mechanisms (over-the-counter markets). Yet fixed income securities allow pension funds and other asset managers to mitigate interest rate risk, which can be much, much higher than that in equity markets. Interest rates have also very interesting business cycle properties, which policy-makers rely on while trying to predict the business cycle: for this reason, Chapter 13 also contains some links to fascinating topics arising in macroeconomics. Finally, Chapter 14 deals with the evaluation of debt subject to default risk and derivatives based thereon. It is a chapter focussed on practical aspects, with the exception of a few junctures devoted to the analysis of strategic default or the origins of the 2007-subprime crisis.

Engineering can be civil or electronic engineering, amongst many others. Likewise, a would-be "financial engineering" field should not be only about derivatives. For example, it should also deal with such issues as portfolio optimization in contexts with short-sale constraints, time-varying volatility (ARCH models, for example), Bayesian learning, and also with variance shrinkage methods. In general, this field should be a camaleon, just as it happens in the physical sciences: it should take the colors of the specific set of problems that is helping to address, in order to facilitate financial transactions, information processes (including, for instance, the design of volatility indexes with data stemming out from over-the-counter markets) and methods both in the buy-side and sell-side worlds. Some of these topics are covered throughout the lectures although their systematic treatment goes well beyond the scope of this work.

C. Discussion of related work

Financial economics has evolved while crossing a variety of boundaries. How did we track this progress? Ingersoll (1987), Huang and Litzenberger (1988) and Duffie (2001) are the first classics organizing more than thirty years of conceptual analysis, while O'Hara (1995) is the first classic organizing the theories of the 1980s on liquidity and market microstructure.

These works led to a sophisticated and consistent framework at the basis of subsequent progress, progress that occurred mostly in response to empirical challenges faced by the initial analyses; for example, during the mid 1990s and the early 2000s, new models were proposed to explain how aggregate market behavior links to the business cycle. Cochrane (2005), Back (2010) and Campbell (2017) provided further momentum to standardization of knowledge, teaching how part of this progress relates to the initial analyses. The early work of Campbell, Lo and MacKinlay (1997) offered a quite exhaustive overview of many statistical instruments that are still of paramount importance in the empirical modeling or statistical testing of financial markets phenomena.

Foucault, Pagano and Röell (2013) summarized further progress related to studies of market liquidity and microstructure. During the 2000s, Amaro de Matos (2001) and Tirole (2006) provided us with the first attempts at organizing knowledge acquired in the theory of corporate finance, while during the 1990s, Freixas and Rochet (1997) had already written a classic in the theory of banking. In the references section of this Introduction, I list additional references on works that attempt at organizing knowledge in financial economics.

I like to represent this book as being complementary to these very important works. The added value of mine is to provide general perspective into a large variety of topics, as well as details of the historical progress leading to our current understanding of each of these topics.

To illustrate, the Handbooks of the Economics of Finance (Constantinides, Harris and Stultz, 2003, 2013) currently undertake the ambitious task of dealing with many disparate topics arising in financial economics. However, these works are contributed by several authors and they are only partially coordinated. This book provides explicit linkages across chapters, which may help a reader interested in learning or reviewing several topics while accessing to a common language. For example, the book may be used as a reference in several courses in advanced training programs: adopting this book for a single course (e.g., a course in macro-finance) should allow the reader to access material for related courses (one in financial markets with frictions, say) while relying on a style that he or she is already familiar with. But the most remarkable feature of the efforts I have tried to accomplish in this book is the ambition to cover a wide range of topics, the milestones in the history of thought in this field, while maintaining a balance across theory, empirical evidence, historical contexts and, also, market practice. As explained, in the references section of this Introduction, I would like to bring a few additional textbook treatments of the field to the reader's attention, noting that these represent my own preferred readings, and apologizing with all omitted authors.

D. Audience and pre-requisites

The main audience for this book will be academics studying, teaching and researching in financial economics. The book also aims to appeal to applied researchers and other professionals servicing investment banks, institutional investors, central banks and governments. The inclusion of policy makers as part of the audience for this book is motivated by the widespread acknowledgment of the many interconnections between financial markets and the macroeconomy. Macro-financial linkages arising through the business cycle (or, say, market liquidity, microstructure and volatility) are themes that have motivated important work by financial economists; this work is very useful reading for those engaged in designing supervisory standards and macroprudential policies. The main audience for this book will be financial economists, though. I'll return to additional factors explaining potential audience in Section E below.

The style of the book is eminently academic. It is primarily quantitative, even while, on many occasions, the book provides descriptions of markets and historical contexts. But appreciating this book in its entirety relies on a predisposition to quantitative reasoning. At the same time, quantitative reasoning is the means, not the goal of this work. Thus, in general, appreciating this book also requires to be genuinely passionate about economics. Finally, the book is suitable for both theorists and empiricists or fellows searching for concrete applications. Applied researchers will have access to a clear theoretical background needed prior to undertaking meaningful empirical research. Theorists will learn the nature of the empirical regularities and puzzles that have characterized our field since its very beginnings, knowledge that is indispensable prior to undertaking meaningful theoretical research.

Reading the book requires knowledge of economics and mathematics at a level required from a candidate to a Master of Science in Finance and Economics at the London School of Economics or to a PhD in Finance at the Swiss Finance Institute, two programs where I delivered many of the lectures inspiring this work. I also believe that the book may be accessed by a well-motivated student enrolled into a program such as, say, a Bachelor of Science in Economics

and Statistics at University College London. Therefore, reading this book requires maturity in both microeconomics and macroeconomics at the level of Varian (1992) and Blanchard (2017) textbooks, respectively. Moreover, some readers might already have gained motivation for finance while exposed to introductory finance textbooks such as Berk and DeMarzo (2016) and Bodie, Kane and Marcus (2014). Additional pre-requisites include knowledge of calculus, basic knowledge of time series and statistics, and an introductory exposure to stochastic calculus. Many technicalities are introduced and explained in appropriate junctures of the book, along with references to more advanced material.

E. Usage of the book

The reader of this book is a scholar in financial economics, a market practictioner, a policymaker, or a scholar in related fields such as macroeconomics.

A scholar in financial economics may (i) recommend this book to a specialized readership for a survey of work linked to his research articles; (ii) recommend portions of the book to advanced graduate or PhD classes as complements to his lecture notes; and, finally, (iii) value a book that attempts at a synthesis: for example, reading this book may help a young scholar develop a critical view of our current understanding of financial markets, thereby stimulating further and hopefully important research.

A market practictioner or a policy maker with an appropriate background (see Section D) may find a source of valuable information in this book. For example, certain parts of the book (Part III as well as some chapters in Part II) may provide a quantitative strategist or a risk manager with guidance on elaborating, estimating and implementing models for signal generation. Even more important, the book may help gain intellectual perspective into the many details that arise in market or policy practice.

Finally, the average reader may be a scholar in other fields. For example, a macroeconomist might be interested in financial economics from a broad and still rigorous perspective; this book may help shed light into his own research and, perhaps, lead him to recommend portions of it to his PhD classes. For example, some financial economists grew up by learning from some of the classics described in Section C of this Introduction, while at the same time reading the beautiful *Lectures on Macroeconomics* of Blanchard and Fischer (1989), which helped shape some research into macro-finance. Similarly, one objective of this book is to attract the attention of scholars in other fields. Oftentimes did scholars in other fields make "excursions" into financial economics and provide marvelous contributions, e.g., on the role of information in securities markets, idiosyncratic risk or financial accelerator mechanisms. I hope that the general perspective I endeavor to follow in this book may attract scholars from other fields and help render such excursions more frequent.

The book contains material that may be accessed to while learning about a number of topics, and/or be used as a reference for a number of courses, such as:

- Foundations of portfolio selection (Chapter 1)
- Foundations of financial economics (selected portions of Chapters 2, 3, 4, 5 and 10)
- Introduction to quantitative methods in finance (selected portions of Chapters 3; Chapter 5)
- Statistical methods of financial model validation (Chapters 6 and 7)

- Financial markets and the macroeconomy (Chapters 8 and 9)
- Information and financial markets (Chapters 4 and 10)
- Financial markets, debt and frictions (Chapters 4 and 10)
- Option pricing and volatility trading (Chapter 11)
- Fixed income markets and derivatives (Chapters 12 and 13)
- Credit markets and derivatives (Chapter 14)
- Derivatives and financial engineering (selected portions of Chapters 11 through 14)

The chapters indicated in parenthesis contain the relevant material for the suggested courses, which I have indeed experimented during the last 20 years or so.

References

[1] Amaro de Matos, J. (2001): *Theoretical Foundations of Corporate Finance*. Princeton: Princeton University Press.

[2] Bachelier, L. (1900): "Théorie de la spéculation." *Annales Scientifiques de l'École Normale Supérieure* 3 (17), 21-86.

[3] Back, K. (2010): *Asset Pricing and Portfolio Choice*. Oxford: Oxford University Press.

[4] Berk, J. and P. DeMarzo (2016): *Corporate Finance*. Boston: Pearson Education.

[5] Bodie, Z., A. Kane and A.J. Marcus (2014): *Investments*. Berkshire, U.K.: McGraw-Hill Education.

[6] Blanchard, O. (2017): *Macroeconomics*. Boston: Pearson Education.

[7] Blanchard, O. and S. Fisher (1989): *Lectures on Macroeconomics*. Cambridge: MIT Press.

[8] Campbell, J.Y., A. W. Lo and C. MacKinlay (1997): *The Econometrics of Financial Markets*. Princeton: Princeton University Press.

[9] Campbell, J. (2017): *Financial Decisions and Markets*: A Course in Asset Pricing. Princeton: Princeton University Press (forthcoming).

[10] Cochrane, J. (2005): *Asset Pricing*. Princeton: Princeton University Press.

[11] Constantinides, G.M., M. Harris and R. Stultz (2003): *Handbooks of the Economics of Finance* (Volume 1). Amsterdam: Elsevier.

[12] Constantinides, G.M., M. Harris and R. Stultz (2013): *Handbooks of the Economics of Finance* (Volume 2). Amsterdam: Elsevier.

[13] Duffie, D. (2001): *Dynamic Asset Pricing Theory*. Princeton: Princeton University Press.

[14] Foucault T., M. Pagano and A. Röell (2013): *Market Liquidity: Theory, Evidence and Policy*. Oxford: Oxford University Press.

[15] Freixas, X. and J-Ch. Rochet (1997): *Microeconomics of Banking*. Cambridge: MIT Press.

[16] Huang, C-f. and R.H. Litzenberger (1988): *Foundations for Financial Economics*. New York: North-Holland.

[17] Ingersoll, J.E. (1987): *Theory of Financial Decision Making*. New York: Rowman & Littlefield Publishers.

[18] Keynes, J.M. (1921): *A Treatise on Probability*. London: MacMillan and Co.

[19] Keynes, J.M. (1936): *The General Theory of Employment, Interest and Money*. London: Palgrave Macmillan.

[20] Knight, F.H. (1921): *Risk, Uncertainty, and Profit*. New York: Houghton Mifflin.

[21] O'Hara, M. (1995): *Market Microstructure Theory*. Cambridge: Blackwell Publishers.

[22] Tirole, J. (2006): *The Theory of Corporate Finance*. Princeton: Princeton University Press.

[23] Varian, H. (1992): *Microeconomic Analysis*. New York: W.W. Norton & Company.

Further readings

Generic references

[24] LeRoy, S.F. and J. Werner (2014): *Principles of Financial Economics*. Cambridge: Cambridge University Press.

[25] Lengwiler, Y. (2014): *Microfoundations of Financial Economics*. Princeton: Princeton University Press.

[26] Munk, C. (2013): *Financial Asset Pricing Theory*. Oxford: Oxford University Press.

[27] Pennacchi, G. (2008): *Theory of Asset Pricing*. Boston: Pearson Addison Wesley.

Financial markets and the macroeconomy

[28] Altug, S. and P. Labadie (2008): *Asset Pricing for Dynamic Economies*. Cambridge: Cambridge University Press.

[29] Hansen, L.P. and T.J. Sargent (2008): *Robustness*. Princeton: Princeton University Press.

Information and other market frictions

[30] Allen, F. and D. Gale (2007): *Understanding Financial Crises*. Oxford: Oxford University Press.

[31] Bianconi, M. (2003): *Financial Economics, Risk and Information*. Singapore: World Scientific Publishing.

[32] Brunnermeier, M.K. (2001): *Asset Pricing under Asymmetric Information*. Oxford: Oxford University Press.

[33] Duffie, D. (2012): *Dark Markets*. Princeton: Princeton University Press.

[34] Schleifer, A. (2000): *Inefficient Markets*. Oxford: Oxford University Press.

[35] Shin, H.S. (2010): *Risk and Liquidity*. Oxford: Oxford University Press.

[36] Vives, X. (2008): *Information and Learning in Markets*. Princeton: Princeton University Press.

Derivatives

[37] Brigo, D. and F. Mercurio (2006): *Interest Rate Models*: Theory and Practice. New York: Springer Verlag.

[38] Derman, E. and M.B. Miller (2016): *The Volatility Smile*. Hoboken, NJ: John Wiley & Sons.

[39] Duffie, D. and K.J. Singleton (2003): *Credit Risk*. Princeton: Princeton University Press.

[40] Lando, D. (2004): *Credit Risk Modeling*. Princeton: Princeton University Press.

[41] McDonald, R.L. (2006): *Derivatives Markets*. Boston: Pearson Addison Wesley.

[42] Veronesi, P. (2010): *Fixed Income Securities*. Hoboken, NJ: John Wiley & Sons.

Disclaimer

This manuscript is still incomplete. Economic motivation and intuition are not always developed as they would deserve, some derivations are inelegant and, sometimes, the presentation is still a bit informal. Moreover, I still have to include (additional) material on monetary models of asset prices, theories of the nominal and the real term structure of interest rates, bubbles, asset prices implications of overlapping generations models, limits to arbitrage, theories of debt and capital structure and agency problems in continuous time, or financial frictions and their interconnections with business cycle developments. Finally, I still need to include more extensive surveys for each topic I cover, especially in Chapters 1, 3, 4, 6 and 7. Of the 14 Chapters I have drafted, Chapter 7 (and the present Preface!) are those in need of the most serious revamp. Moreover, I only started drafting Chapter 4. I am working towards revising these notes and filling these gaps. Meanwhile, any comments on this version of my work are very welcome.

Antonio Mele
December 2017

Part I

Foundations

1
The classic capital asset pricing model

1.1 Introduction

An investor is considering creating a portfolio of securities. How should he choose the proportions of each asset based on his attitude towards risk? What are the asset pricing implications of all the investors' portfolio choices? Are these choices consistent with absence of arbitrage opportunities?

This chapter deals with these issues in the context of the static market idealized in the first contributions that would give birth to financial economics. The next section explains how portfolio selection works when our investor optimizes a mean-variance criterion, as in the seminal approach of Markovitz (1952). According to this criterion, an investor seeks to maximize the portfolio expected return for any given variance constraint. The ensuing asset allocation leads, in equilibrium (supply equals demand for all assets), to a notion of "market portfolio" as well as a first theory of asset prices, the celebrated "Capital Asset Pricing Model" of Sharpe (1964), Lintner (1965) and Mossin (1966), or CAPM for short (see Section 1.3).

The market portfolio is the first instance of a "preference-free" construct in financial economics: the asset composition of this portfolio is independent of the investors' risk aversion. Remarkably, any desired profile in the investors' risk-return trade-off is obtained as a combination of a purely safe asset and the market portfolio (hence, the name). We shall explain that there are nuances to be made regarding these preference-free attributes, but the notion of a market portfolio is important, representing as it does a benchmark for each individual asset evaluation: the striking conclusion of the CAPM is that each asset expected return is proportional to the expected return on the market portfolio. The volatility (i.e., riskiness) of each asset return then increases with the asset exposure to market movements.

These explanations are elegant albeit quite coarse descriptions of asset markets. Section 1.4 develops the "Arbitrage Pricing Theory" model of Ross (1976), or APT. The APT provides refinements of the CAPM, predicting that each asset expected return relates to a number of factors, not only the market, as the original CAPM suggests. For example, there might be macroeconomic forces explaining stock returns, because investors require compensation (i.e., a risk premium) for investing in risky assets that perform poorly when the economy performs equally poorly, as is usually the norm. The APT formalizes these concepts very elegantly.

How successful were these initial explanations of asset prices fluctuations? It is well known that the single-factor CAPM cannot explain the cross section of stock returns. It is one of the first asset pricing puzzles that financial economists had to face while developing the field. We shall explain that the empirical evidence suggests that at least three factors explain the cross-section of stock returns. In fact, the literature has produced a huge amount of evidence regarding many additional factors. We shall succinctly discuss this evidence and the extent to which some of them are mere "lucky factors."

The important point of this empirical evidence is that it provides motivation and guidance to the formulation of new models. These models can be seen as dynamic extensions of the APT. For example, a recurrent theme in financial economics is how "cyclical forces" affect asset evaluation and, in turn, how capital market developments feed back the business cycle. The original APT framework was the initial reference for thinking about these themes, as studied in detail in more advanced parts of these lectures (see, e.g., Chapters 8 and 9 in Part II). More generally, it was a milestone for all subsequent work aimed to address various "asset pricing puzzles."

The empirical literature abounds with additional instances of asset pricing "anomalies," i.e., difficulties the CAPM (and related static models) encounters while explaining market behavior. Some of these anomalies are now understood to link to frictions that are behind the very price formation process. Information problems, agency issues, limited rationality, capital immobility, or decentralized markets, are all examples of hindrances to market efficiency. In fact, the notion of "market efficiency" is elusive and financial economists actually agree that some (rigourously defined) inefficiency is needed to have markets function in the first place—an "equilibrium degree of disequilibrium," in the words of Grossman and Stiglitz (1980) contained in a contribution on capital markets equilibrium with information asymmetries (see Chapter 10). For example, bid-ask spreads are needed to incentivize financial institutions to trade securities in their roles of intermediaries. These topics are dealt with in many junctures of these lectures (see, e.g., Chapters 4 and 10), and it is important to emphasize they will since this very first chapter.

But the CAPM has many undisputed potentially useful features. A striking feature of the portfolio allocation rules derived in this chapter is their analytical simplicity. We have already explained that the CAPM is the reference to further developments in the history of thought. But how successful have been the asset allocation rules of this chapter in the market practice? Mention (i) risk-parity, (ii) Black-Litterman, (iii) smart beta, (iv) factor investing.

[In progress]

1.2 Portfolio selection

This section describes the process of allocation of wealth for a representative investor who cares about "mean-variance efficiency." This notion of efficiency constraints our investor to choices of portfolios that produce the highest expected return for a given level of risk.

1.2.1 Wealth constraints

Our investor can invest into m risky assets and one safe asset. Let $S = [S_1, \cdots, S_m]$ be the risky assets price vector, and let S_0 be the price of the riskless asset. We wish to determine the value of a portfolio of all these assets. Let $\theta = [\theta_1, \cdots, \theta_m]$, where θ_i is the number of the i-th risky asset, and let θ_0 be the number of the riskless asset, in this portfolio. The initial wealth is, $w = S_0\theta_0 + S \cdot \theta$. Terminal wealth is $w' = x_0\theta_0 + x \cdot \theta$, where x_0 is the payoff promised by the riskless asset, and $x = [x_1, \cdots, x_m]$ is the vector of the payoffs pertaining to the risky assets, i.e., x_i is the payoff promised by the i-th risky asset.

Let $R \equiv \frac{x_0}{S_0}$ and $\tilde{R}_i \equiv \frac{x_i}{S_i}$ denote the gross returns from investing into the safe asset and the i-th risky asset. Accordingly, define $r \equiv R - 1$ as the safe interest rate, $\tilde{b} \equiv [\tilde{b}_1, \cdots, \tilde{b}_m]^\top$, where $\tilde{b}_i \equiv \tilde{R}_i - 1$ is the return on the i-th risky asset, and $b \equiv E(\tilde{b})$, the vector of the expected returns on the risky assets. Finally, we let $\pi = [\pi_1, \cdots, \pi_m]^\top$, where $\pi_i \equiv \theta_i S_i$ is the wealth invested in the i-th asset. We have

$$w' = x_0\theta_0 + \sum_{i=1}^{m} x_i\theta_i \equiv R\pi_0 + \sum_{i=1}^{m} \tilde{R}_i\pi_i \quad \text{and} \quad w = \pi_0 + \sum_{i=1}^{m} \pi_i. \tag{1.1}$$

Combining the two expressions for w' and w leaves:

$$w' = \pi^\top(\tilde{R} - \mathbf{1}_m R) + Rw = \pi^\top(b - \mathbf{1}_m r) + Rw + \pi^\top(\tilde{b} - b).$$

We use the decomposition, $\tilde{b} - b = B \cdot \tilde{u}$, where B is a $m \times d$ "volatility" matrix, with $m \leq d$, and $\tilde{u}$ is a random vector with expectation zero and variance-covariance matrix equal to the identity matrix. With this decomposition, we can rewrite the budget constraint in Eq. (1.1) as follows

$$w' = \pi^\top(b - \mathbf{1}_m r) + Rw + \pi^\top B\tilde{u}. \tag{1.2}$$

We can now use Eq. (1.2) and determine the expected return and the variance of the portfolio value. We have

$$E\left[w'(\pi)\right] = \pi^\top\left(b - \mathbf{1}_m r\right) + Rw \quad \text{and} \quad var\left[w'(\pi)\right] = \pi^\top\Sigma\pi, \tag{1.3}$$

where $\Sigma \equiv BB^\top$. Let $\sigma_i^2 \equiv \Sigma_{ii}$. We assume that Σ has full-rank, and that for all i, j, $\sigma_i^2 > \sigma_j^2 \Longrightarrow b_i > b_j$. The last condition does actually hold in the equilibrium of markets in which investors are averse to risk, a topic dealt with several times in this chapter; furthermore, note, more trivially, that this condition implies that $r < \min_j(b_j)$.

1.2.2 Portfolio choice: the "Capital Market Line"

We assume that the investor maximizes the expected return on his portfolio conditionally on a given level of uncertainty. That is, let $w^2 \cdot v_p^2$ define the maximum level of dollar variance

the investor is willing to accept. We consider the following program based on Eq. (1.3) and the uncertainty constraint:

$$\hat{\pi}\left(v_p\right) = \arg\max_{\pi\in\mathbb{R}^m} E\left[w'(\pi)\right] \quad \text{s.t. } var\left[w'(\pi)\right] = w^2 \cdot v_p^2. \tag{1.P1}$$

The first order conditions for [1.P1] are,

$$\hat{\pi}\left(v_p\right) = \left(2\nu\right)^{-1}\Sigma^{-1}\left(b - \mathbf{1}_m r\right) \quad \text{and} \quad \hat{\pi}^\top\Sigma\hat{\pi} = w^2 \cdot v_p^2,$$

where ν is a Lagrange multiplier for the variance constraint. Plugging the first condition into the second, we obtain, $\left(2\nu\right)^{-1} = \mp\frac{w\cdot v_p}{\sqrt{\text{Sh}}}$, where

$$\text{Sh} \equiv \left(b - \mathbf{1}_m r\right)^\top \Sigma^{-1}\left(b - \mathbf{1}_m r\right), \tag{1.4}$$

is the *Sharpe market performance*. To ensure efficiency, we take the positive solution. Substituting the positive solution for $\left(2\nu\right)^{-1}$ into the first order condition, we obtain that the portfolio that solves [1.P1] is

$$\frac{\hat{\pi}\left(v_p\right)}{w} \equiv \frac{\Sigma^{-1}\left(b - \mathbf{1}_m r\right)}{\sqrt{\text{Sh}}} \cdot v_p. \tag{1.5}$$

We are ready to determine the value of [1.P1], $E\left[w'(\hat{\pi}\left(v_p\right))\right]$ and, hence, the expected portfolio return, defined as,

$$\mu_p(v_p) \equiv \frac{E\left[w'(\hat{\pi}(v_p))\right] - w}{w} = r + \sqrt{\text{Sh}} \cdot v_p, \tag{1.6}$$

where the last equality follows by a simple calculation. Eq. (1.6) describes what is known as the *Capital Market Line* (CML).

1.2.3 Without the safe asset: the "Efficient Portfolio Frontier"

Next, assume that the investor's choice space does not include the riskless asset. His current wealth is now $w = \sum_{i=1}^m \pi_i$, while his terminal wealth is $w' = \sum_{i=1}^m \tilde{R}_i\pi_i$, such that, and relying on the definition of $\tilde{b}_i$,

$$w' = \sum_{i=1}^m \tilde{b}_i\pi_i + \sum_{i=1}^m \pi_i = \pi^\top b + w + \pi^\top B\tilde{u}, \tag{1.7}$$

where B and $\tilde{u}$ are as defined as in Eq. (1.2). We can use Eq. (1.7) to determine the expected return and the variance of the portfolio value, which are

$$E\left[w'(\pi)\right] = \pi^\top b + w, \text{ where } w = \pi^\top\mathbf{1}_m \quad \text{and} \quad var\left[w'(\pi)\right] = \pi^\top\Sigma\pi. \tag{1.8}$$

The program our investor solves, now, is

$$\hat{\pi}\left(v_p\right) = \arg\max_{\pi\in\mathbb{R}} E\left[w'(\pi)\right] \quad \text{s.t. } var\left[w'(\pi)\right] = w^2 \cdot v_p^2 \text{ and } w = \pi^\top\mathbf{1}_m. \tag{1.P2}$$

In the Appendix, we show that provided $AC - D^2 > 0$ (a second order condition, as explained in the Appendix), the solution to [1.P2] is

$$\frac{\hat{\pi}\left(v_p\right)}{w} = \frac{C\mu_p(v_p) - D}{AC - D^2}\Sigma^{-1}b + \frac{A - D\mu_p(v_p)}{AC - D^2}\Sigma^{-1}\mathbf{1}_m, \tag{1.9}$$

where $A \equiv b^\top \Sigma^{-1} b$, $D \equiv \mathbf{1}_m^\top \Sigma^{-1} b$ and $C \equiv \mathbf{1}_m^\top \Sigma^{-1} \mathbf{1}_m$, and $\mu_p(v_p)$ is the expected portfolio return, defined as in Eq. (1.6). In the Appendix, we also show that

$$v_p^2 = \frac{1}{C}\left[1 + \frac{1}{AC - D^2}\left(\mu_p(v_p)C - D\right)^2\right]. \tag{1.10}$$

Based on Eq. (1.10), we define the *global minimum variance portfolio* as the portfolio that achieves a variance equal to $v_p^2 = C^{-1}$ and an expected return equal to $\mu_p = D/C$. We shall return to this portfolio below (see Section 1.2.4).

Note that for each v_p, there are two values of $\mu_p(v_p)$ that solve Eq. (1.10). The optimal choice for our investor is that with the highest μ_p. We define the *efficient portfolio frontier* as the set of values (v_p, μ_p) that solve Eq. (1.10) with the highest μ_p. It has the following expression

$$\mu_p(v_p) = \frac{D}{C} + \frac{1}{C}\sqrt{\left(Cv_p^2 - 1\right)(AC - D^2)}. \tag{1.11}$$

The efficient portfolio frontier is increasing and concave in risk, v_p. It can be interpreted as a "production function," i.e., a technology through which "expected returns" are obtained using varying "levels of risk" as inputs (see, e.g., the two-asset case depicted in Figure 1.1). Which portfolio on this frontier is selected by an investor depends on the investor's attitudes towards risk. We shall return to this selection problem in Section 1.3).

1.2.4 Risk-return trade-offs in the two asset case

Consider a simple example in which there are only two risky assets, $m = 2$. This simplified version of the model illustrates some crucial points regarding diversification very clearly, and is also at the basis of famous contributions going beyond financial economics, notably in macroeconomics (see Section 1.3.5.2). In this two-asset case, the efficient portfolio frontier is the risk-return trade-off resulting from any feasible combination of the two assets, and does not require any optimization, as above: the budget constraint, $\frac{\pi_1}{w} + \frac{\pi_2}{w} = 1$, pins down an unique relation between the portfolio expected return and variance. Precisely, we have: $\mu_p = \frac{E[w'(\pi)] - w}{w} = \frac{\pi_1}{w} b_1 + \frac{\pi_2}{w} b_2$, or

$$\begin{cases} \mu_p = b_1 + (b_2 - b_1)\dfrac{\pi_2}{w} \\ v_p^2 = \left(1 - \dfrac{\pi_2}{w}\right)^2 \sigma_1^2 + 2\left(1 - \dfrac{\pi_2}{w}\right)\dfrac{\pi_2}{w}\sigma_{12} + \left(\dfrac{\pi_2}{w}\right)^2 \sigma_2^2 \end{cases}$$

whence

$$v_p = \frac{1}{b_2 - b_1}\sqrt{\left(b_2 - \mu_p\right)^2 \sigma_1^2 + 2\left(b_2 - \mu_p\right)\left(\mu_p - b_1\right)\rho\sigma_1\sigma_2 + \left(\mu_p - b_1\right)^2 \sigma_2^2}.$$

Note that when the two asset returns are perfectly correlated, $\rho = 1$, we have that

$$\mu_p = b_1 + \frac{b_2 - b_1}{\sigma_2 - \sigma_1}\left(v_p - \sigma_1\right).$$

That is, the risk-return trade-off is linear. However, diversification pays as soon as the asset returns are not perfectly positively correlated: for any given expected return, the portfolio return volatility decreases as the asset correlation decreases. Figure 1.1 actually reveals that there are portfolios that are even less risky than the less risky asset, and risk can actually be zeroed in the extreme case where $\rho = -1$.

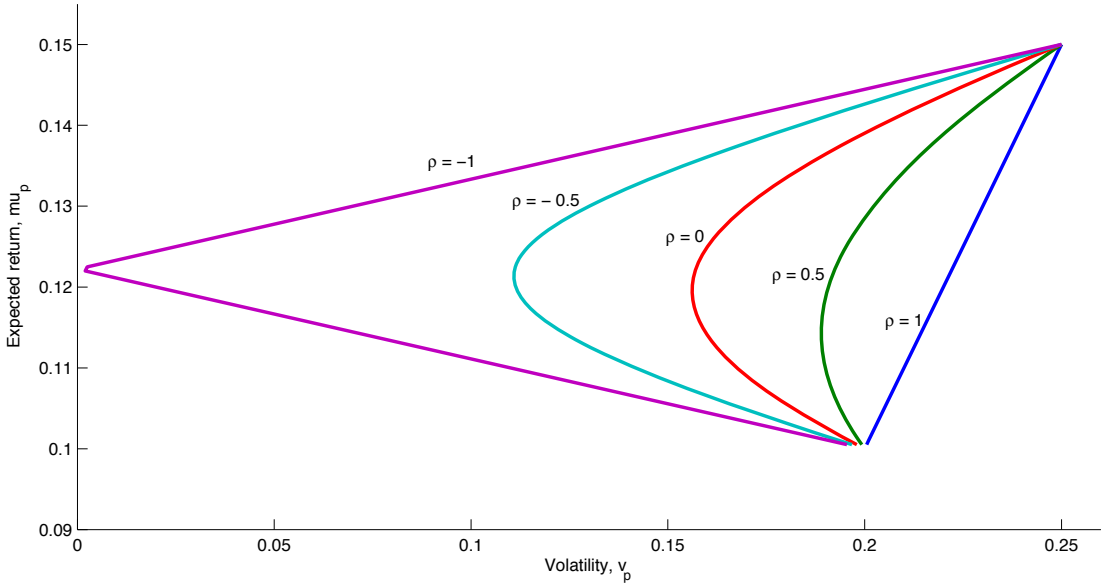


FIGURE 1.1. From top to bottom: portfolio frontiers corresponding to $\rho = -1$, -0.5, 0, 0.5, 1. Remaining parameter values are $b_1 = 0.10$, $b_2 = 0.15$, $\sigma_1 = 0.20$, $\sigma_2 = 0.25$. For each portfolio frontier, the efficient portfolio frontier includes those portfolios that yield the highest expected return for a given level of volatility.

Naturally, in these examples, the portfolio frontier is given, in that the parameter values are given. However, these values should be determined in equilibrium. For example, the portfolio with zero volatility seems to offer investors too good opportunities: it yields positive expected returns with no risk. Can this be sustainable? Many investors would spot this opportunity, and implement a trade that would make this opportunity disappear. Let us illustrate. If $\rho = -1$, the portfolio that achieves zero volatility has $\frac{\pi_2}{w} = \frac{\sigma_1}{\sigma_1+\sigma_2}$, and the expected return is $\mu_{o,p} \equiv \frac{\sigma_2}{\sigma_1+\sigma_2} b_1 + \frac{\sigma_1}{\sigma_1+\sigma_2} b_2$. But as more investors go long this portfolio to exploit this nice opportunity, the two asset prices go up and their returns decrease. This process eventually stops once b_1 and b_2 adjust in a way such that $\mu_{o,p} = r$, the risk-free rate.

Naturally, it is a simple example, which relies on hypothetical price adjustments. However, this example illustrates the type of standard restrictions arising in financial economics, which make prices economically viable, or free from arbitrage opportunities, as we shall explain in many junctures of these lectures. For example, in this chapter, Eq. (1.13) is the restriction that generalizes the previous reasoning to the case in which assets returns are not perfectly and negatively correlated.

1.2.5 The global minimum variance portfolio

Note that the portfolio in Eq. (1.9) can be decomposed into two components, as follows:

$$\frac{\hat{\pi}(v_p)}{w} = \ell(v_p)\frac{\pi_d}{w} + [1 - \ell(v_p)]\frac{\pi_g}{w}, \quad \ell(v_p) \equiv \frac{D\left(\mu_p(v_p)C - D\right)}{AC - D^2}, \tag{1.12}$$

where

$$\frac{\pi_d}{w} \equiv \frac{\Sigma^{-1}b}{D}, \qquad \frac{\pi_g}{w} \equiv \frac{\Sigma^{-1}\mathbf{1}_m}{C}.$$

Therefore, any portfolio on the efficient portfolio frontier can be constructed by choosing a convex combination of two portfolios, π_d and π_g, with weights equal to $\ell(v_p)$ and $1-\ell(v_p)$. It's a *mutual fund (or separation) theorem*.[1] We shall use this representation of efficient portfolios to derive the zero-beta CAPM of Section 1.3.2.

Regarding the two portfolios of the decomposition in (1.12), the first, π_d, achieves the expected return that makes $\ell(v_p)=1$, that is, $(v_p,\mu_p)=\left(\frac{\sqrt{A}}{D},\frac{A}{D}\right)$. The second portfolio, π_g, is the *global minimum variance portfolio*, for we know from Eq. (1.10) that the minimum variance occurs at $(v_p,\mu_p)=\left(\sqrt{\frac{1}{C}},\frac{D}{C}\right)$, that is, when $\ell(v_p)=0$.[2] This portfolio displays a few interesting properties. *Risk-parity* practice.

[In progress]

1.2.6 The market portfolio

The *market portfolio* is the portfolio at which the CML in Eq. (1.6) and the efficient portfolio frontier in Eq. (1.11) intersect. In fact, the market portfolio is the point at which the CML is *tangent* to the efficient portfolio frontier. For this reason, the market portfolio is also referred to as the "tangent portfolio." In Figure 1.2, the market portfolio is at point M and has volatility equal to v_M and expected return equal to μ_M. At this point, the CML is tangent to the efficient portfolio frontier, AMC.[3] We now develop the reasons for referring to the tangent portfolio as the market portfolio.

1.2.6.1 Two-fund separation

Figure 1.2 illustrates that the CML dominates the efficient portfolio frontier AMC. Let us explain. On the one hand, the CML is the value of the investor's problem, [1.P1], obtained while investing in all the risky assets *and* the riskless asset. On the other hand, the efficient portfolio frontier is the value of the investor's problem, [1.P2], obtained while investing in the risky assets *only*.[4] Then, the CML and the efficient portfolio frontier can only be tangent to each other. For suppose not. Then, there would exist a point on the efficient portfolio frontier that dominates some portfolio on the CML, a contradiction. Likewise, the CML must have a portfolio in common with the efficient portfolio frontier—the portfolio that does not include the safe asset. Below, we shall rely on these insights and provide an anlytical characterization of the market portfolio.

So why do we term the market portfolio in this way? Figure 1.2 reveals that any portfolio on the CML can be obtained as a combination of the safe asset and the market portfolio M (a portfolio containing only the risky assets). Investors with high risk aversion would like to choose a portfolio such as Q, and investors with low risk aversion would perhaps pick a portfolio such as P. But no matter how risk averse an investor is, his optimal choice is a mix of the safe

[1] But then any efficient portfolio can be constructed through a convex combination of any two arbitrary but distinct efficient portfolios (see Appendix 1).

[2] The covariance of the global minimum variance portfolio with any other portfolio is always C^{-1} (see Appendix 1).

[3] The existence of the market portfolio requires a restriction on r, derived in Eq. (1.13) below.

[4] Figure 1.2 also depicts the dotted line MZ, which is the value of the investor's problem when he invests a proportion higher than 100% in the market portfolio, leveraged at an interest rate for borrowing higher than the interest rate for lending. In this case, the CML coincides with rM, up to the point M. From M onwards, the CML coincides with the highest between MZ and MA.

asset and the market portfolio M. Thus, the market portfolio allows an investor to calibrate his desired, overall exposure to risk: he will invest those portions of his wealth in the market portfolio that make his overall exposure to risk consistent with his risk appetite. It's a *two-fund separation theorem*. We shall make these statements more rigorous in the following sections.

This separation theorem has equilibrium implications that naturally lead us to refer the tangent portfolio to as "market portfolio." We know that any portfolio can be attained by investing in the portfolio M and by lending or borrowing funds in zero net supply. In equilibrium, then, every investor must hold some proportions of M. But in aggregate, there is no net borrowing or lending and all investors must then have portfolio holdings that sum up to the market portfolio, which is therefore the value-weighted portfolio of all the existing risky assets. This argument is formally developed in the appendix.

Note, finally, that the market portfolio is a mere convex combination of all the existing risky assets. At first sight, it appears to be "preference-free," in that its construction is independent of the investors' risk attitudes. However, this proposition needs to be nuanced: below, we shall explain that the vector of expected returns b still needs to be estimated and that models of risk and risk aversion might be needed to characterize b.

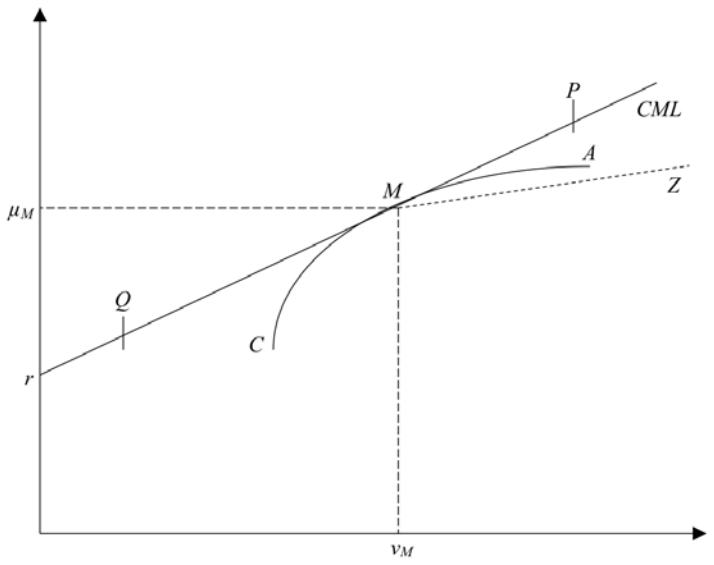


FIGURE 1.2. The Capital Market Line and the market portfolio.

1.2.6.2 Analytical characterization

We turn to characterize the market portfolio. We need to assume that the interest rate is sufficiently low to allow the CML to be tangent at the efficient portfolio frontier. The condition that ensures this is that the return on the safe asset be less than the expected return on the global minimum variance portfolio, viz

$$r < \frac{D}{C}. \tag{1.13}$$

That is, we cannot find risky portfolios that yield less than the safe asset.

Next, let π_M be the market portfolio. To identify π_M, we note that it belongs to AMC if $\pi_M^\top \mathbf{1}_m = w$, where π_M also belongs to the CML and, therefore, by Eq. (1.5), is such that

$$\frac{\pi_M}{w} = \frac{\Sigma^{-1}(b - \mathbf{1}_m r)}{\sqrt{\text{Sh}}} v_M. \tag{1.14}$$

Therefore, we search for the value v_M that solves

$$w = \mathbf{1}_m^\top \pi_M = \mathbf{1}_m^\top w \frac{\Sigma^{-1}(b - \mathbf{1}_m r)}{\sqrt{\text{Sh}}} v_M,$$

that is

$$v_M = \frac{\sqrt{\text{Sh}}}{D - Cr}. \tag{1.15}$$

Finally, we plug this value of v_M into the expression for π_M in Eq. (1.14) and obtain,

$$\frac{\pi_M}{w} = \frac{1}{D - Cr} \Sigma^{-1}(b - \mathbf{1}_m r). \tag{1.16}$$

Once again, the market portfolio belongs to the efficient portfolio frontier.[5] Indeed, on the one hand, the market portfolio cannot be above the efficient portfolio frontier, as this would contradict the efficiency of the AMC curve (obtained by investing in the risky assets only). On the other hand, by construction, the market portfolio belongs to the CML and so it cannot be below the efficient portfolio frontier, as the CML dominates the efficient portfolio frontier. The Appendix makes this reasoning rigorous, and shows that the market portfolio does indeed satisfy the tangency condition.

1.2.6.3 Asset allocation puzzles

Are these predictions observed in reality? Financial advisors are known to recommend young investors to hold more risky positions; they are also known to advice less conservative investors to increase their stock holdings compared to bonds. However, according to the theory in this section, the stocks-bonds mix in the market portfolio should be the same, regardless of the investors' attitude to risk (see the discussion around Figure 1.2). It is an instance of "asset allocation puzzles."

Campbell and Viceira (2002) provide an early synthesis of how these puzzles can be addressed through dedicated extensions of the framework in this section. These or additional analysts' recommendations could be understood while assuming that investors face more risks than those directly related to the price fluctuations of the assets in their portfolios. Some of these risks (not necessarily traded, e.g., labor income or random volatility) could indeed be hedged by investing more in risky positions than predicted by the simple theory in this section. Chapter 5 provides discussions of these cases while framing them into subsequently developed notions, according to which, agents are assumed to have access to stochastic opportunity sets.

1.3 The CAPM

This section provides analysis that goes beyond the investors' efficient portfolio choices, and derives the first asset evaluation formula appearing in the literature, based on the celebrated

[5] While the market portfolio depends on r, this portfolio does not obviously include any share in the safe asset.

Capital Asset Pricing Model (CAPM). First, we derive the CAPM, relying on arguments that have the same flavor as those in the original derivation of Sharpe (1964). Second, we derive the zero-beta CAPM of Black (1972), a model that does not require the existence of pure risk-free asset. Third, we explain how the CAPM predictions relate to the behavior of risk averse investors. Finally, this section applies the ideas of the CAPM to deal with such disparate topics as project evaluation or "speculative" demand of money.

1.3.1 Restrictions on securities expected returns

Consider a portfolio comprising a proportion x of wealth invested in any asset i and the remaining proportion $1-x$ invested in the market portfolio. That is, we consider a portfolio that is "parametrized" by x. The expected return and volatility of this portfolio are

$$\begin{cases} \tilde{\mu}_p \equiv xb_i + (1-x)\mu_M \\ \tilde{v}_p \equiv \sqrt{(1-x)^2\sigma_M^2 + 2(1-x)x\sigma_{iM} + x^2\sigma_i^2} \end{cases} \tag{1.17}$$

where we have defined $\sigma_M \equiv v_M$. Clearly, the market portfolio M is a special case of this portfolio. Relying on the examples of Section 1.2.4, we know that the curve in (1.17) has the same shape as the curve $A'Mi$ in Figure 1.3. The curve $A'Mi$ lies below the efficient portfolio frontier AMC, because the latter results from optimizing a mean-variance criterion over all the existing assets, which then dominates any portfolio that only comprises the two assets i and M.

For example, assume the $A'Mi$ curve intersects the AMC curve. Then, a feasible combination of assets in the $A'Mi$ curve would dominate points on AMC, a contradiction, as AMC is the most efficient feasible combination of all existing assets. On the other hand, the $A'Mi$ curve has a point in common with the AMC curve, which is M (the portfolio obtained with $x=0$). Therefore, the curve $A'Mi$ is tangent to the efficient portfolio frontier AMC at M, which is tangent to the CML at M, as we know.

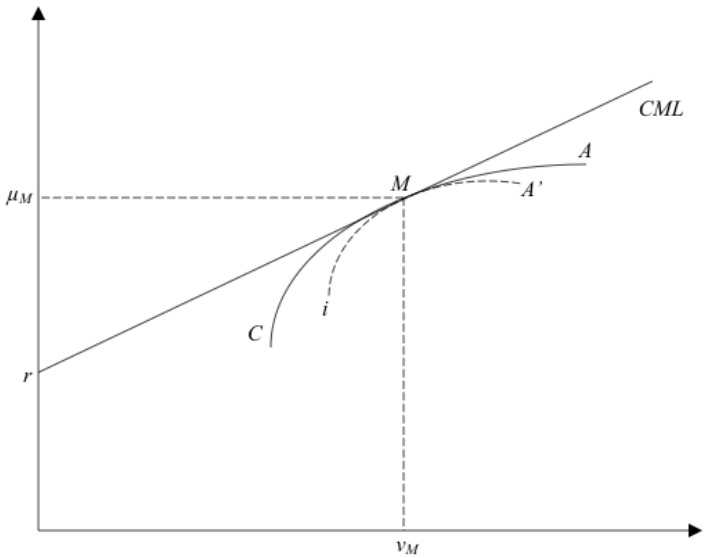


FIGURE 1.3. Construction of the Capital Asset Pricing Model.

That is, at M, the slopes of the $A'Mi$ curve is the same as that of the efficient portfolio frontier AMC. This condition provides a restriction on the expected return b_i on any asset i, as we now show. Consider the x-parametrized curve (1.17). We determine its slope at M through the two slopes, $d\tilde{\mu}_p/\,dx$ and $d\tilde{v}_p/\,dx$, both evaluated at $x=0$. We have

$$\frac{d\tilde{\mu}_p}{dx} = b_i - \mu_M, \quad \left.\frac{d\tilde{v}_p}{dx}\right|_{x=0} = -\frac{-(1-x)\sigma_M^2 + (1-2x)\sigma_{iM} + x\sigma_i^2|_{x=0}}{\tilde{v}_p|_{x=0}} = \frac{1}{\sigma_M}\left(\sigma_{iM} - \sigma_M^2\right).$$

Therefore,

$$\left.\frac{d\tilde{\mu}_p(x)}{d\tilde{v}_p(x)}\right|_{x=0} = \frac{b_i - \mu_M}{\frac{1}{\sigma_M}\left(\sigma_{iM} - \sigma_M^2\right)}. \tag{1.18}$$

On the other hand, the slope of the CML is $(\mu_M - r)/\,\sigma_M$ which, equated to the slope in Eq. (1.18), yields

$$b_i - r = \beta_i\left(\mu_M - r\right), \quad \beta_i \equiv \frac{\sigma_{iM}}{v_M^2}, \quad i = 1, \cdots, m. \tag{1.19}$$

Eq. (1.19) is the celebrated *Security Market Line* (SML) (see Appendix 2 for an alternative derivation). The assets for which $\beta_i > 1$ are usually referred to as "aggressive" and the assets for which $\beta_i < 1$ are called "conservative."[6] The rationale is the following. Assuming a positive beta, the higher this beta, the more volatile (and risky) the asset, and the higher its expected return. Assets for which beta is negative return on average less than the safe interest, because they provide diversification. Thus, according to the model, assets that command an expected return less than the safe interest rate can still be quite volatile, notably when their (negative) beta is sizeable.

The SML can be interpreted as a projection of the excess return on asset i (i.e. $\tilde{b}_i - r$) on the excess returns on the market portfolio (i.e. $\tilde{b}_M - r$). In other words,

$$\tilde{b}_i - r = \beta_i(\tilde{b}_M - r) + \varepsilon_i, \quad i = 1, \cdots, m. \tag{1.20}$$

Eq. (1.20) leads to the following decomposition of the volatility regarding the return in the i-th asset:

$$\sigma_i^2 = \beta_i^2 v_M^2 + var\left(\varepsilon_i\right), \quad i = 1, \cdots, m.$$

The quantity $\beta_i^2 v_M^2$ is known as *systematic* risk. The quantity $var\left(\varepsilon_i\right) \geq 0$ does, instead, capture the notion of *idiosyncratic* risk. In the next section, we shall show that idiosyncratic risk can be eliminated through a "well-diversified" portfolio—roughly, a portfolio that contains a large number of assets.

1.3.2 Restrictions on alternative investment returns

The CAPM is a founding stone in financial economics. Section 1.5 surveys reasons for motivating extensions of this model aimed to a realistic description of asset returns. Additional motivation to extend the model arises when the purpose is to assess returns on alternative investments.

[6] Interestingly, this notion of beta links to the following classical regression-based hedging context. Suppose we hold a (perhaps highly illiquid) asset that promises to deliver a return equal to $\tilde{z}$. To hedge against this stochastic return, go long a portfolio comprising a proportion π in the market portfolio and $1-\pi$ in a safe asset. We minimize the variance of the overall position by choosing $\hat{\pi} = \arg\min_x var[\tilde{z} - ((1-\pi)\,r + \pi\tilde{b}_M)] = \beta_{\tilde{z}} \equiv cov(\tilde{z}, \tilde{b}_M)/v_m^2$. That is, the beta of the asset is the proportion of the market portfolio that we use to hedge the asset against systematic market movements.

How do we assess investment performance relating to alternative investments such as, say, private equity, real estate, private debt, art, or hedge funds? Hedge funds typically employ "non-linear" strategies to exploit a risk-premium arising due to some particular event or market conditions. For example, these strategies can be event-driven (i.e., driven by corporate events such as an earnings call, or by a scheduled announcement by a central bank), or they can be non-directional (as it is typically the case when trying to express a view on market or exchange rate volatility; see Chapter 11); furthermore, to implement these trades, one typically relies on derivatives instruments that have non-linear payoffs, and some of these derivatives may actually trade on relatively illiquid venues.

The CAPM extensions dealt with in Section 1.5 regard pricing confined to standard assets traded in relatively liquid venues, such as equity traded on well-organized exchanges. [Mention references to work that aim to assess strategies leading to nonlinear returns in the AI world.] [In progress]

1.3.3 Zero-beta CAPM

The CAPM predicts that the market portfolio is a benchmark for all the assets required returns. We can actually use any benchmark on the efficient portfolio frontier, not only the market portfolio, to gauge each asset expected return. In particular, we pick an arbitrary portfolio on the efficient frontier and consider another portfolio that has returns uncorrelated with the initial arbitrary portfolio. The spread generated by these two portfolios becomes the evaluation benchmark. Because we only deal with portfolios that have no riskless assets, the procedure may be useful when the risk-free asset is not available for trading.

So, consider any portfolio on the efficient frontier, which can be generated by Eq. (1.12). Given a fixed weigth equal to l, define from Eq. (1.12) $v_p : \ell(v_p) = l$, and let $\hat{\pi}_l \equiv \hat{\pi}\left(v_p^{-1}(l)\right)$, where $v_p^{-1}(\cdot)$ denote the inverse function of ℓ. The return generated by the portfolio $\frac{\hat{\pi}_l}{w}$ is obviously $b_l \equiv \frac{\hat{\pi}_l^\top b}{w}$. The vector of the covariances of all the asset returns with b_l is

$$cov(\tilde{b}, \tilde{b}_l) = \Sigma \frac{\hat{\pi}_l}{w} = \frac{l}{D} b + \frac{1-l}{C} \mathbf{1}_m. \tag{1.21}$$

In particular, we have that for any asset i,

$$cov(\tilde{b}_i, \tilde{b}_l) = \frac{l}{D} b_i + \frac{1-l}{C}, \tag{1.22}$$

and, then,

$$var(\tilde{b}_l) = cov(\tilde{b}_l, \tilde{b}_l) = \frac{l}{D} b_l + \frac{1-l}{C}. \tag{1.23}$$

Eqs. (1.22) and (1.23) can be solved for $\frac{l}{D}$ and $\frac{1-l}{C}$, and the solutions can be replaced back into Eq. (1.21), leaving

$$b = \frac{\left(var(\tilde{b}_l) b_i - cov(\tilde{b}_i, \tilde{b}_l) b_l\right) \mathbf{1}_m - cov(\tilde{b}, \tilde{b}_l)(b_l - b_i)}{var(\tilde{b}_l) - cov(\tilde{b}_i, \tilde{b}_l)}.$$

The previous expression simplifies for assets that have returns uncorrelated with $\tilde{b}_l$, say some asset $z_l : cov(\tilde{b}_{z_l}, \tilde{b}_l) = 0$. We have

$$b = b_{z_l} \mathbf{1}_m + \frac{cov(\tilde{b}, \tilde{b}_l)}{var(\tilde{b}_l)} (b_l - b_{z_l}). \tag{1.24}$$

Eq. (1.24) is the Black's (1972) zero-beta CAPM. We may express expected returns in terms of "benchmark" returns, just as with the SML of Eq. (1.19), without however having to rely on the market portfolio and a riskless asset. The spread referred to as at the beginning of this section is $b_l - b_{z_l}$, where b_{z_l} is the expected return on an asset that has zero correlation with the benchmark portfolio on the efficient frontier (whence, "zero-beta"). It goes without saying that Eq. (1.24) does not rely on whether riskless assets are actually available for trading or not.

1.3.4 An excursion into risk-premiums and certainty equivalents

1.3.4.1 Project evaluation

The CAPM is a model that determines the required return for any asset and, thus, is a very first tool that we can use to evaluate any risky project, not necessarily a very liquid one. The main prediction of the model is that the market is a pricing factor, being the only source of risk against which we assess the expected return on any other asset. Let $\tilde{C}$ denote the future, random, cash flow promised by a given project. Its return is $\tilde{b}_C = \frac{\tilde{C}}{V} - 1$, where V is the current value of the project, and by the CAPM, the risk-adjusted discount rate for this project is

$$b_C \equiv E(\tilde{b}_C) = r + \beta_C \left(\mu_M - r\right) = r + \frac{1}{V}\frac{cov(\tilde{C}, \tilde{b}_M)}{v_M}\lambda, \tag{1.25}$$

where $\lambda \equiv \frac{\mu_M - r}{v_M}$ is the unit market risk-premium and $\beta_C = cov(\tilde{b}_C, \tilde{b}_M)/v_M^2$ is the project beta. The value of the project is

$$V = \frac{E(\tilde{C})}{1 + b_C}, \tag{1.26}$$

and replacing the expression for b_C in Eq. (1.25) into Eq. (1.26) and rearranging terms, leaves

$$V = \frac{E(\tilde{C}) - \frac{\lambda}{v_M} cov(\tilde{C}, \tilde{x}_M)}{1 + r}. \tag{1.27}$$

The project value is the discounted expectation of the random cash flow, $\tilde{C}$, "corrected for risk," in the following sense. According to the CAPM, the market is the only source of risk relevant for pricing, as explained. If the project cash flows are positively correlated with the market, investing in the project does not insulate the investments from this source of risk: the higher this correlation, the more aggressive the project is, just as we said in previous sections. But the more aggressive the project, the higher its expected return and, hence, the lower the project value. The risk-adjustment term, $\frac{\lambda}{v_M} cov(\tilde{C}, \tilde{x}_M)$, determines how lower the project value has to be to induce an investor to undertake this project. It is a *risk-premium* required to invest.

Were the project cash flows uncorrelated with the market (or were the market return the same as the riskless asset, $\lambda = 0$), Eq. (1.27) would collapse to an actuarially fair evaluation formula, by which the project value is simply the discounted expectation of the future cash flows. Quite naturally, these statistical correlations occur under the probability laws generating returns and cash flows.

Is there any probability law, such that the project value is the same as in Eq. (1.27), but with the numerator replaced by the cash flow expectation under these hypothetical new laws, and without any risk-adjustments? The answer is in the affirmative. For obvious reasons, these laws are known as "risk-neutral." They constitute indeed one fundamental pillar of asset evaluation,

which these lectures heavily rely on.[7] Section 1.4 provides an introductory discussion, although the next chapters provide a systematic treatment of these foundational issues.

1.3.4.2 With utility functions

Consider an individual with an initial wealth equal to w, who is considering entering into a gamble that yields a random payoff $\tilde{x}$ and costs V. We can think of this "gamble" as the investment decision in the previous section, and set $r = 0$ to simplify the analysis. We assume the investor will enjoy a random utility $u(w - V + \tilde{x})$ from the project, where the function u is increasing and concave. The property that the utility function is concave is tantamount to the assumption our decision maker is risk averse. That is, the investor would prefer receiving the expected outcome of the project rather than exposing himself to "gambling for his utility." Formally, by Jensen's inequality, $E\left(u\left(w - V + \tilde{x}\right)\right) \leq u\left(w - V + E\left(\tilde{x}\right)\right)$.

We ask, what is the value V that induces the investor to enter into this project? Clearly, it cannot be higher than the solution V to

$$u\left(w\right) = E\left(u\left(w - V + \tilde{x}\right)\right), \tag{1.28}$$

for otherwise the investor would be better off while staying away from the gamble (and experiencing utility equal to $u\left(w\right)$). We also assume perfect competition: should the project value be less than V, many additional investors with the same attitude for risk would step in and make this value increase to the extent that Eq. (1.28) holds.

Next, define the de-meaned payoff $\epsilon \equiv \tilde{x} - E(\tilde{x})$. We study the investor's attitude vis-à-vis risk in relation to the fair gamble, i.e., the lottery that has payoff equal to ϵ. We have

$$u\left(w\right) = E\left(u\left(w + \Lambda + \epsilon\right)\right) \leq u\left(w + \Lambda\right), \tag{1.29}$$

where we have defined $\Lambda \equiv E(\tilde{x}) - V$. The first equality is simply Eq. (1.28) (with the change in notation), and the inequality is Jensen's inequality again. Thus, Λ is positive. That is, due to risk aversion, the investor is willing to pay less than the expected cash flow to enter the project. For obvious reasons, we refer to Λ as the risk-premium required to invest in the project.

Suppose, for example, that a representative investor has exponential utility, $u(w) = -e^{-\tau w}$, and that the cash flow is normally distributed, $\tilde{C} \sim N(E(\tilde{C}), \sigma_\epsilon^2)$, such that by Eq. (1.28),

$$V = E(\tilde{C}) - \frac{1}{2}\tau\sigma_\epsilon^2. \tag{1.30}$$

The project value is at discount compared to the expected cash flow by a factor equal to $\Lambda = \frac{1}{2}\tau\sigma_\epsilon^2$, the risk-premium. Note that this risk-premium relates to the investor's risk aversion coefficient, τ: the higher τ, the higher the discount. We shall elaborate on this risk aversion coefficient below.

These conclusions are the utility counterparts to those of the CAPM in the previous section. According to the CAPM, the risk-premium relates to the project exposure to market movements (see Eq. (1.27)). In the model of this section, the risk-premium arises to compensate a risk averse investor for the risk he would undertake in this specific project (see Eq. (1.30)). Granted, in equilibrium, the market movements may well be related to the investors' attitude vis-à-vis

[7] More formally, the project value can be expressed as $V = \frac{\mathbb{E}(\tilde{C})}{1+r}$, where $\mathbb{E}(\cdot)$ denotes the expectation under this risk-neutral probability, say Q. For the CAPM, the relation between Q and the original, natural probability, say P, is given by the so-called Radon-Nikodym derivative $\frac{dQ}{dP} = 1 - \frac{\lambda}{v_M}(\tilde{x}_M - x_M)$ (see Section 1.4.2 for introductory details).

risk. However, the link to risk aversion is direct in this model. Section 1.3.5 clarifies how asset prices link to risk aversion in an equilibrium model, one in which agents choose amongst several alternatives, as in the CAPM.

Apart from exceptions (e.g., the previous Gaussian case), no closed-form solutions are available for V. We can, however, approximate Λ when the risks regarding the fair lottery are "small," that is, when the variance of ϵ, σ_ϵ^2 say, is small. Appendix 3 provides the following approximation to Λ in this small risks case:

$$\Lambda \approx \hat{\Lambda} \equiv -\frac{1}{2}\frac{u''(w)}{u'(w)}\sigma_\epsilon^2. \tag{1.31}$$

For example, if the representative investor has exponential utility, $\hat{\Lambda} = \frac{1}{2}\tau\sigma_\epsilon^2$, which is the exact risk-premium in the previous Gaussian case (see Eq. (1.30)). Appendix 3 provides one additional example of evaluation models, in which no closed-form solutions are available for V, and discusses the quality of the approximations to V based on Eq. (1.31).

1.3.4.3 With utility functions II: certainty equivalents

An alternative albeit closely related notion of risk-premium relies on the following reasoning. Assume that an individual *is entering* into a gamble. That is, he has no choice but being about to experience a random perturbation to his wealth, leaving him with $w+\tilde{x}$ dollars.[8] To illustrate, suppose that the individual expects a loss from this lottery, $E(\tilde{x}) < 0$. We ask, how much of his wealth would he be willing to give up if hypothetically asked to *avoid* the gamble?

If the individual were risk-neutral, he would be glad to give up the expected loss. But if he is risk averse, he would be willing to give up more than the expected loss, say $-E(\tilde{x}) + \Pi$, for some positive Π interpreted as a risk-premium. In other words, the individual would be glad to accept a *sure* reduction of his wealth to $w + E(\tilde{x}) - \Pi$, in exchange of avoiding to gamble,

$$u(w + E(\tilde{x}) - \Pi) = E(u(w + E(\tilde{x}) + \epsilon)) \leq u(w + E(\tilde{x})), \tag{1.32}$$

where the inequality is Jensen's inequality, and shows that Π is positive. We refer to $\text{CE}(w, \tilde{x}) \equiv E(\tilde{x}) - \Pi$ as the *certainty equivalent* for $\tilde{x}$ for an initial wealth equal to w. Naturally, the sign of $E(\tilde{x})$ can be anything.

The risk-premium Π has an "insurance flavor." It relates to an economic premium that the individual is willing to pay to avoid a loss, rather than to an economic incentive to undertake a risky project. However, it closely relates to the risk-premium Λ defined in Eq. (1.29). Consider, for example, the Gaussian project case in the previous subsection: if $\tilde{C}$ is Gaussian and the utility is exponential with risk aversion τ, the risk-premium in this section is $\Pi = \frac{1}{2}\tau\sigma_\epsilon^2$, i.e., the same as Λ in the previous section (see Eq. (1.30)). Furthermore, and just as with the approximation in Eq. (1.31), the risk-premium Π relates to the curvature of the utility function u, in that, in the small risks case, it can be approximated as follows (see Appendix 3):

$$\Pi \approx \hat{\Pi} \equiv -\frac{1}{2}\frac{u''(w + E(\tilde{x}))}{u'(w + E(\tilde{x}))}\sigma_\epsilon^2. \tag{1.33}$$

Note that $\hat{\Pi}$ and $\hat{\Lambda}$ are the same as soon as the initial lottery is fair, $E(\tilde{x}) = 0$ (see Eq. (1.31)). We now discuss alternative measures of risk-premiums.

[8] As an example, think of these wealth fluctuations as those arising during the business cycles.

1.3.4.4 Constant and relative risk aversion

The notions of risk premiums investigated so far relate to *dollar* incentives to undertake (or avoid) risks. The expressions in Eqs. (1.31) and (1.33) reveal that they are proportional to these risks (i.e., σ_ϵ^2), and that they increase with the "normalized curvature" of the utility function, $A(\cdot) \equiv -u''(\cdot)/u'(\cdot)$, a function we refer to as *absolute risk aversion*. The numerator of this function naturally measures how far the investor is from being risk-neutral. The denominator of A makes this notion invariant to any positive and affine transformation of the original utility function: if two risk averse investors have utilities u_1 and u_2, they have the same absolute risk aversion if and only if $u_1 = \varphi(u_2)$ and φ is an affine function. The exponential utility function hypothesized in the previous section is one for which absolute risk aversion is constant and equal to τ; it is known as CARA utility for this reason.

These notions can be extended to cover *relative* incentives to undertake (or avoid) risks. Consider, for example, the insurance risk-premium of the previous section, one, however, regarding a multiplicative risk, i.e., a lottery with an outcome equal to $(w + E(\tilde{x}))(1 + \tilde{g})$ for some zero-mean random variable $\tilde{g}$, not $w + E(\tilde{x}) + \epsilon$ (an additive risk). The question we ask now is, what *proportion* of his wealth would the investor be willing to give up if hypothetically asked to avoid this lottery? The indifference condition in (1.32) is now replaced with

$$u(w + E(\tilde{x})(1 - \Pi_r)) = E(u((w + E(\tilde{x}))(1 + \tilde{g}))) \leq u(w + E(\tilde{x})), \tag{1.34}$$

where Π_r is a constant, relative (or percentage) risk-premium. The previous inequality reveals that Π_r is positive. In Appendix 3, we show that this relative risk-premium can approximated as follows

$$\Pi_r \approx \hat{\Pi}_r \equiv -\frac{1}{2}\frac{u''(w + E(\tilde{x}))(w + E(\tilde{x}))}{u'(w + E(\tilde{x}))}\sigma_g^2, \tag{1.35}$$

where σ_g^2 denotes the variance of $\tilde{g}$.

We refer the function $R_r(x) \equiv -u''(x)x/u'(x)$ to as *relative risk aversion*. Consider the utility function $u(x) = \frac{x^{1-\gamma}}{1-\gamma}$ for some constant $\gamma \neq 1$, or $u(x) = \ln x$ when $\gamma = 1$. In this case, relative risk aversion is constant and equal to γ; it is known as CRRA utility for this reason. Naturally, a CARA investor has increasing relative risk aversion. The implication of this property is that an investor keeps the same amount of dollars invested in risky assets as his wealth increases. However, the relative proportion of risky assets in his portfolio decreases with his wealth, as explained next. These notions of risk aversion are known since at least Pratt (1964) and Arrow (1965).

1.3.5 Back to CAPM: Equilibrium with expected utility

We explained that the investors' risk aversion determines the proportion of the market portfolio to hold and achieve any desired point of the risk-return trade-off. We formalize these explanations. We assume that our investors are risk averse: they prefer higher expected returns than lower, but for any expected portfolio return, they like portfolios that provide less volatility. We assume investors are all the same, and the representative investor chooses portfolio holdings that maximize his utility derived from the expected returns and variance, viz

$$\pi^* = \arg\max_{\pi} U(E[w'(\pi)], var[w'(\pi)]), \tag{1.P3}$$

where $U(\cdot,\cdot)$ is a concave utility function satisfying $U_1(\cdot,\cdot) > 0$ and $U_2(\cdot,\cdot) < 0$, subscripts denote partial derivatives, and the expressions for the portfolio expected return and variance are given in Eq. (1.3).

Rearranging the first order conditions to problem [1.P3] leads to the following portfolio choice

$$\pi^* = \tau_*^{-1}\Sigma^{-1}\left(b - \mathbf{1}_m r\right), \quad \tau_* \equiv \frac{-2U_2^*}{U_1^*}, \tag{1.36}$$

where $U_j^* \equiv U_j\left(E\left[w'\left(\pi^*\right)\right], var\left[w'\left(\pi^*\right)\right]\right)$ and subscripts denote partial derivatives. That is, the investor's choice is a proportion of the market portfolio (see Eq. (1.16)). Moreover, the more risk averse the investor is (i.e., the higher τ_*), the lower the proportion he desires to hold of the market portfolio.

The investor wishes to lend or to borrow according to whether $w - \mathbf{1}_m^\top \pi^*$ is positive or negative. If the riskless security is in zero net supply, then, in equilibrium, $w - \mathbf{1}_m^\top \pi^* = 0$, that is, the representative investor holds exactly the market portfolio, which is the value-weighted portfolio, $\pi_i^* = \bar{\theta}_i S_i$, where $\bar{\theta}_i$ is the number of the i-th outstanding securities. Note, indeed, that if the riskless security is in zero net supply,

$$w = \mathbf{1}_m^\top \pi^* = \tau_*^{-1}\mathbf{1}_m^\top \Sigma^{-1}\left(b - \mathbf{1}_m r\right) = \tau_*^{-1}\left(D - Cr\right). \tag{1.37}$$

Replacing the previous relation into Eq. (1.36) gives exactly the market portfolio, $\frac{\pi^*}{w} = \frac{\pi_M}{w}$, where π_M is as in Eq. (1.16). Note that Eq. (1.37) places a restriction on the safe interest rate, to which we shall return below.

1.3.5.1 Equilibrium expected returns

What are the equilibrium expected returns? By Eq. (1.36), we have that

$$b - \mathbf{1}_m r = \tau_* w \cdot \Sigma \frac{\pi^*}{w} = \tau_* w \cdot cov(\tilde{b}, \tilde{b}_M), \tag{1.38}$$

where $\tilde{b}_M = \tilde{b}^\top \frac{\pi^*}{w}$ denotes the market return, and the second equality follows by results in Appendix 2 (see Eq. (1A.12)). Moreover, note that, in equilibrium,

$$b_M = \frac{\pi^{*\top}}{w} b = r + \tau_* w \cdot \frac{\pi^{*\top}}{w} \Sigma \frac{\pi^*}{w} = r + \tau_* w \cdot v_M^2, \tag{1.39}$$

where the second equality follows by replacing b from Eq. (1.38). Eqs. (1.38) and (1.39) do obviously lead to the CAPM in Eq. (1.19),

$$b - \mathbf{1}_m r = \frac{cov(\tilde{b}, \tilde{b}_M)}{v_M^2}\left(b_M - r\right),$$

but they now also lead to a determination of the overall system of the expected excess returns: by Eqs. (1.38) and (1.39), the expected excess returns on all assets (including the market portfolio) link to the representative agent's attitudes vis-à-vis risk.

1.3.5.2 A parametric example

A simple example helps illustrate these conclusions. Suppose the investor maximizes the expected utility of wealth, taken to be negative exponential, $u(w) = -e^{-\tau w}$, where τ is his absolute risk aversion. Assume that the asset returns are normally distributed, such that

$$\pi^* = \arg\max_{\pi} E\left[u\left(w'\left(\pi\right)\right)\right] = \arg\max_{\pi}\left(E\left[w'\left(\pi\right)\right] - \frac{\tau}{2} var\left[w'\left(\pi\right)\right]\right). \tag{1.40}$$

The solution to this problem is the same as that in Eqs. (1.36), where τ^* collapses to the constant τ.[9] In equilibrium, the safe interest rate satisfies Eq. (1.37), such that, now, the investor's equilibrium asset holdings coincide with the market portfolio.

We may elaborate on these results with a further simplification of this example, one with two risky assets that have uncorrelated payoffs. Assume, also, that the asset return volatilities are entirely driven by the asset payoffs, in that $\sigma_i^2 S_i^2 = \sigma_{x_i}^2$, where $\sigma_{x_i}^2$ is the variance of the payoff of the i-th asset. It is easy to see that each asset demand is $\pi_i^* = (\tau\sigma_i^2)^{-1}(b_i - r) = \bar{\theta}_i S_i$, where the second equality is the equilibrium condition, and by the definition of b_i, the equilibrium price is then

$$S_i = S_0\left(E(\tilde{x}_i) - \tau\bar{\theta}_i\sigma_{x_i}^2\right), \tag{1.41}$$

where we have set $x_0 \equiv 1$, such that $S_0 \equiv (1+r)^{-1}$. We assume that these prices are strictly positive.

To dermine the equilibrium riskless rate r (or, equivalently, S_0), we use the zero-net supply condition for the safe asset, that is, we aggregate the value of the asset holdings and set the result equal to the initial wealth, $\bar{\theta}_1 S_1 + \bar{\theta}_2 S_2 = w$, leaving

$$S_0 = \frac{w}{\bar{\theta}_1 E(\tilde{x}_1) + \bar{\theta}_2 E(\tilde{x}_2) - \tau(\bar{\theta}_1^2\sigma_{x_1}^2 + \bar{\theta}_2^2\sigma_{x_2}^2)}. \tag{1.42}$$

The single risky asset case is equally instructive. Assume that the second asset is not available for trading. In this case, the wealth of the representative investor equals $w = \bar{\theta}_1 S_1$, where S_1 is as in Eq. (1.41), and the the price of the safe asset is given by Eq. (1.42), with $\bar{\theta}_2 \equiv 0$. That is, the equilibrium riskless rate simply forces the value of the risky asset holdings to equal initial wealth, $S_0 : \bar{\theta}_1 S_1 = w$.[10]

1.3.5.3 Tobin's re-interpretation of Keynesian money demand

One of the earliest applications of the ideas of portfolio selection occurred in monetary economics. In 1958, Tobin provided a new framework to think about money demand, aimed to provide microeconomic foundations to some traits of the monetary theory of Keynes (1936). In this novel framework, money demand arises as the agent invests his savings while optimizing over his risk-returns trade-offs based on his risk aversion. Tobin's work would actually inspire the very same Sharpe's ideas leading to the CAPM. Let us review the main points underlying Tobin's contribution.

Tobin explains that taken at face-value, Keynes' explanations of money demand would imply that agents end up making dichotomic choices: they would either hold money or bonds. That is, at the individual level, each agent holds money or bonds based on his own expectations about future interest rate levels. Yet at the aggregate level, money demand does inversely relate to nominal interest rates while being flat in correspondence of small rate values—a "liquidity trap." The Appendix provides a parameteric example that clarifies the details of how these mechanisms operate.

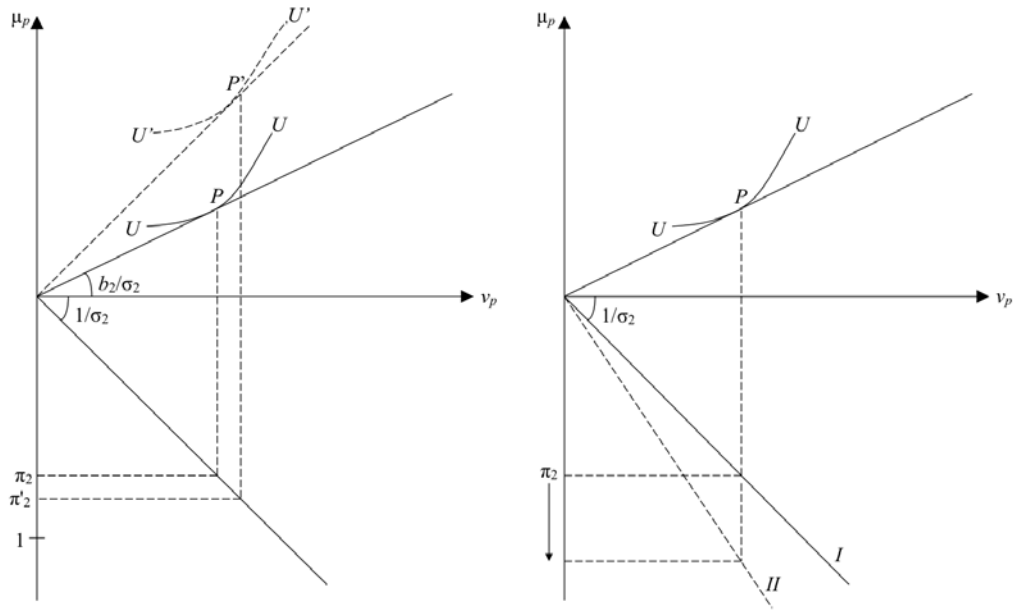
[9] Note that the investor desires to hold relative proportions of the risky assets (π^*/w) that are decreasing in w. This property follows by the assumption that this investor has constant absolute risk-aversion (see Section 1.3.3).

[10] Consider applying the notion of (insurance) risk-premium in Section 1.3.3 to our current portfolio problem, and set $r = 0$ and $\bar{\theta}_1 = 1$ to simplify. The insurance premium Π does then solve Eq. (1.32), with $\tilde{x} \equiv w' - w$. That is, $u\left(E(w') - \Pi\right) = E(u(w'))$, whence $\Pi = \frac{1}{2}\tau var\left(w'\right)$, where $var\left(w'\right) = \sigma_1^2$ in equilibrium. However, as the investor optimizes his risk-return tradeoff, the equilibrium expected *excess* return on his wealth is $E\left(\tilde{w}\right) - w = \tau\sigma_1^2$. This required return relates to, but differs from, Π; it is the equilibrium outcome of a diversification process, i.e., the optimality condition regarding the allocation of wealth amongst safe and risky assets.

Tobin formulates a theory of money demand in which agents do not make previous dichotomic choices. Consider the following specification of the examples in Section 1.2.4. We interpret the safe asset as "money," which is therefore such that its return and volatility are $b_1 \equiv 0$ and $\sigma_1 \equiv 0$; instead, bonds are risky, in that they provide a superior return but at the cost of some volatility. Therefore, the expected return and volatility of a portfolio comprising money and bonds are $\mu_p = b_2\pi_2$ and $v_p = \pi_2\sigma_2$, with straightforward notation. We have

$$\mu_p = \frac{b_2}{\sigma_2} \cdot v_p, \quad \pi_2 = \frac{1}{\sigma_2} \cdot v_p. \tag{1.43}$$

The top panels of next two pictures plots the first of Eqs. (1.43), along with indifference curves of an hypothetical representative agent. The agent finds his optimum is achieved at P, which is the point of tangency between the first of Eqs. (1.43) and the indifference curve UU. Money demand, $1 - \pi_2$, is determined by the second of Eqs. (1.43), and is shown in the bottom panels. As the expected return on the bond, b_2, increases, the new optimum shifts to P', the tangency point between the new relation in Eqs. (1.43) and the indifference curve $U'U'$. Money demand decreases as a result.



This framework of analysis can be used to study the effects of a decreased interest rate volatility. Suppose the central bank has the power to lower both interest rates *and* interest rate volatility in such a way to keep the ratio b_2/σ_2 unchanged. As the second picture above illustrates, the optimum is still P, although then the second of Eqs. (1.43) becomes steeper than before the policy action, by shifting from the I line to the II line. Money demand decreases as a result. We can say more. As is clear, money demand decreases as interest rate volatility, σ_2, decreases. Therefore, in this model, the central bank might keep money supply constant and achieve lower interest rates by simply targeting low interest rate volatility.

1.4 The APT

1.4.1 Exact APT

1.4.1.1 No-arb restrictions on expected returns

Suppose that the m asset returns we observe are generated by the following *linear factor model*,

$$\underset{m\times 1}{\tilde{b}} = \underset{m\times 1}{b} + \underset{m\times k}{B} \cdot \underset{k\times 1}{f} \equiv b + cov(\tilde{b}, f)[var(f)]^{-1} \cdot f \tag{1.44}$$

where b and B are a vector and a matrix of constants, and f is a k-dimensional vector of factors supposed to affect the asset returns, with $k \leq m$. The vector f is the vector of risks that affects developments in asset returns.

Let us normalize $[var(f)]^{-1} = I_{k\times k}$, so that $B = cov(\tilde{b}, f)$. With this normalization, we have,

$$\tilde{b} = b + \begin{bmatrix} cov(\tilde{b}_1, f) \\ \vdots \\ cov(\tilde{b}_m, f) \end{bmatrix} \cdot f = b + \begin{bmatrix} \sum_{j=1}^{k} cov(\tilde{b}_1, f_j) f_j \\ \vdots \\ \sum_{j=1}^{k} cov(\tilde{b}_m, f_j) f_j \end{bmatrix}. \tag{1.45}$$

Next, consider a portfolio of m risky assets and a riskless asset. The wealth generated by this portfolio is given by Eq. (1.2), which applied to this model leaves,

$$w' = \pi^\top (b - \mathbf{1}_m r) + Rw + \pi^\top B f. \tag{1.46}$$

An arbitrage opportunity arises, in this context, if there exists some portfolio π such that the wealth generated by this portfolio, w' in Eq. (1.46) is certain, and different from the safe gross interest rate R, i.e. if $\exists \pi : \pi^\top B = 0$ and $\pi^\top (b - \mathbf{1}_m r) \neq 0$. Mathematically, this is ruled out whenever $\exists \lambda \in \mathbb{R}^k : b = \mathbf{1}_m r + B\lambda$. Substituting this relation into Eq. (1.44) leaves,

$$\tilde{b} = \mathbf{1}_m r + B\lambda + Bf = \mathbf{1}_m r + cov(\tilde{b}, f)\lambda + cov(\tilde{b}, f) f.$$

Taking expectations, we have, that for each asset return,

$$b_i = r + (B\lambda)_i = r + \sum\nolimits_{j=1}^{k} \underbrace{cov(\tilde{b}_i, f_j)}_{\equiv \beta_{i,j}} \lambda_j, \qquad i = 1, \cdots, m. \tag{1.47}$$

Let f_j be the j-th component of the vector of risks in f. The economic interpretation of Eq. (1.47) is that to be induced to invest in a risky market, I need a a risk-premium, which compensates me beyond the risk-free rate. This risk premium is $\sum_{j=1}^{k} \beta_{i,j}\lambda_j$. For each asset i, it is a linear combination of the "betas" $\beta_{i,j}$, which are the exposures of the asset return i to the sources of risks f_j, times the "unit risk-premiums" λ_j relating to each f_j. So naturally, the unit risk-premiums are common to all asset returns, although of course the return exposures, $\beta_{i,j}$, which can vary across i.

1.4.1.2 Project evaluation

Project evaluation under the exact APT obtains under a straightforward generalization of Eq. (1.26). For any project with random cash-flow equal to $\tilde{C}_i$, we have that its random return is $\tilde{b}_i = \frac{\tilde{C}_i}{V_i} - 1$, such that the value, V_i, is given by

$$V_i = \frac{E(\tilde{C}_i)}{1 + b_i}, \tag{1.48}$$

where b_i is as in Eq. (1.47). In Section 1.4.2, we shall explain how to represent the value of any project, in terms of an alternative probability.

1.4.1.3 APT and CAPM

The APT collapses to the CAPM, once we assume that the only factor affecting the returns is the market portfolio. To show this, we must normalize the market portfolio return so that its variance equals one, consistently with Eq. (1.47). So let $\tilde{r}_M$ be the normalized market return, defined as $\tilde{r}_M \equiv v_M^{-1}\tilde{b}_M$, so that $var(\tilde{r}_M) = 1$. We have

$$\tilde{b}_i = b + \beta_i \tilde{r}_M, \qquad i = 1, \cdots, m,$$

where $\beta_i = cov(\tilde{b}_i, \tilde{r}_M) = v_M^{-1} cov(\tilde{b}_i, \tilde{b}_M)$. Then, we have

$$b_i = r + \beta_i \lambda, \qquad i = 1, \cdots, m. \tag{1.49}$$

In particular, $\beta_M = cov(\tilde{b}_M, \tilde{r}_M) = v_M^{-1} var(\tilde{b}_M) = v_M$, and so, by Eq. (1.49),

$$\lambda = \frac{b_M - r}{v_M}, \tag{1.50}$$

which is known as the *Sharpe ratio* for the market portfolio, or the market price of risk.

By replacing $\beta_i = v_M^{-1} cov(\tilde{b}_i, \tilde{b}_M)$ and the expression for λ above into Eq. (1.49), we obtain,

$$b_i = r + \frac{cov(\tilde{b}_i, \tilde{b}_M)}{v_M^2}\left(b_M - r\right), \qquad i = 1, \cdots, m.$$

This is simply the SML in Eq. (1.19).

1.4.2 Risk-neutral tilts, or the fundamental theorem of asset pricing

Only if the market were risk-neutral, would Eq. (1.47) predict that the expected return on any risky asset collapse to the safe interest rate, $b_i = r$. But, is there a way to construct a risk-neutral market, i.e. one with a zero risk-premium, from the original market where the risk-premium $\sum_{j=1}^{k} \beta_{i,j}\lambda_j$ is different from zero? The answer is in the affirmative. It is a quite fundamental theme in financial economics. It links to the celebrated *fundamental theorem of asset pricing*—often referred to as the "FTAP."

1.4.2.1 Probability twists

Let us develop intuition on such a beautiful result, by elaborating a simple example. Assume that each of the risks f_j, is standard normal: $P\left(df_j\right) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} f_j^2) df_j$ and next, that we tilt their densities by a factor equal to $\zeta\left(f_j\right) = \exp(-\frac{1}{2}\lambda_j^2 - \lambda_j f_j)$, where λ_j is the unit-risk premium, as defined in Section 1.4.1. This tilt defines a new probability under which each factor f_j is distributed. Let us determine this probability, by tilting P through ζ,

$$Q(df_j) = \zeta(f_j)P(df_j) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}f_j^2 - \frac{1}{2}\lambda_j^2 - \lambda_j f_j\right) df_j = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\tilde{f}_j^2\right) d\tilde{f}_j, \tag{1.51}$$

where

$$\tilde{f}_j = f_j + \lambda_j. \tag{1.52}$$

Note that the new density, Q, is still that of standard normal variate. Yet under Q, it is $\tilde{f}_j$ to have zero expectation, not f_j. In other words, we have that under Q, (i) $\tilde{f}_j$ is standard normal and (ii) f_j is normal with unit variance, but expectation equal to $-\lambda_j$. That is, assuming that

$\lambda_j > 0$, f_j has a lower expectation under Q than under P—under P, this expectation is zero, and under Q, it is $-\lambda_j$.

We label the new probability Q risk-neutral probability, for the following reasons. Consider Eqs. (1.45) and (1.47), which say that under P,

$$\tilde{b}_i = r + \sum_{j=1}^{k} \beta_{i,j}\lambda_j + \sum_{j=1}^{k} \beta_{i,j} f_j. \tag{1.53}$$

We also know that under the new probability Q, each factor f_j is normally distributed with mean $-\lambda_j$. That is, replacing Eq. (1.52) into Eq. (1.53), we have that under Q,

$$\tilde{b}_i = r + \sum_{j=1}^{k} \beta_{i,j}\tilde{f}_j,$$

To summarize, the return on each assets $\tilde{b}_j$, has the following distributions under P and under Q,

$$\tilde{b}_j \sim N_P\left(r + \textstyle\sum_{j=1}^{k} \beta_{i,j}\lambda_j, \sigma_j^2\right), \quad \tilde{b}_j \sim N_Q\left(r, \sigma_j^2\right), \tag{1.54}$$

where $\sigma_j^2 = \Sigma_{jj}$, $\Sigma \equiv BB^\top$, as in Eq. (1.3), and N_P and N_Q denote the Normal densities under P and under Q. That is, under Q, the expected return on each asset equals r, whence the "risk-neutral probability" label. In later chapters, we shall use the celebrated Girsanov's theorem to elaborate on these topics, and label the tilt ζ in Eq. (1.51), as the Radon-Nikodym derivative of P against Q (see Chapter 5, Section 5.3.3).

1.4.2.2 Derivative asset evaluation

Why complicating everything through the previous probability changes? In fact, the entire building block underlying asset evaluation, relies on similar risk-neutral tilts. Consider the following example of derivative evaluation. We wish to price a "quadratic derivative," i.e. one that pays off the square of the cash-flow promised by the first asset, $\tilde{C}_1^2$. It is challenging to evaluate this derivative through "APT software" in this Gaussian market, because $\tilde{C}_1^2$ is obviously not normally distributed, which complicates the reasoning underlying its exposure to the factors f_j. In fact, in this Gaussian market, we cannot restrict the behavior of the expected return on this derivative, without assuming something more. Let us explain.

Suppose we want to construct a portfolio of the existing assets to replicate the payoff of the quadratic derivative, for each possible value this derivative could take. Can we do this? The answer is in the negative. We cannot use a finite number of assets to span an asset payoff, which could take a continuum of values, such as $\tilde{C}_1^2$. We say markets are *incomplete* in this context. In the next chapter, we shall see that the price of such and related derivatives, can be found "in a preference-free format," as soon as the number of assets is at least as large as the number of states—markets are, then, *complete*. In this case, a portfolio that replicates the derivative's payoff can be found, and its value is the same as the derivative's, for two assets are worth the same whenever they promise the same payoff.

Chapter 2 explains that in a world with complete markets, the price of the existing traded assets can be inverted for the shadow value of some elementary assets, those that pay off one unit of numéraire in a given state of the world, and zero otherwise. The price of these elementary securities can then be used to price any derivative, which is redundant indeed.

Chapter 5 explains how these results can be generalized to markets with a continuum of states, as soon as we assume that there exist a number of sufficiently diverse "elementary" securities, which guarantee a payoff could be delivered for each state of nature. To illustrate through the example of this section, consider the following elementary security, which promises the following payoff,

$$\Pi(s) = \begin{cases} 1, & \text{if } \tilde{C}_1 \in (s, s+ds) \\ 0, & \text{otherwise} \end{cases} \tag{1.55}$$

and let $\phi(s)ds$ be its current price. We shall refer these securities as "Arrow-Debreu" securities in these lectures, for reasons explained in the next chapter.

We could utilize all of these Arrow-Debreu securities, i.e. for all $s \in \mathbb{R}$, and replicate any generic function of the state, $\psi(s)$, including our original payoff, $\psi(s) = s^2$, $s \in \mathbb{R}$. Indeed, note that by purchasing $\psi(s)$ units of the security that pays off $\Pi(s)$ in Eq. (1.55), we pay $\psi(s)\phi(s)ds$ today, and are guaranteed to receive $1 \times \psi(s)$ tomorrow in state $\tilde{C}_1 \in (s, s+ds)$, and zero otherwise. Therefore, by purchasing all the securities that span $\mathbb{R}$, we shall receive $\psi(\tilde{C}_1)$ for any possible value of $\tilde{C}_1$, for sure, tomorrow, and pay, today,

$$\mathcal{C}_\psi \equiv \int_{-\infty}^{\infty} \psi(s)\phi(s)ds. \tag{1.56}$$

We call P_ψ such a portfolio. We claim that the value of the derivative, V_ψ say, is just $\mathcal{C}_\psi$. For suppose not, and assume, for instance, that $V_\psi > \mathcal{C}_\psi$. Then, we could sale short the derivative for V_ψ, invest $\mathcal{C}_\psi$ into the portfolio P_ψ, and retain an arbitrage profit equal to $V_\psi - \mathcal{C}_\psi$. It is an arbitrage profit, because the portfolio P_ψ delivers the exact payoff we need to honour the short-sale of the derivative.

The crucial point is to determine the value of ϕ. We claim that,

$$\phi(s) = \frac{1}{1+r} q\left(s; V_1(1+r), V_1^2\sigma_1^2\right), \tag{1.57}$$

where $q(s; m, \sigma^2)$ denotes the density of the risk-neutral normal distribution N_Q in (1.54), but with mean and variances m and σ^2 as given in Eq. (1.57). Indeed, let us take expectations of $\tilde{C}_1 = V_1(1+\tilde{b}_1)$ under Q, such that

$$V_1 = \frac{E_Q(\tilde{C}_1)}{1+r}. \tag{1.58}$$

On the other hand, let us apply Eq. (1.56) to determine the value of the "derivative" that pays off, $\psi(\tilde{C}_1) = \tilde{C}_1$, which is

$$V_1 = \int_{-\infty}^{\infty} s\phi(s)ds. \tag{1.59}$$

Comparing Eq. (1.58) and Eq. (1.59) yields Eq. (1.57).[11]

We are now ready to evaluate the quadratic derivative, by relying on Eq. (1.56), and the expression of ϕ in Eq. (1.57). We have for $\psi(s) = s^2$, that the value of the derivative, say $V_\mathcal{C}$, is:

$$V_\mathcal{C} = \int_{-\infty}^{\infty} s^2\phi(s)ds = \frac{E_Q(\tilde{C}_1^2)}{1+r}. \tag{1.60}$$

[11] We can check that Eq. (1.57) is consistent with the pricing of a pure discount bond. Such a bond has a payoff equal to $\psi(s) = 1$ for all s, such that by Eq. (1.56), $\frac{1}{1+r} = \int_{-\infty}^{\infty}\phi(s)ds$, which it does, by Eq. (1.57).

The expression in Eq. (1.60) tells us that all we have to do is to re-cast the initial evaluation problem in terms of this new, risk-neutral, setup. To fix ideas, suppose that the unit prices of risk λ_j are all positive. In this setup, the discounting factor is the safe interest rate, r. To compensate with such as generous discounting, the expectation in the numerator of Eq. (1.60) is lower than that under P, because the distribution of $\tilde{C}_1$ and, hence, $\tilde{C}_1^2$, is more skewed to the left under Q than under P, due to the factors driving the value of $\tilde{C}_1 = V_1(1+\tilde{b}_1)$ being skewed to take more "pessimistic" values under Q, on average, by a factor of $-\lambda_j$.

Let us proceed with the determination of $V_{\mathcal{C}}$, by using Eq. (1.60). We have that $E_Q(\tilde{C}_1^2) = V_1^2 E_Q((1+\tilde{b}_1)^2)$, where $1+\tilde{b}_1 \sim N_Q\left(1+r, \sigma_1^2\right)$, such that, by a direct calculation,

$$V_{\mathcal{C}} = V_1^2 \left(\frac{\sigma_1^2}{1+r} + 1 + r \right),$$

where V_1 is the value of the first asset, determined as usual through Eq. (1.48). Note how simple this formula is. It links the value of the derivative to the square of the value of the underlying risk, V_1^2, and the discounted expectation of σ_1^2, reflecting that after all, a quadratic derivative is about a play in volatility.[12]

Remarkably, the price of this derivative does not require any knowledge of the risk-premium components, λ_j. It is, thus, a "preference-free" formula. It is the task of the next chapter to develop the deep reasons for which derivatives can sometimes be expressed in such a simple way. To anticipate, the derivative relies on a risk, which is already traded in the market, in that the cash flow $\tilde{C}_1$ is traded at V_1. All risks have, then, already been embedded into the market price, V_1. Naturally, the market value V_1, may not be preference free: indeed, V_1 is as in Eq. (1.58), and the expected cash flow under the risk-neutral prob may well depend on risk-premium adjustments. However, the "percentage change" $\frac{E_Q(\tilde{C}_1)-V_1}{V_1}$ is always r, a property that we shall use in the next subsection.

1.4.3 Uncertainty and asset evaluation

How asset prices relate to the volatility of fundamentals? It is a very old issue in financial economics.[13] Consider a two-period market for a cash ψ to be paid in the second period, where we assume that $\psi = \psi(\tilde{C})$ for some random variable $\tilde{C}$. Accordingly, and generalizing Eq. (1.60), $V_{\mathcal{C}} \equiv (1+r)^{-1} E_Q[\psi(\tilde{C})]$ is the current value of a derivative that pays $\psi(\tilde{C})$.

We ask: how does the value $V_{\mathcal{C}}$ change when uncertainty regarding $\tilde{C}$ increases? For example, how does $V_{\mathcal{C}}$ change after a change in the variance of $\tilde{C}$? It is a complex question, and the answers to it depend on the assumptions we make about the uncertainty changes surrounding $\tilde{C}$. In this chapter, we begin with the simplest situation, one in which a change in the *uncertainty* of $\tilde{C}$ does not entail a change in the *expected value* of $\tilde{C}$. This case is actually very relevant. Suppose that that $\tilde{C}$ is the cash flow of a traded security with a given value V, such that its current value is, consistently with results in the previous sections, $E_Q(\tilde{C}) = V(1+r)$, which is clearly independent of the uncertainty surrounding $\tilde{C}$ once we take V to be given, just as we are doing.

[12] In continuous-time, the price at time t of a quadratic derivative, which promises to pay off the square of an asset price, $S^2(T)$ at T, is given by $S^2(t)\, e^{(\sigma^2+r)(T-t)}$, provided S is a Geometric Brownian motion with volatility parameter equal to σ (see Chapter 5, Section 5.3.3.1).

[13] See, e.g., the classical work of Malkiel (1979), Pindyck (1984), Poterba and Summers (1985), Abel (1988), Barsky (1989), among others.

Situations in which changes in the volatility of a random variable do not affect its mean are known as *mean-preserving spreads*. Appendix 4 contains a succinct summary of these situations and develops results that predict that $V_{\mathcal{C}}$ does indeed increase after a mean-preserving spread in $\tilde{C}$, provided ψ is a convex function of $\tilde{C}$. Intuitively, a convex function "exaggerates" favorable realizations of $\tilde{C}$ and "dampens" the poor ones. The theory of mean-preserving spreads was introduced by Rothschild and Stiglitz (1970, 1971) and will be used at different junctures of these lectures.

1.4.4 The APT with idiosyncratic risk and a large number of assets

[Ross (1976), and Connor (1984), Huberman (1983).]

How can idiosyncratic risk be eliminated? Consider, for example, Eq. (1.20). Intuitively, we may form portfolios with a large number of assets, so as to make idiosyncratic risk negligible, by the law of large numbers. But would the beta-relation still hold, in this case? More in general, would the APT relation in Eq. (1.47) be still valid? The answer is in the affirmative, although it deserves some qualifications.

Consider the APT equation (1.44), and "add" a vector of idiosyncratic returns, ε, which are independent of f, and have mean zero and variance σ_{ε}^2:

$$\tilde{b} = b + B \cdot f + \varepsilon.$$

We wish to show that in the absence of some appropriate notion of arbitrage, to be defined below, it must be that the number of assets such that Eq. (1.47) does *not* hold, $N(m)$ say, is bounded as m gets large, i.e.:

$$|b_i - ((B\lambda)_i + r)| > 0, \quad i = 1, \cdots, N(m), \tag{1.61}$$

where

$$\lim_{m\to\infty} N(m) < \infty. \tag{1.62}$$

In other words, we wish to show that in a "large" market, Eq. (1.47) does indeed hold for most of the assets, an approach close to that in Huang and Litzenberger (1988, p. 106-108).

By the same arguments leading to Eq. (1.1), the wealth generated by a portfolio of the assets satisfying (1.61), $w'_{N(m)}$ say, is,

$$w'_{N(m)} = \pi_{N(m)}^{\top}\left(b_{N(m)} - \mathbf{1}_{N(m)} r\right) + R w_{N(m)} + \pi_{N(m)}^{\top}\left(B_{N(m)} f + \varepsilon_{N(m)}\right),$$

where b_N, B_N and ε_N are (i) the vector of the expected returns, (ii) the return volatility (or factor exposures) matrix and (iii) the vector of idiosyncratic return components affecting these assets, and, finally, π_N and w_N are the portfolio and the initial wealth invested in these assets.

In this context, we may define an arbitrage as the portfolio $\pi_{N(m)}$ that in the limit, as the number of the existing assets m gets large, is riskless and yet delivers an expected return strictly larger than the safe interest rate, viz

$$\lim_{m\to\infty} \frac{E[w'_{N(m)}]}{w_{N(m)}} > R, \quad \text{and} \quad \lim_{m\to\infty} var[w'_{N(m)}] \to 0. \tag{1.63}$$

We want to show that this situation does not arises, under the condition in (1.62), thereby establishing that the linear APT relation in Eq. (1.47) is valid for most of the assets, in a large market.

So suppose the linear relation, $b_N - \mathbf{1}_N r = B_N \lambda$, doesn't hold. Then, there exists a portfolio $\underline{\pi}$ such that,

$$\underline{\pi}^\top B_N = 0 \quad \text{and} \quad \underline{\pi}^\top (b_N - \mathbf{1}_N r) \neq 0. \tag{1.64}$$

Consider the portfolio:

$$\hat{\pi}_N = \frac{1}{N} \cdot sign\left(\underline{\pi}^\top (b_N - \mathbf{1}_N r)\right) \cdot \underline{\pi},$$

where $\underline{\pi}$ is as in (1.64). With this portfolio we have, clearly, that $E[w'_N] = \hat{\pi}_N^\top (b_N - \mathbf{1}_N r) + Rw_N > Rw_N$, for each N, and even for N large. That is, $\lim_{m\to\infty} E[w'_{N(m)}]/w_{N(m)} > R$, which is the first condition in (1.63). As regards the second condition in (1.63), we have that

$$var[w'_N] = \hat{\pi}_N^\top \left(B_N B_N^\top + \sigma_\varepsilon^2 I_{N\times N}\right) \hat{\pi}_N = \sigma_\varepsilon^2 \hat{\pi}_N^\top \hat{\pi}_N,$$

where the second equality follows by the first relation in (1.64). Clearly, $\lim_{m\to\infty} var[w'_{N(m)}] \to 0$ as $N(m) \to \infty$. Hence, in the absence of arbitrage, the condition in (1.62) must hold.

1.4.5 Systematic risk

Portfolio diversification & systematic versus non-systematic risk —> Well-diversified portfolios. [In progress]

1.5 Empirical evidence

[In progress]

1.5.1 Fama & Mac Beth

How to estimate Eq. (1.20)? Consider a slightly more general version of Eq. (1.20), where the safe interest rate is time-varying:

$$\tilde{b}_{i,t} - r_t = \beta_i(\tilde{b}_{M,t} - r_t) + \varepsilon_{i,t}, \quad i = 1, \cdots, m,$$

where $\varepsilon_{i,t}$ denote "time-series residuals." Fama and MacBeth (1973) consider the following procedure. In a first step, one obtains estimates of the exposures to the market, $\hat{\beta}_i$ say, for all stocks, using, for example, monthly returns, and approximating the market portfolio with some broad stock market index.[14] In a second step, one runs cross-sectional regressions, one for each month,

$$\tilde{b}_{i,t} - r_t = \alpha_{it} + \lambda_t \hat{\beta}_i + \eta_{i,t}, \quad t = 1, \cdots, T,$$

where T is the sample size and $\eta_{i,t}$ denote "cross-sectional residuals." The time-series of cross-sectional estimates of the intercept $\alpha_{i,t}$ and the price of risk λ_t, $\hat{\alpha}_{i,t}$ and $\hat{\lambda}_t$ say, are, then, used to make statistical inference. For example, time-series averages and standard errors of $\hat{\alpha}_{i,t}$ and $\hat{\lambda}_t$ lead to point estimates and standard errors for $\alpha_{i,t}$ and λ_t. If the CAPM holds, estimates of α_i should not be significantly different from zero.

[14] In tests of the CAPM, one uses proxies of the market portfolio, such as, say, the S&P 500. However, the market portfolio is unobservable. Roll (1977) points out that as a result, the CAPM is inherently untestable, as any test of the CAPM is a joint test of the model itself and of the closeness of the proxy to the market portfolio.

1.5.2 Macroeconomic forces

Chen, Roll and Ross (1986) use the Fama-MacBeth two-step procedure to estimate a multifactor APT model, such as that in Section 1.4. They identify "macroeconomic forces" driving asset returns with the innovations in variables such as the term spread, expected and unexpected inflation, industrial production growth, or the corporate spread. They find that these sources of variation in the cross-section of asset returns are significantly priced.

1.5.3 Fama & French

Consider the Security Market Line in Eq. (1.19), which predicts that each asset display an average excess return lying precisely on the SML. Assets delivering average excess returns and betas above the SML, as the points A, B, C, and D in Figure 1.4 below, would be simply evidence that this single factor version of the APT does not work. Consider, for example, the asset corresponding to point A. A regression of the excess return of this asset onto the excess return on the market would produce a positive intercept, some $\alpha > 0$, such that its average excess return would equal $\alpha + \beta_i (\mu_M - r)$, thereby invalidating Eq. (1.19).

Fama and French (1992, 1993) provide evidence that the standard CAPM is invalidated by the presence of two factors:

(i) *Size* (Banz, 1981): Average returns for "small firms," or low capitalized firms (in terms of market equity, defined as stock price times outstanding shares) are too high given their beta.

(ii) *Value* (Stattman, 1980; Rosenberg, Reid and Lanstein, 1985): Average returns on stocks of firms with high book-to-market (BM, henceforth) ratios, or "value stocks," are too high given their beta. In general, average returns on value stocks are higher than those on "growth" stocks, i.e. those stocks with low BM ratios. As an example, the points D, C, B, and A in Figure 1.4 might typically refer to stocks with low-to-high BM ratios.

A third factor providing evidence against the standard CAPM is "momentum":

(iii) *Momentum* (Jegadeesh and Titman, 1993): Stocks with the highest returns in the previous twelve months will outperform in the next future.

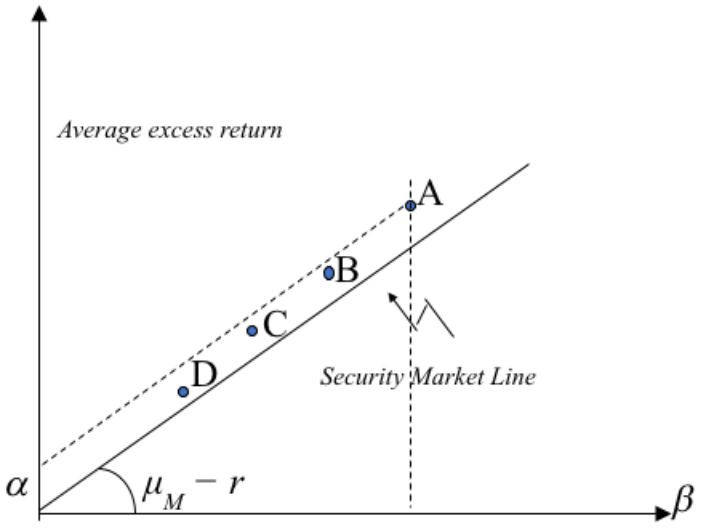


FIGURE 1.4.

The standard CAPM does not have power in explaining the cross-section of asset returns, sorted by size, BM or momentum. Assets sorted in this way command a size premium, a value premium, and a momentum premium. For example, one can create portfolios sorted by size and BM, say 25 portfolios, out of a 5×5 matrix with dimensions given by size and BM. The puzzle, then, at least from the standard CAPM perspective, is that this model cannot explain the returns on these porfolios. Fama and French (1993) show that the returns on these portfolios can be very much better understood by means of a multifactor model, where both size and value premiums are explicitly taken into account. They consider three factors: (i) the excess return on the market; (ii) an "HML" factor, defined as the monthly difference between the returns on assets with high and low BM ratios ("high minus low"); an "SMB" factor, defined as the difference between the asset returns of firms with small and big size ("small minus big"). The HML and SMB factors are defined as the differences between the returns on the appropriate cells of a 2×3 matrix, obtained through percentiles of the distribution of asset returns over the previous year.

	Book-to-Market		
Size	L	M	H
S			
L			

The resulting model is the celebrated Fama-French three factor model. Carhart (1997) extends this model to a four-factor model with a momentum factor: the monthly difference between the returns on the high and low prior return portfolios.

1.5.4 The high-beta stocks anomaly

High-beta stocks should command higher returns, to compensate for their higher volatility. Yet historically, it is low-beta stocks to have performed better, on a risk-adjusted basis (i.e., in terms of alphas). Theoretically, then, one could buy low-beta stocks and leverage them through debt. It's indeed feasible when you've got the opportunity to do so. Two important pieces are Frazzini, Kabiller and Pedersen (2012) and Frazzini and Pedersen (2012).

1.5.5 "Lucky factors"

[In progress]

1.6 Market practice

1.6.1 Risk-parity

1.6.2 Black-Litterman

1.6.3 Smart beta

1.6.4 Factor investing

1.7 Appendix 1: Details regarding portfolio choice

We derive Eq. (1.9), i.e., the optimal portfolio allocation in the case without the safe asset. We solve two programs: (i) the primal program [1.P2] in the main text (maximize portfolio expected return for a given variance); and (ii) a dual program, to be introduced below, by which we minimize the portfolio return variance for a given expected return.

1.7.1 The primal program

Given Eq. (1.8), the Lagrangian function associated to [1.P2] is

$$L = \pi^\top b + w - \nu_1(\pi^\top \Sigma \pi - w^2 v_p^2) - \nu_2(\pi^\top \mathbf{1}_m - w),$$

where ν_1 and ν_2 are two Lagrange multipliers. The first order conditions are

$$\hat{\pi} = \frac{1}{2\nu_1}\Sigma^{-1}\left(b - \nu_2 \mathbf{1}_m\right), \quad \hat{\pi}^\top \Sigma \hat{\pi} = w^2 v_p^2, \quad \hat{\pi}^\top \mathbf{1}_m = w. \tag{1A.1}$$

Using the first and the third conditions leaves

$$w = \mathbf{1}_m^\top \hat{\pi} = \frac{1}{2\nu_1}(\underbrace{\mathbf{1}_m^\top \Sigma^{-1} b}_{\equiv D} - \nu_2 \underbrace{\mathbf{1}_m^\top \Sigma^{-1} \mathbf{1}_m}_{\equiv C}) \equiv \frac{1}{2\nu_1}(D - \nu_2 C).$$

We solve for ν_2, obtaining

$$\nu_2 = \frac{D - 2w\nu_1}{C}.$$

By replacing the solution for ν_2 into the first condition in (10.113) leaves

$$\hat{\pi} = \frac{w}{C}\Sigma^{-1}\mathbf{1}_m + \frac{1}{2\nu_1}\Sigma^{-1}\left(b - \frac{D}{C}\mathbf{1}_m\right). \tag{1A.2}$$

Next, we derive the value of the program [1.P2]. We have

$$E\left[w'(\hat{\pi})\right] - w = \hat{\pi}^\top b = \frac{w}{C}D + \frac{1}{2\nu_1}\left(A - \frac{D^2}{C}\right), \quad A \equiv b^\top \Sigma^{-1} b. \tag{1A.3}$$

It is easy to show that

$$\begin{aligned} var\left[w'(\hat{\pi})\right] &= w^2 v_p^2 \\ &= \hat{\pi}^\top \Sigma \hat{\pi} \\ &= \left[\frac{w}{C}\mathbf{1}_m^\top \Sigma^{-1} + \frac{1}{2\nu_1}\left(b^\top - \frac{D}{C}\mathbf{1}_m^\top\right)\Sigma^{-1}\right]\left[\frac{w}{C}\mathbf{1}_m + \frac{1}{2\nu_1}\left(b - \frac{D}{C}\mathbf{1}_m\right)\right] \\ &= \frac{w^2}{C} + \left(\frac{1}{2\nu_1}\right)^2\left(A - \frac{D^2}{C}\right). \end{aligned} \tag{1A.4}$$

Let us gather Eqs. (1A.3) and (1A.4):

$$\begin{cases} \mu_p(v_p) \equiv \dfrac{E\left[w'(\hat{\pi})\right] - w}{w} = \dfrac{D}{C} + \dfrac{1}{2\nu_1 w}\left(A - \dfrac{D^2}{C}\right) \\ v_p^2 = \dfrac{1}{C} + \left(\dfrac{1}{2\nu_1 w}\right)^2\left(A - \dfrac{D^2}{C}\right) \end{cases} \tag{1A.5}$$

where we have emphasized the dependence of μ_p on v_p, which arises through the Lagrange multiplier ν_1, as formally seen below (see Eq. (1A.7)).

The first equation in (1A.5) can be solved for ν_1, as follows

$$\frac{1}{2\nu_1 w} = \left(AC - D^2\right)^{-1}\left(C\mu_p(v_p) - D\right). \tag{1A.6}$$

We use Eq. (1A.6) and express $\hat{\pi}$ in Eq. (1A.2) in terms of the portfolio expected return, $\mu_p(v_p)$. We have

$$\frac{\hat{\pi}}{w} = \frac{\Sigma^{-1}\mathbf{1}_m}{C} + \left(AC - D^2\right)^{-1}\left(C\mu_p(v_p) - D\right)\Sigma^{-1}\left(b - \frac{D}{C}\mathbf{1}_m\right).$$

By rearranging terms in the previous equation, we obtain Eq. (1.9) in the main text.

Finally, we substitute Eq. (1A.6) into the second equation of (1A.5), obtaining

$$v_p^2 = \frac{1}{C}\left[1 + \left(AC - D^2\right)^{-1}\left(C\mu_p(v_p) - D\right)^2\right],$$

which is Eq. (1.10) in the main text. Note, also, that the second condition in (1A.5) reveals that

$$\left(\frac{1}{2\nu_1 w}\right)^2 = \frac{Cv_p^2 - 1}{AC - D^2}. \tag{1A.7}$$

Given that $AC - D^2 > 0$, the previous equation confirms the properties of the *global minimum variance portfolio* stated in the main text.

1.7.2 The dual program

Next, consider the following program:

$$\hat{\pi} = \arg\min_{\pi\in\mathbb{R}^m} var\left[\frac{w'(\pi)}{w}\right] \quad \text{s.t. } E\left[w'(\pi)\right] = E_p \text{ and } w = \pi^\top \mathbf{1}_m, \tag{1A.P2-D}$$

for some constant E_p. The first order conditions are

$$\frac{\hat{\pi}}{w} = \frac{\eta_1 w}{2}\Sigma^{-1}b + \frac{\eta_2 w}{2}\Sigma^{-1}\mathbf{1}_m, \qquad \hat{\pi}^\top b = E_p - w, \qquad w = \hat{\pi}^\top \mathbf{1}_m, \tag{1A.8}$$

where η_1 and η_2 are two Lagrange multipliers.

Combining the first and second conditions in (1A.8) leaves

$$E_p - w = \hat{\pi}^\top b = w^2\left(\frac{\eta_1}{2}A + \frac{\eta_2}{2}D\right), \tag{1A.9}$$

and combining the first and third conditions in (1A.8) produces

$$w = \hat{\pi}^\top \mathbf{1}_m = w^2\left(\frac{\eta_1}{2}D + \frac{\eta_2}{2}C\right). \tag{1A.10}$$

Next, let $\mu_p \equiv \frac{E_p - w}{w}$. By Eqs. (1A.9) and (1A.10), the solutions for η_1 and η_2 are

$$\frac{\eta_1 w}{2} = \frac{C\mu_p - D}{AC - D^2}, \qquad \frac{\eta_2 w}{2} = \frac{A - D\mu_p}{AC - D^2}.$$

Therefore, the solution for the portfolio in Eq. (1A.8) is

$$\frac{\hat{\pi}}{w} = \frac{C\mu_p - D}{AC - D^2}\Sigma^{-1}b + \frac{A - D\mu_p}{AC - D^2}\Sigma^{-1}\mathbf{1}_m.$$

Finally, the value of the program is

$$var\left[\frac{w'(\hat{\pi})}{w}\right]=\frac{\hat{\pi}^{\top}}{w}\Sigma\frac{\hat{\pi}}{w}=\frac{\hat{\pi}^{\top}}{w}\left(\frac{C\mu_p-D}{AC-D^2}b+\frac{A-D\mu_p}{AC-D^2}\mathbf{1}_m\right)=\frac{C\mu_p^2-2D\mu_p+A}{AC-D^2},$$

where we have used the fact that (i) $\frac{\hat{\pi}^{\top}}{w}b=\mu_p$ and (ii) $\frac{\hat{\pi}^{\top}}{w}\mathbf{1}_m=1$. Rearranging terms leaves Eq. (1.10) in the main text.

To check the second-order conditions apply, consider the bordered Hessian

$$H=\begin{bmatrix}\frac{\partial^2 L}{\partial\pi\partial\pi} & \frac{\partial^2 L}{\partial\pi\partial\eta_1} & \frac{\partial^2 L}{\partial\pi\partial\eta_2}\\ \frac{\partial^2 L}{\partial\eta_1\partial\pi} & \frac{\partial^2 L}{\partial\eta_1\partial\eta_1} & \frac{\partial^2 L}{\partial\eta_1\partial\eta_2}\\ \frac{\partial^2 L}{\partial\eta_2\partial\pi} & \frac{\partial^2 L}{\partial\eta_2\partial\eta_1} & \frac{\partial^2 L}{\partial\eta_2\partial\eta_2}\end{bmatrix}=\begin{bmatrix}-2\eta_1\Sigma & -2\Sigma\pi & -\mathbf{1}_m\\ -2\pi\Sigma & 0 & 0\\ \mathbf{1}_m^{\top} & 0 & 0\end{bmatrix}.$$

It is negative (semi) definite whenever the leading principal minors (formed through the last k columns and corresponding rows, for $k=4,\cdots,m+2$) have determinants with signs that alternate, with the first one (formed with the last 4 rows and corresponding columns) having the sign of $(-1)^2=+1$. This is possible whenever $\eta_1>0$, which is true, by Eq. (1A.6), whenever $AC>D^2$.

1.7.3 Efficient portfolios generate efficient portfolios

Let π_1 and π_2 be two portfolios on the efficient portfolio frontier. Then, $\pi_i=\ell_i\pi_d+(1-\ell_i)\pi_g$ for some ℓ_i, and $i=1,2$, and solving for π_d and π_g leaves

$$\pi_d=\frac{1-\ell_1}{\ell_2-\ell_1}\pi_2-\frac{1-\ell_2}{\ell_2-\ell_1}\pi_1,\quad \pi_g=\frac{\ell_2}{\ell_2-\ell_1}\pi_1-\frac{\ell_1}{\ell_2-\ell_1}\pi_2.$$

Replacing these expressions for π_d and π_g into (1.12) leaves, for any given weight ℓ,

$$\hat{\pi}=\ell\pi_d+(1-\ell)\pi_g=\omega(\ell_1,\ell_2)\pi_1+(1-\omega(\ell_1,\ell_2))\pi_2,\quad \omega(\ell_1,\ell_2)\equiv\frac{\ell_2-\ell}{\ell_2-\ell_1}.$$

That is, any arbitrary efficient portfolio can be generated by any other two arbitrary efficient portfolios, $\ell\neq\ell_1\neq\ell_2$.

1.7.4 Covariance of global minimum variance and efficient portfolio returns

Let $\tilde{b}^p=\frac{\pi_p^{\top}}{w}\tilde{b}$ and $\tilde{b}^g=\frac{\pi_g^{\top}}{w}\tilde{b}$ the returns that can be obtained by any portfolio π_p, and π_g. The covariance of the global minimum variance portfolio with any other portfolio is

$$cov\left(\frac{\pi_p^{\top}}{w}\tilde{b},\frac{\pi_g^{\top}}{w}\tilde{b}\right)=\frac{\pi_p^{\top}}{w}cov(\tilde{b}\tilde{b}^{\top})\frac{\pi_g}{w}=\frac{\pi_p^{\top}}{w}\frac{\mathbf{1}_m}{C}=\frac{1}{C}.$$

1.8 Appendix 2: The market portfolio and the security market line

1.8.1 The tangent portfolio is the market portfolio

Let us define the market capitalization for any asset i as the value of all the assets i that are outstanding in the market, viz

$$\text{Cap}_i \equiv \bar{\theta}_i S_i, \quad i = 1, \cdots, m,$$

where $\bar{\theta}_i$ is the number of assets i outstanding in the market. The market capitalization of all the assets is simply

$$\text{Cap}_M \equiv \sum_{i=1}^{m} \text{Cap}_i.$$

The market portfolio, then, is the portfolio with relative weights given by,

$$\bar{\pi}_{M,i} \equiv \frac{\text{Cap}_i}{\text{Cap}_M}, \quad i = 1, \cdots, m.$$

Next, suppose there are N investors and that each investor j has wealth w_j, which he invests in two funds, a safe asset and the tangent portfolio. Let w_j^f be the wealth investor j invests in the safe asset and $w_j - w_j^f$ the remaining wealth the investor invests in the tangent portfolio. The tangent portfolio is defined as $\bar{\pi}_T \equiv \frac{\pi_T}{w_j}$, for some π_T solution to [1.P2], and is obviously independent of w_j (see Eq. (1.16) in the main text). The equilibrium in the stock market requires that

$$\text{Cap}_M \cdot \bar{\pi}_M = \sum_{j=1}^{N} \left(w_j - w_j^f \right) \bar{\pi}_T = \sum_{j=1}^{N} w_j \cdot \bar{\pi}_T = \text{Cap}_M \cdot \bar{\pi}_T.$$

where the second equality follows because the safe asset is in zero net supply and, hence, $\sum_{j=1}^{N} w_j^f = 0$; and the third equality holds because all the wealth in the economy is invested in stocks, in equilibrium.

1.8.2 Tangency conditions

We check that the CML and the efficient portfolio frontier have the same slope in correspondence of the market portfolio. Let us impose the following tangency condition of the CML to the efficient portfolio frontier in Figure 1.2, AMC, at the point M:

$$\sqrt{\text{Sh}} = \frac{AC - D^2}{C\mu_M - D} v_M. \tag{1A.11}$$

The left hand side of this equation is the slope of the CML, obtained through Eq. (1.6). The right hand side is the slope of the efficient portfolio frontier, obtained by differentiating $\mu_p(v)$ in the expression for the portfolio frontier in Eq. (1.11), and setting $v = v_M$ in

$$\frac{d\mu_p(v)}{dv} = \sqrt{\left(Cv^2 - 1\right)^{-1} (AC - D^2)} v = \frac{AC - D^2}{C\mu_p(v) - D} v,$$

and where the second equality follows, again, by Eq. (1.11). By Eqs. (1A.11) and (1.15), we need to show that,

$$\frac{C\mu_M - D}{AC - D^2} = \frac{1}{D - Cr}.$$

By plugging $\mu_M = r + \sqrt{\text{Sh}} \cdot v_M$ into the previous equality and rearranging terms,

$$v_M = \frac{\sqrt{\text{Sh}}}{D - Cr},$$

where we have made use of the equality $\text{Sh} = A - 2Dr + Cr^2$, obtained by elaborating on the definition of the Sharpe market performance Sh given in Eq. (1.4). This is indeed the variance of the market portfolio given in Eq. (1.15).

1.8.3 An alternative derivation of the SML

The vector of covariances of the m asset returns with the market portfolio are:

$$cov(\tilde{b}, \tilde{b}_M) \equiv cov(\tilde{b}, \tfrac{\pi_M^\top}{w}\tilde{b}) = cov(\tilde{b}\tilde{b}^\top)\tfrac{\pi_M}{w} = \Sigma\frac{\pi_M}{w} = \frac{1}{D - Cr}\left(b - \mathbf{1}_m r\right), \tag{1A.12}$$

where we have used the expression for the market portfolio given in Eq. (1.16). Next, premultiply the previous equation by $\frac{\pi_M^\top}{w}$ in Eq. (1.14) and obtain

$$v_M^2 = \frac{\pi_M^\top}{w}\Sigma\frac{\pi_M}{w} = \underbrace{\frac{\left(b - \mathbf{1}_m r\right)^\top \Sigma^{-1}}{\sqrt{\text{Sh}}} v_M}_{=\frac{\pi_M^\top}{w}} \cdot \underbrace{\frac{1}{D - Cr}\left(b - \mathbf{1}_m r\right)}_{=\Sigma\frac{\pi_M}{w}}, \tag{1A.13}$$

or $v_M = \frac{\sqrt{\text{Sh}}}{D - Cr}$, which confirms Eq. (1.15).

Let us rewrite Eq. (1A.12) component by component. That is, for $i = 1, \cdots, m$,

$$\sigma_{iM} \equiv cov(\tilde{b}_i, \tilde{b}_M) = \frac{1}{D - Cr}\left(b_i - r\right) = \frac{v_M}{\sqrt{\text{Sh}}}\left(b_i - r\right) = \frac{v_M^2}{\mu_M - r}\left(b_i - r\right),$$

where the last two equalities follow by Eq. (1A.13) and by the capital market line relation $\sqrt{\text{Sh}} = \frac{\mu_M - r}{v_M}$. By rearranging terms, we obtain Eq. (1.19).

1.9 Appendix 3: Risk and risk aversion

1.9.1 Modeling and approximating risk-premiums

Project value in a representative investor model. We determine the value of the project discussed in Section 1.3.3 in a simple model with a representive investor. This value is defined as the solution V to Eq. (1.28), and we assume that the representative investor has initial wealth $w = 1$, exponential utility, $u(w) = -e^{-\tau w}$ (with τ set equal to 1), and that the random cash flow $\tilde{C}$ is either 1 or $C > 1$ with equal probability. In this example, a closed-form solution is not available for the value of V that satisfies Eq. (1.28).

Figure 1.A1 depicts the numerical solution for V as a function of C. Note that due to risk aversion, V is lower than the expected cash flow, $E(\tilde{C}) = \frac{1}{2}(1 + C)$. An analytical approximation to the value of V based on Eq. (1.31) is $\hat{V} \equiv E(\tilde{C}) - \hat{\Lambda}$, where the approximated premium, $\hat{\Lambda}$, is defined as

$$\Lambda \approx \hat{\Lambda} \equiv \frac{1}{2}\sigma_\epsilon^2 = \frac{1}{8}(C-1)^2,$$

and the second equality follows by a simple calculation. Figure 1A.1 shows that the accuracy of the approximation, $\hat{V}$, deteriorates as the cash flow risk increases, i.e., as the deviation of the cash flow in the good state of nature increases from that in the bad (i.e., one). We now turn to provide the rationale underlying the approximation in Eq. (1.31).

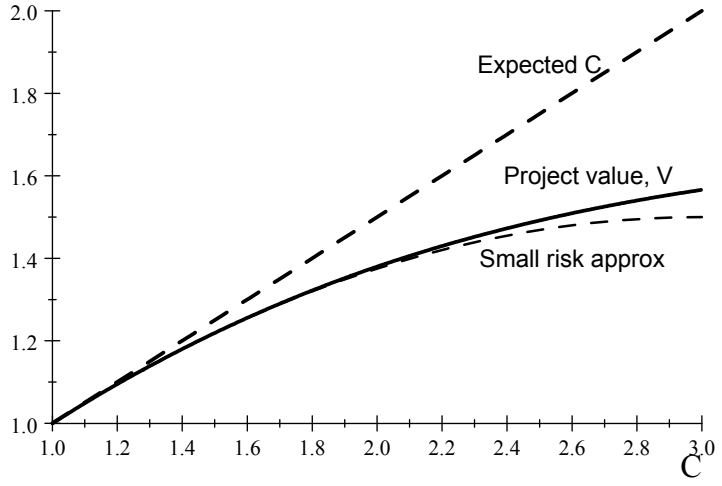


FIGURE 1.A1. The solid line depicts the value of the project, V; the thick dashed line depicts the expected value of the project, $E(\tilde{C}) = \frac{1}{2}(1 + C)$, and the thin dashed line is the "small risk approximation" to V in this example, obtained as $\hat{V} \equiv E(\tilde{C}) - \hat{\Lambda}$, where $\hat{\Lambda} \equiv \frac{1}{8}(C-1)^2$. The true risk premium is the difference $\Lambda = E(\tilde{C}) - V$.

The approximation in Eq. (1.31). Consider a second-order Taylor's approximation of the R.H.S. of the equality in Eq. (1.29) around $\epsilon = 0$,

$$E(u(w + \Lambda + \epsilon)) \approx u(w) + u'(w)\Lambda + \frac{1}{2}u''(w)\sigma_\epsilon^2,$$

where σ_ϵ^2 denotes the variance of ϵ, and where we have disregarded second order terms in Λ. Using this approximation in Eq. (1.29) leaves Eq. (1.31).

The approximation in Eq. (1.33). The rationale behind this approximation is similar to that underlying Eq. (1.31). By a second-order Taylor's approximation of the R.H.S. of the equality in Eq. (1.32) around $\epsilon = 0$,

$$E\left(u\left(w + E(\tilde{x}) + \epsilon\right)\right) \approx u\left(w + E(\tilde{x})\right) + \frac{1}{2}u''\left(w + E(\tilde{x})\right)\sigma_\epsilon^2,$$

and by a first-order approximation of the L.H.S.,

$$u\left(w + E(\tilde{x}) - \Pi\right) \approx u\left(w + E(\tilde{x})\right) - u'\left(w + E(\tilde{x})\right)\Pi.$$

Using these approximations in Eq. (1.32) produces Eq. (1.33).

The approximation in Eq. (1.35). Approximate the R.H.S. of the equality in Eq. (1.34) with a second-order Taylor's expansion around $\tilde{g} = 0$,

$$E\left(u\left(\left(w + E(\tilde{x})\right)\left(1 + \tilde{g}\right)\right)\right) \approx u\left(w + E(\tilde{x})\right) + \frac{1}{2}u''\left(w + E(\tilde{x})\right)\left(w + E(\tilde{x})\right)^2\sigma_g^2,$$

and by a first-order approximation of the L.H.S.,

$$u\left(w + E(\tilde{x})\left(1 - \Pi_r\right)\right) \approx u\left(w + E(\tilde{x})\right) - u'\left(w + E(\tilde{x})\right)\left(w + E(\tilde{x})\right)\Pi_r.$$

Using these approximations in Eq. (1.34) leaves Eq. (1.35).

Certainty equivalents: examples. Consider a risk averse individual who is about to enter a gamble by which he will receive the outcome of rolling a dice. We assume that he has initial wealth $w = 1$ and utility function $u\left(w\right) = \ln w$. The expect payoff from the gamble is $E\left(\tilde{x}\right) = \frac{1}{6}\sum_{j=1}^{6} j = 3.5$, and the variance is $\sigma_\epsilon^2 = 2.9167$. The certainty equivalent is defined as

$$\text{CE}\left(w, \tilde{x}\right) : u(w + \text{CE}\left(w, \tilde{x}\right)) = E\left(u\left(w + \tilde{x}\right)\right) = \frac{1}{6}\sum_{j=1}^{6}\ln\left(1 + j\right) = 1.4209,$$

where the last equality follows by the assumption that $w = 1$. That is, $\text{CE}\left(1, \tilde{x}\right) = 3.1408$. Accordingly, the (insurance) risk-premium is $\Pi = E\left(\tilde{x}\right) - \text{CE}\left(1, \tilde{x}\right) = 0.3592$, whereas its approximation in Eq. (1.33) is

$$\hat{\Pi} = -\frac{1}{2}\frac{u''\left(w + E(\tilde{x})\right)}{u'\left(w + E(\tilde{x})\right)}\sigma_\epsilon^2 = \frac{1}{2}\frac{1}{1 + 3.5}2.9167 = 0.3241.$$

Alternatively, assume the individual has exponential utility $u\left(w\right) = -e^{-\tau w}$, with $\tau = 1$, and initial wealth still equal to 1. The certainty equivalent is now solution to

$$\text{CE}\left(w, \tilde{x}\right) : u(w + \text{CE}\left(w, \tilde{x}\right)) = E\left(u\left(w + \tilde{x}\right)\right) = \frac{1}{6}\sum_{j=1}^{6}(-e^{-(1+j)}) = -3.5594 \times 10^{-2}.$$

That is, $\text{CE}\left(1, \tilde{x}\right) = 2.3356$, such that the risk-premium is now equal to $\Pi = E\left(\tilde{x}\right) - \text{CE}\left(1, \tilde{x}\right) = 3.5 - 2.3356 = 1.1644$, whereas its approximation through Eq. (1.33) is $\hat{\Pi} = 1.4584$.

Note that due to constant absolute risk aversion, certainty equivalents remain the same as the individual's initial wealth changes. It is obviously not the case in the log case: for example, if $w = 0$, we have $\text{CE}\left(0, \tilde{x}\right) = 2.9937$, $\Pi = 0.5063$ and $\hat{\Pi} = 0.41667$, i.e., the risk-premium is inversely related to wealth, as explained in the main text.

1.9.2 Stochastic dominance and mean-preserving spreads

Variance is the notion of risk underlying the classical CAPM and other models in this chapter. But there are situations studied in these lectures in which choices of an expected utility maximizer are well understood while relying on a generalized notion of risk due to Rothschild and Stiglitz (1970, 1971). Moreover, this notion of risk can be utilized to characterize quite simply the relation between asset prices and the volatility of fundamentals described in Section 1.4.3.

We begin with the following definition:

DEFINITION 1.1 (First-order stochastic dominance). *$\tilde{x}_2$ dominates $\tilde{x}_1$ if, for any utility function, i.e. u increasing and concave, we have that $E[u(\tilde{x}_2)] \geq E[u(\tilde{x}_1)]$.*

We have:

THEOREM 1A.1. *The following statements are equivalent: (a) $\tilde{x}_2$ dominates $\tilde{x}_1$, or $E[u(\tilde{x}_2)] \geq E[u(\tilde{x}_1)]$ for every increasing function u; (b) for each $x > 0$, $F_2(x)$ is more likely than $F_1(x)$ to pay more than x, i.e. $F_1(x) \geq F_2(x)$, where $F_i(x)$ denotes the cumulative distribution of x_i.*

PROOF. We prove this result in the case the support is compact, say $[a, b]$. First, we show that $(b) \Rightarrow (a)$. By integrating by parts,

$$E[u(x)] = \int_a^b u(x)dF(x) = u(b) - \int_a^b u'(x)F(x)dx,$$

where we have used the fact that: $F(a) = 0$ and $F(b) = 1$. Therefore,

$$E[u(\tilde{x}_2)] - E[u(\tilde{x}_1)] = \int_a^b u'(x)[F_1(x) - F_2(x)]dx.$$

Next, we show that $(a) \Rightarrow (b)$. Indeed (a) implies that $\int_a^b u(x)(dF_2(x) - dF_1(x)) \geq 0$ for $u(x) = \mathbb{I}_{x \geq y}$ and $y \in (a, b)$, or $0 \leq \int_a^b \mathbb{I}_{x \geq y}(dF_2(x) - dF_1(x)) = \int_y^b (dF_2(x) - dF_1(x)) = F_1(y) - F_2(y)$. ■

There is an alternative characterization of first-order stochastic dominance. Suppose there exists a strictly positive random variable η by which $\tilde{x}_2$ exceeds $\tilde{x}_1$, viz,

$$\eta > 0 : \tilde{x}_2 = \tilde{x}_1 + \eta. \tag{1A.14}$$

Then, we have that $\forall t \in [a, b]$, $F_1(t) \equiv \Pr(\tilde{x}_1 \leq t) = \Pr(\tilde{x}_2 \leq t + \eta) \geq \Pr(\tilde{x}_2 \leq t) \equiv F_2(t)$. That is, $\tilde{x}_2$ dominates $\tilde{x}_1$ if it can be expressed as in Eq. (1A.14). It is quite an intuitive property, but at the same time, it does not insulate the pure component of risk. Instead, we would like to perform the thought experiment to ask every expected utility maximizer to choose between two distributions with the same mean. Consider the following definition of "second order stochastic dominance."

DEFINITION 1.2. (Second-order stochastic dominance). *Let $\tilde{x}_1$ and $\tilde{x}_2$ be two random variables with the same expectation. We say that $\tilde{x}_1$ is more risky than $\tilde{x}_2$ if, for every concave function u, we have that $E[u(\tilde{x}_1)] \leq E[u(\tilde{x}_2)]$.*

Next, consider a case in which the distribution of one variable $\tilde{x}_1$ is obtained from that of another variable $\tilde{x}_2$, as follows: we take weights from the middle part of the density and move them towards the tails, by making sure that the new density has the same mean as the initial. This is equivalent to requiring that condition (b) in the following theorem holds true. We have:

THEOREM 1A.2. *The following statements are equivalent: (a) $\tilde{x}_1$ is more risky than $\tilde{x}_2$; (b) $\tilde{x}_1$ has more weight in the tails than $\tilde{x}_2$, i.e. $\forall t$, $\int_{-\infty}^{t} [F_1(x) - F_2(x)]\, dx \geq 0$.*

PROOF. As for $(a) \Rightarrow (b)$, consider the function, $-b_y(x) = -\max\{y - x, 0\}$. It is increasing and concave and, hence, a candidate utility function. Therefore, it satisfies,

$$\int_a^b (-b_y(x))\, [dF_1(x) - dF_2(x)] \leq 0.$$

That is, using the definition of b_y,

$$\begin{aligned} 0 &\leq \int_a^y (y - x)\, [dF_1(x) - dF_2(x)] \\ &= y\, [F_1(y) - F_2(y)] - \int_a^y x\, [dF_1(x) - dF_2(x)] \\ &= \int_a^y [F_1(x) - F_2(x)]\, dx, \end{aligned}$$

where the last equality follows by an integration by parts. Next we prove that $(b) \Rightarrow (a)$. We have:

$$\begin{aligned} E[u(\tilde{x}_1)] - E[u(\tilde{x}_2)] &= \int_a^b u(x)\, [f_1(x) - f_2(x)]\, dx \\ &= u(x)\, [F_1(x) - F_2(x)]\big|_a^b - \int_a^b u'(x)\, [F_1(x) - F_2(x)]\, dx \\ &= -\int_a^b u'(x)\, [F_1(x) - F_2(x)]\, dx \\ &= -\left[u'(x)\, \left[\bar{F}_1(x) - \bar{F}_2(x)\right]\big|_a^b - \int_a^b u''(x)\, \left[\bar{F}_1(x) - \bar{F}_2(x)\right] dx \right] \\ &= \int_a^b u''(x)\, \left[\bar{F}_1(x) - \bar{F}_2(x)\right] dx - u'(b)\, \left[\bar{F}_1(b) - \bar{F}_2(b)\right] \\ &= \int_a^b u''(x)\, \left[\bar{F}_1(x) - \bar{F}_2(x)\right] dx, \end{aligned}$$

where $\bar{F}_i(x) = \int_a^x F_i(u) du$. The last equality follows, because, by integrating by parts,

$$-\int_a^b [F_1(x) - F_2(x)]\, dx = \int_a^b x\, [dF_1(x) - dF_2(x)] = 0,$$

where the last equality follows by the assumption that F_1 and F_2 have the same mean. Now, by $u'' < 0$, and $\bar{F}_1(x) > \bar{F}_2(x)$, the previous relation implies that $E[u(\tilde{x}_1)] < E[u(\tilde{x}_2)]$, i.e., $\tilde{x}_1$ is more risky than $\tilde{x}_2$. ■

Finally, we explain a link between the previous notions of risk and that of "mean-preserving spreads" utilized in Section 1.4.4 of this chapter.

MEAN-PRESERVING SPREADS. We can now consider random variables that "add up risk" without affecting the mean: suppose that there exists a random variable $\epsilon : \tilde{x}_1$ has the same distribution as $\tilde{x}_2 + \epsilon$, and $E(\epsilon | \tilde{x}_2 = x_2) = 0$. We can think of an experiment in which after receiving a random payoff $\tilde{x}_2$, another random payoff could be offered to us, which has conditional expectation zero, thereby adding randomness without boosting the overall expected return. That is, $\tilde{x}_1$ is a *mean-preserving spread* of

$\tilde{x}_2$. A variable that is a mean-preserving spread of another displays the properties stated in Theorem 1A.2. Indeed,

$$\begin{aligned} E\left[u\left(\tilde{x}_1\right)\right] &= E\left[u\left(\tilde{x}_2+\epsilon\right)\right] \\ &= E\left[E\left(u\left(\tilde{x}_2+\epsilon\right)\middle|\,\tilde{x}_2=x_2\right)\right] \\ &\leq E\left[u\left(E\left(\tilde{x}_2+\epsilon\middle|\,\tilde{x}_2=x_2\right)\right)\right] \\ &= E\left[u\left(E\left(\tilde{x}_2\middle|\,\tilde{x}_2=x_2\right)\right)\right] \\ &= E\left[u\left(\tilde{x}_2\right)\right], \end{aligned}$$

where the inequality follows by concavity of u and by Jensen's inequality.

1.10 Appendix 4: Money demand and liquidity traps

We provide the parametric example mentioned in the main text, in which agents make dichotomic choices while either holding money or bonds according to their expectations regarding future interest rates. Consider a perpetuity with fixed coupons equal to 1, priced at $b_0 = \sum_{h=1}^{\infty} 1 \cdot (1+i_0)^{-h} = i_0^{-1}$, where i_0 is the current nominal rate, assumed to be flat over all maturities. We assume that there is a continuum of agents on $[0,1]$ and, crucially, that each agent $j \in [0,1]$ "expects" that within a certain reference period, i_0 will converge to a "normal" rate, say $i_e(j) \equiv b_e^{-1}(j)$, with probability one, for all maturities. From the perspective of this agent, the gain from holding the perpetuity over this period is

$$g_j(i_0) \equiv \frac{b_e(j) - b_0 + 1}{b_0} = \frac{i_0}{i_e(j)} - 1 + i_0.$$

It is easy to see that there exists a value of i_0 for each j, such that $g_j(i_0) = 0$, given by

$$\hat{i}_0(j) \equiv \frac{i_e(j)}{1 + i_e(j)}.$$

It is a critical rate, in that agent j would only hold money if $\hat{i}_0(j) > i_0$, would only invest in bonds if $\hat{i}_0(j) < i_0$ and, finally, would be indifferent about cash or bonds should $\hat{i}_0(j) = i_0$.

How do open market operations affect interest rates in this economy? Re-shuffle the agents such that $\hat{i}_0(j)$ is monotonically decreasing, and for a given i_0, let $\bar{j} : \hat{i}_0(\bar{j}) = i_0$, such that investors "on the left" of $\bar{j}$ (i.e., $j < \bar{j}$) only hold money, and investors "on the right" of $\bar{j}$ (i.e., $j > \bar{j}$) only hold bonds. Suppose, now, that the central bank enters the market as buyer of bonds. Given the current i_0, the investors "on the right" are obviously happy with *not* tendering any of their bond holdings to the bank, because they all anticipate rates to lower to an extent of generating profits to them.

To incentivize bond-holders to sell against money, the central bank must be willing to buy these bonds at a price higher than b_0, so as to allow the bond-holders to break-even given their expectations. That is, the nominal interest rate needs to lower. For example, if the central bank bids up the bond price at $b_0' \equiv (i_0 - \epsilon)^{-1}$, for a small $\epsilon > 0$, then, the break-even condition $g_j(i_0 - \epsilon) = 0$ implies that the investors "on the right" are those $j : j > \bar{j}' > \bar{j}$, where $\hat{i}_0(\bar{j}') = i_0 - \epsilon$. In other words, a mass of investors equal to $\bar{j}' - \bar{j}$ is willing to sell the bonds at a price b_0'. So at the aggregate level, money demand decreases with the nominal interest rates.

When does this process end? Once the current interest rate reaches $i_0 = \xi \equiv \hat{i}_0(1)$, there are no agents who are currently holding any bond apart from the "marginal" investor $j = 1$, as all agents now believe that interest rates can only rise within the reference period and lead to a sure loss according to their perspective. This situation is known as a *liquidity trap*: when $i_0 = \xi$, changes in money supply, be they positive (and of any size) or negative (but with small size), do not affect interest rates. Indeed, even while offering a higher purchasing price, there are no investors who could sell, except of course the marginal investor $j = 1$. So if the central bank increases money supply by purchasing the bonds, this marginal investor would be perfectly ready to accept this new money and tender the bonds, as he is obviously indifferent between investing in bonds or hoarding money. Likewise, if the central bank decreases money supply through a bonds sale, the marginal investor would buy these bonds.

The important point of Keynesian theory is that money demand is negatively sloped at an aggregate level. To illustrate with an analytical example, assume that the distribution of $i_e(j)$ is uniform, i.e.,

$$i_e(j) = \bar{\xi} - (\bar{\xi} - \underline{\xi})j, \quad j \in [0,1],$$

for some two constants $\underline{\xi}$ and $\bar{\xi} > \underline{\xi}$. Then,

$$\hat{i}_0(j) = \frac{\bar{\xi} - (\bar{\xi} - \underline{\xi})j}{1 + \bar{\xi} - (\bar{\xi} - \underline{\xi})j}.$$

The j-th agent money demand, $m^d(j)$ say, is dichotomic, in that

$$m^d(j) = \begin{cases} 1, & \text{if } \hat{\imath}_0(j) > i_0 \\ 0, & \text{otherwise} \end{cases}$$

Yet aggregate money demand is

$$M^d \equiv \int_0^1 m^d(j)\, dj = \int_0^1 \mathbb{I}_{\{\hat{\imath}_0(j) > i_0\}} dj = \int_{j:\hat{\imath}_0(j) > i_0} dj = \frac{\bar{\xi} - \left(1 + \bar{\xi}\right) i_0}{(\bar{\xi} - \underline{\xi})(1 - i_0)},$$

where $\mathbb{I}_{\cdot}$ is the indicator function. Note that M^d is decreasing in i_0; moreover, $M_d > 0$ provided $i_0 < \frac{\bar{\xi}}{1+\bar{\xi}}$.

References

Abel, A.B. (1988): "Stock Prices under Time-Varying Dividend Risk: An Exact Solution in an Infinite-Horizon General Equilibrium Model." *Journal of Monetary Economics* 22, 375-393.

Arrow, K.J. (1965): "The Theory of Risk Aversion." In *Aspects of the Theory of Risk Bearing*. Helsinki: Yrjö Jahnssonin Säätiö.

Banz, R.W. (1981): "The Relationship Between Return and Market Value of Common Stocks." *Journal of Financial Economics* 9, 3-18.

Barsky, R.B. (1989): "Why Don't the Prices of Stocks and Bonds Move Together?" *American Economic Review* 79, 1132-1145.

Black, F. (1972): "Capital Market Equilibrium with Restricted Borrowing." *Journal of Business* 45, 444-454.

Campbell, J.Y. and L.M. Viceira (2002): *Strategic Asset Allocation*. Oxford: Oxford University Press.

Carhart, M. (1997): "On Persistence of Mutual Fund Performance." *Journal of Finance* 52, 57-82.

Chen, N-F., R. Roll and S.A. Ross (1986): "Economic Forces and the Stock Market." *Journal of Business* 59, 383-403.

Connor, G. (1984): "A Unified Beta Pricing Theory." *Journal of Economic Theory* 34, 13-31.

Fama, E.F. and J.D. MacBeth (1973): "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 38, 607-636.

Fama, E. F. and K. R. French (1992): "The Cross-Section of Expected Stock Returns." *Journal of Finance* 47, 427-465.

Fama, E. F. and K. R. French (1993): "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33, 3-56.

Frazzini, A., D. Kabiller and L. Pedersen (2012): "Buffet's Alpha." Working paper.

Frazzini, A. and L. Pedersen (2012): "Betting Against Beta." Working paper.

Grossman, S.J. and J.E. Stiglitz (1980): "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70, 393-408.

Huang, C-f. and R.H. Litzenberger (1988): *Foundations for Financial Economics*. New York: North-Holland.

Huberman, G. (1983): "A Simplified Approach to Arbitrage Pricing Theory." *Journal of Economic Theory* 28, 1983-1991.

Jegadeesh, N. and S. Titman (1993): "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *Journal of Finance* 48, 65-91.

Keynes, J. M. (1936): *The General Theory of Employment, Interest and Money*. London: Palgrave Macmillan.

Kyle, A.S. (1985): "Continuous Auctions and Insider Trading." *Econometrica* 53, 1335-1335.

Lintner, J. (1965): "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *Review of Economics and Statistics* 47, 13-37.

Malkiel, B. (1979): "The Capital Formation Problem in the United States." *Journal of Finance* 34, 291-306.

Markovitz, H. (1952): "Portfolio Selection." *Journal of Finance* 7, 77-91.

Mossin, J. (1966): "Equilibrium in a Capital Asset Market." *Econometrica* 34, 768-783.

Pindyck, R. (1984): "Risk, Inflation and the Stock Market." *American Economic Review* 74, 335-351.

Poterba, J. and L. Summers (1985): "The Persistence of Volatility and Stock Market Fluctuations." *American Economic Review* 75, 1142-1151.

Pratt, J. W. (1964): "Risk Aversion in the Small and in the Large." *Econometrica* 32, 122-136.

Roll, R. (1977): "A Critique of the Asset Pricing Theory's Tests Part I: On Past and Potential Testability of the Theory." *Journal of Financial Economics* 4, 129-176.

Rosenberg, B. K. Reid and R. Lanstein (1985): "Persuasive Evidence of Market Inefficiency." *Journal of Portfolio Management* 11, 9-17.

Ross, S. (1976): "Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory* 13, 341-360.

Rothschild, M. and J. Stiglitz (1970): "Increasing Risk: I. A Definition." *Journal of Economic Theory* 2, 225-243.

Rothschild, M. and J. Stiglitz (1971): "Increasing Risk: II. Its Economic Consequences." *Journal of Economic Theory* 5, 66-84.

Sharpe, W. F. (1964): "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance* 19, 425-442.

Stattman, D. (1980): "Book Values and Stock Returns." *The Chicago MBA: A Journal of Selected Papers* 4, 25-45.

Tobin, J. (1958): "Liquidity Preference as Behavior Towards Risk." *Review of Economic Studies* 25, 65-86.

2
Arbitrage, equilibrium and pricing

2.1 Introduction

This chapter develops asset pricing implications arising while requiring that markets are free from arbitrage. What are gambles and what are financial securities? Typically, a gamble has a payoff linked to a risk that is "not traded" and that may not necessarily relate to the activity of entities such as private corporations, households, or governments; the notion of "traded risk" is explained below. Equally typically, once entered into a gamble, the gambler encounters difficulties while attempting to exit from it—there aren't "liquid" markets for a gamble. Finally, the value of a gamble is typically determined by supply and demand, thereby resting on factors such as risk-aversion or the bargaining power of both sides of the trade. Some gambles (say those we see in a casino) may be "repeated," so to speak, in which case the Law of Large Numbers may provide us with a rough guidance to their pricing; the guidance is only "rough" because supply and demand or bargaining power still play a critical role.

Financial securities work differently. Firstly, they are issued to finance (or hedge risks arising from) the activities of private corporations, households, and governments. Secondly, financial securities differ from gambles due to how they are priced. Granted, in practice, some financial securities are traded in relatively illiquid markets (see Chapters 4 and 10). But one distinctive feature of these securities is that, at least in theory, there are markets in which they could be priced while only relying on absence of arbitrage. In contrast, there aren't idealized markets in which gambles could be priced in this way—for otherwise we would perhaps begin thinking of them as financial securities. These idealized markets did constitute a useful starting point in financial economics.

How do these idealized markets work? Consider the case in which the value of a security could be "replicated" via a dedicated portfolio of other securities. In this case, the security value equals that of the replicating portfolio, regardless of supply and demand and any Law of Large Numbers. The reason is that two portfolios delivering the same payoff must be worth the same, for otherwise an arbitrage opportunity would arise—the possibility to trade exisiting securities and make riskless profits. We say that in this case, security evaluation is "preference-free," and that the risk underlying the security value is *traded*. It is, of course, an idealized situation, but also the cornerstone in modern asset evaluation.

The purest kind of securities we may imagine is one that only pays off a fixed amount of numéraire in a pre-specified state of the world: say £1 in the given state and zero otherwise. These assets are known as Arrow-Debreu securities, in recognition of the founders of general equilibrium theory with uncertainty. Arrow-Debreu securities are conceptually very useful, as the knowledge of their prices can be utilized to price any other asset. Not surprisingly, these idealized securities are also a useful tool in the market practice of asset pricing, as explained in Part III. They link to what we commonly refer to as risk-neutral probabilities—a concept introduced in the previous chapter and developed in detail in this and subsequent chapters.

Needless to mention, many of the previous optimistic conclusions rely on a number of assumptions such as the agents' symmetric information about the assets fundamentals, market completeness (i.e., the existence of an Arrow-Debreu security for each state of the world), absence of additional capital market frictions or, finally, perfect competition in the goods markets. We shall deal with capital market imperfections in Chapters 4 and 5 and with information problems and other frictions in Chapters 9 and 10. These imperfections allow to address many realistic features although this chapter already contains an introduction to the theory in the context of incomplete markets. In these interesting cases, financial securities could not be replicated—they would not be "preference free." Indeed, there certainly are securities or complex products that, in practice, cannot be replicated, linking as they do to new and not traded risks, that is, risks to which existing securities only partially link (say a future on a volatility index; see Chapter 11). However, in some cases, it may be that for all intents and purposes, some securities could be bunched into a portfolio that mimics sufficiently a given security. We could then design worst-case and best-case scenarios and evaluate the security in this way. Naturally, an alternative to all this is financial innovation—the process of creation of new securities that have the potential to fill in the initially incomplete markets.

This chapter aims to analyze the functioning of these idealized markets in a world with uncertainty—the basis for more advanced analyses in further chapters. We start from very far: we review the classical general equilibrium model in a context without uncertainty. This model has profound implications for the society as a whole, whereby equilibrium allocations are optimal according to Pareto's criterion. Financial securities play a critical role once these economies are subject to uncertainty. When the number of these securities is large enough, resources allocation is still Pareto optimal in these economies, just as it is in the classic general equilibrium model. Intuitively, a number of sufficiently diverse financial securities has the potential to deliver the payoffs we need in each future contingency of the world, thereby making markets function as if we were in a static world.

In fact, even when markets are incomplete, we may still define shadow prices of Arrow-Debreu security, which could be used to evaluate any other asset. The challenge in this case is that when markets are incomplete, requiring absence of arbitrage does not lead to an unique set of Arrow-Debreu prices (as in the complete markets case). Further assumptions are needed to determine these shadow prices. For example, in an economy in which agents optimize their consumption choices, Arrow-Debreu prices are pinned down in equilibrium. Indeed, an important task of this chapter is to link Arrow-Debreu prices to optimal consumption in general equilibrium. It is natural. As financial economists, we would naturally like to understand asset prices from the perspective of an economy where households allocate their endowments across consumption and savings. Our households' objective is to maximize utility of consumption in an uncertain world. This chapter provides introductions to these problems in the context of two-period economies, but many of its insights lead to asset pricing equations that are the logical antecedents to the Euler's equations arising in multiperiod economies dealt with in further chapters.

The chapter is organized as follows. The next section contains a succinct description of the static general equilibrium model and its properties in a static context, and abstracts from decisions taken within the production sphere of the economy. Production-based economies are studied in more detail in Chapter 3 as well as in Part II (e.g., Chapter 9). Section 2.3 illustrates the role financial securities can play in economies with uncertainties, and the very first examples of the meaning and use of Arrow-Debreu securities, and their relations to the risk-neutral probability. Sections 2.4 and 2.5 provide theory, based on the extension of the general equilibrium model of Section 2.2, and include uncertainty. Section 2.4 provides the first important example of how we can evaluate assets solely based on requiring absence of arbitrage opportunities. Section 2.5 develops a more systematic treatment of markets in absence of arbitrage, in particular the implications on asset prices in both complete and incomplete markets. Naturally, absence of arbitrage does not imply equilibrium, although the converse is true, as explained in Section 2.6. Section 2.6 then relates Arrow-Debreu security prices to risk-neutral probabilities, in both complete and incomplete markets, and provides discussion on topics such as the role of financial markets as vehicles risk-sharing for a society, or financial innovation. Section 2.7 provides a very first introduction to the theory of the Consumption-CAPM as well as its predictions on the equity premium. Section 2.8 provides a framework to think about budget constraints in infinite horizon markets. Finally, Section 2.9 develops a few more topics about the theme of incomplete markets, and the appendixes contain material omitted from the main text.

2.2 The static general equilibrium in a nutshell

We consider an economy with n agents and m commodities. Let w_{ij} denote the amount of the i-th commodity the j-th agent is endowed with, and let $w^j = [w_{1j}, \cdots, w_{mj}]$. Let the price vector be $p = [p_1, \cdots, p_m]$, where p_i is the price of the i-th commodity. Let $w_i = \sum_{j=1}^{n} w_{ij}$ be the total endowment of the i-th commodity in the economy, and $W = [w_1, \cdots, w_m]$ the corresponding endowments bundle in the economy.

The j-th agent has utility function $u_j(c_{1j}, \cdots, c_{mj})$, where $(c_{ij})_{i=1}^{m}$ denotes his consumption bundle. We assume the following standard conditions for the utility functions u_j:

Assumption 2.1 (Preferences). *The utility functions u_j satisfy the following properties: (i) Monotonicity; (ii) Continuity; and (iii) Quasi-concavity*: $u_j(x) \geq u_j(y)$, *and* $\forall \alpha \in (0,1)$, $u_j(\alpha x + (1-\alpha)y) > u_j(y)$ *or,* $\frac{\partial u_j}{\partial c_{ij}}(c_{1j}, \cdots, c_{mj}) \geq 0$ *and* $\frac{\partial^2 u_j}{\partial c_{ij}^2}(c_{1j}, \cdots, c_{mj}) \leq 0$.

Let $B_j(p_1, \cdots, p_m) = \{(c_{1j}, \cdots, c_{mj}) : \sum_{i=1}^{m} p_i c_{ij} \leq \sum_{i=1}^{m} p_i w_{ij} \equiv R_j\}$, a bounded, closed and convex set, hence a convex set. Each agent maximizes his utility function subject to the budget constraint:

$$\max_{\{c_{ij}\}} u_j(c_{1j}, \cdots, c_{mj}) \quad \text{subject to } (c_{1j}, \cdots, c_{mj}) \in B_j(p_1, \cdots, p_m). \qquad [2.P1]$$

This problem has certainly a solution, for B_j is compact set and by Assumption 2.1, u_j is continuous, and a continuous function attains its maximum on a compact set. Moreover, the Appendix shows that this maximum is unique.

The first order conditions to [2.P1] are, for each agent j,

$$\begin{cases} \frac{\frac{\partial u_j}{\partial c_{1j}}}{p_1} = \frac{\frac{\partial u_j}{\partial c_{2j}}}{p_2} = \cdots = \frac{\frac{\partial u_j}{\partial c_{mj}}}{p_m} \\ \sum_{i=1}^{m} p_i c_{ij} = \sum_{i=1}^{m} p_i w_{ij} \end{cases} \tag{2.1}$$

These conditions form a system of m equations with m unknowns. Let us denote the solution to this system with $[\hat{c}_{1j}(p, w^j), \cdots, \hat{c}_{mj}(p, w^j)]$. The total demand for the i-th commodity is,

$$\hat{c}_i(p, w) = \sum_{j=1}^{n} \hat{c}_{ij}(p, w^j), \quad i = 1, \cdots, m.$$

We emphasize the economy we consider in this chapter is one that completely abstracts from production. Here, prices are the key determinants of how resources are allocated in the end. The perspective is, of course, radically different from that taken by the Classical school (Ricardo, Marx and Sraffa), for which prices and resources allocation cannot be disentangled from the production side of the economy. In the next chapter and in Part II, we consider the asset pricing implications of production, following the Neoclassical perspective.

2.2.1 Walras' Law

Let us plug the demand functions of the j-th agent into the constraint of [2.P1], to obtain,

$$\forall p, \quad 0 = \sum_{i=1}^{m} p_i \left(\hat{c}_{ij}(p, w^j) - w_{ij} \right). \tag{2.2}$$

Next, define the total excess demand for the i-th commodity as $e_i(p, w) \equiv \hat{c}_i(p, w) - w_i$. By aggregating the budget constraint across all the agents,

$$\forall p, \quad 0 = \sum_{j=1}^{n} \sum_{i=1}^{m} p_i \left(\hat{c}_{ij}(p, w^j) - w_{ij} \right) = \sum_{i=1}^{m} p_i e_i(p, w).$$

The previous equality is the celebrated *Walras' law.*

Next, multiply p by $\lambda \in \mathbb{R}_{++}$. Since the constraint to [2.P1] does not change, the excess demand functions are the same, for each value of λ. In other words, *the excess demand functions are homogeneous of degree zero in the prices,* or $e_i(\lambda p, w) = e_i(p, w)$, $i = 1, \cdots, m$. This property of the excess demand functions is also referred to as *absence of monetary illusion.*

2.2.2 Competitive equilibrium

A *competitive equilibrium* is a vector $\bar{p}$ in $\mathbb{R}^m_+$ such that $e_i(\bar{p}, w) \leq 0$ for all $i = 1, \cdots, m$, with at least one component of $\bar{p}$ being strictly positive. Furthermore, if there exists a $j : e_j(\bar{p}, w) < 0$, then $\bar{p}_j = 0$.

2.2.2.1 Back to Walras' law

Walras' law holds by the mere aggregation of the agents' constraints. But the agents' constraints are accounting identities. In particular, Walras' law holds for any price vector and, a fortiori, it holds for the equilibrium price vector,

$$0 = \sum_{i=1}^{m} \bar{p}_i e_i(\bar{p}, w) = \sum_{i=1}^{m-1} \bar{p}_i e_i(\bar{p}, w) + \bar{p}_m e_m(\bar{p}, w). \tag{2.3}$$

Now suppose that the first $m-1$ markets are in equilibrium, or $e_i(\bar{p}, w) \leq 0$, for $i = 1, \cdots, m-1$. By the definition of an equilibrium, we have that $\text{sign}\left(e_i(\bar{p}, w)\right) \bar{p}_i = 0$. Therefore, by Eq. (2.3), we conclude that if $m-1$ *markets are in equilibrium, then, the remaining market is also in equilibrium.*

2.2.2.2 The notion of numéraire

The excess demand functions are homogeneous of degree zero. Walras' law implies that if $m-1$ markets are in equilibrium, then, the m-th remaining market is also in equilibrium. We wish to link these two results. A first remark is that by Walras' law, the equations that define a competitive equilibrium are not independent. Once $m-1$ of these equations are satisfied, the m-th remaining equation is also satisfied. In other words, there are $m-1$ independent relations and m unknowns in the equations that define a competitive equilibrium. So, there exists an infinity of solutions.

Suppose, then, that we choose the m-th price to be a sort of exogeneous datum. The result is that we obtain a system of $m-1$ equations with $m-1$ unknowns. Provided it exists, such a solution is a function f of the m-th price, $\bar{p}_i = f_i(\bar{p}_m)$, $i = 1, \cdots, m-1$. Then, we may refer to the m-th commodity as the *numéraire*. In other words, general equilibrium can only determine a structure of *relative* prices. The scale of these relative prices depends on the price level of the numéraire. It is easily checked that if the functions f_i are homogeneous of degree one, multiplying p_m by a strictly positive number λ does not change the relative price structure. Indeed, by the equilibrium condition, for all $i = 1, \cdots, m$,

$$\begin{aligned} 0 &\geq e_i\left(\bar{p}_1, \bar{p}_2, \cdots, \lambda\bar{p}_m, w\right) = e_i\left(f_1(\lambda\bar{p}_m), f_2(\lambda\bar{p}_m), \cdots, \lambda\bar{p}_m, w\right) \\ &= e_i\left(\lambda\bar{p}_1, \lambda\bar{p}_2 \cdots, \lambda\bar{p}_m, w\right) = e_i\left(\bar{p}_1, \bar{p}_2 \cdots, \bar{p}_m, w\right), \end{aligned}$$

where the second equality is due to the homogeneity property of the functions f_i, and the last equality holds because the excess demand functions e_i are homogeneous of degree zero. In particular, by defining relative prices as $\hat{p}_j = p_j / p_m$, one has that $p_j = \hat{p}_j \cdot p_m$ is a function that is homogeneous of degree one. In other words, if $\lambda \equiv \bar{p}_m^{-1}$, then,

$$0 \geq e_i\left(\bar{p}_1, \cdots, \bar{p}_m, w\right) = e_i\left(\lambda\bar{p}_1, \cdots, \lambda\bar{p}_m, w\right) \equiv e_i\left(\frac{\bar{p}_1}{\bar{p}_m}, \cdots, 1, w\right).$$

2.2.3 Optimality

Let $c^j = (c_{1j}, \cdots, c_{mj})$ be the allocation to agent j, $j = 1, \cdots, n$. The following definition is the well-known concept of a desirable resource allocation within a society, according to Pareto.

DEFINITION 2.2 (Pareto optimum). *An allocation $\bar{c} = (\bar{c}^1, \cdots, \bar{c}^n)$ is a Pareto optimum if it is feasible, $\sum_{j=1}^{n} (\bar{c}^j - w^j) \leq 0$, and if there are no other feasible allocations $c = (c^1, \cdots, c^n)$ such that $u_j(c^j) \geq u_j(\bar{c}^j)$, $j = 1, \cdots, n$, with one strict inequality for at least one agent.*

We have the following fundamental result:

THEOREM 2.3 (First welfare theorem). *Every competitive equilibrium is a Pareto optimum.*

PROOF. Let us suppose on the contrary that $\bar{c}$ is an equilibrium but not a Pareto optimum. Then, there exists a $c : u_{j^*}(c^{j^*}) > u_{j^*}(\bar{c}^{j^*})$, for some j^*. Because $\bar{c}^{j^*}$ is optimal for agent j^*, $c^{j^*} \notin B_j(\bar{p})$, or $\bar{p}c^{j^*} > \bar{p}w_j$ and, by aggregating: $\bar{p}\sum_{j=1}^n c^j > \bar{p}\sum_{j=1}^n w^j$, which is unfeasible. It follows that c can not be an equilibrium. ‖

Next, we show that any Pareto optimal allocation can be "decentralized." That is, corresponding to a given Pareto optimum $\bar{c}$, there exist ways of redistributing endowments around, and a price vector $\bar{p} : \bar{p}\bar{c} = \bar{p}w$, which is an equilibrium for the initial set of resources.

THEOREM 2.4 (Second welfare theorem). *Every Pareto optimum can be decentralized.*

PROOF. In the appendix.

The previous theorem can be interpreted as one that supports an *equilibrium with transfer payments*. For any given Pareto optimum $\bar{c}^j$, a social planner can always give $\bar{p}w^j$ to each agent (with $\bar{p}\bar{c}^j = \bar{p}w^j$, where w^j is chosen by the planner), and agents choose $\bar{c}^j$. Figure 2.1 illustratres such a decentralization procedure within the Edgeworth's box. Suppose that the objective is to achieve $\bar{c}$. Given an initial allocation w chosen by the planner, each agent is given $\bar{p}w^j$. Under *laissez faire*, $\bar{c}$ will obtain. In other words, agents are given a constraint of the form $pc^j = \bar{p}w^j$. If w^j and $\bar{p}$ are chosen so as to induce each agent to choose $\bar{c}^j$, then $\bar{p}$ is a supporting equilibrium price. In this case, the marginal rates of substitutions are identical, as established by the following celebrated result:

THEOREM 2.5 (Characterization of Pareto optima: I). *A feasible allocation $\bar{c} = (\bar{c}^1, \cdots, \bar{c}^n)$ is a Pareto optimum if and only if there exists a $\tilde{\phi} \in \mathbb{R}^{m-1}_{++}$ such that*

$$\tilde{\bigtriangledown} u_j = \tilde{\phi},\ j = 1, \cdots, n, \quad \textit{where } \tilde{\bigtriangledown} u_j \equiv \left(\frac{\frac{\partial u_j}{\partial c_{2j}}}{\frac{\partial u_j}{\partial c_{1j}}}, \cdots, \frac{\frac{\partial u_j}{\partial c_{mj}}}{\frac{\partial u_j}{\partial c_{1j}}} \right). \tag{2.4}$$

PROOF. A Pareto optimum satisfies:

$$\bar{c} \in \arg\max_{c \in \mathbb{R}^{m \cdot n}_+} u_1\left(c^1\right) \quad \text{subject to } \begin{cases} u_j(c^j) \geq \bar{u}_j,\ j = 2, \cdots, n & (\lambda_j,\ j = 2, \cdots, n) \\ \sum_{j=1}^n (c^j - w^j) \leq 0 & (\phi_i,\ i = 1, \cdots, m) \end{cases}$$

The Lagrangian function associated with this program is

$$L = u_1(c^1) + \sum_{j=2}^n \lambda_j \left(u_j(c^j) - \bar{u}_j \right) - \sum_{i=1}^m \phi_i \sum_{j=1}^n \left(c_{ij} - w_{ij} \right),$$

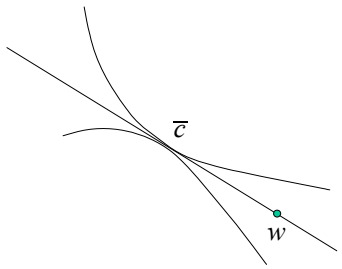


FIGURE 2.1. Decentralizing a Pareto optimum

and the first order conditions are

$$\begin{cases} \dfrac{\partial u_1}{\partial c_{11}} = \phi_1 \\ \cdots \\ \dfrac{\partial u_1}{\partial c_{m1}} = \phi_m \end{cases} \quad \text{and for } j = 2, \cdots, n, \quad \begin{cases} \lambda_j \dfrac{\partial u_j}{\partial c_{1j}} = \phi_1 \\ \cdots \\ \lambda_j \dfrac{\partial u_j}{\partial c_{mj}} = \phi_m \end{cases}$$

In each of the previous two systems, we divide each equation by the the first, obtaining exactly Eq. (2.4), with $\tilde{\phi} = \left(\frac{\phi_2}{\phi_1}, \cdots, \frac{\phi_m}{\phi_1}\right)$. The converse is straightforward. ‖

There is a simple and appealing interpretation of the Kuhn-Tucker multipliers ϕ of the constraints of Theorem 2.5. Note that by Eq. (2.1), in the competitive equilibrium,

$$\tilde{\nabla} u_j = \tilde{p} \equiv \left(\frac{p_2}{p_1}, \cdots, \frac{p_m}{p_1}\right).$$

But because a competitive equilibrium is also a Pareto optimum, then, by Theorem 2.5,

$$\tilde{\nabla} u_j = \tilde{\phi} \equiv \left(\frac{\phi_2}{\phi_1}, \cdots, \frac{\phi_m}{\phi_1}\right).$$

Hence, $\tilde{\phi}$ represents the vector of relative, shadow prices arising within the centralized allocation process.

We provide a further characterization of Pareto optimal allocations.

THEOREM 2.6 (Characterization of Pareto optima: II). *A feasible allocation $\bar{c} = (\bar{c}^1, \cdots, \bar{c}^n)$ is a Pareto optimum if and only if there exists $\ell > 0$ such that $\bar{c}$ is solution to the following program:*

$$u(w, \ell) = \max_{c^1, \cdots, c^n} \sum_{j=1}^{n} \ell_j u_j\left(c^j\right) \quad \text{subject to } \sum_{j=1}^{n} c^j \leq w \quad (\psi_j,\, j = 1, \cdots, m) \qquad [2.P2]$$

PROOF. The if part is simple and at the same time instructive. Let us solve the program in [2.P2]. The Lagrangian is,

$$L = \sum_{j=1}^{n} \ell_j u_j\left(c^j\right) - \sum_{i=1}^{m} \psi_i \sum_{j=1}^{n} \left(c_{ij} - w_{ij}\right),$$

and the first order conditions are, for $j = 1, \cdots, n$,

$$\ell_j \nabla u_j = \psi \equiv (\psi_1, \cdots, \psi_m)^\top, \quad \nabla u_j \equiv \left(\frac{\partial u_j}{\partial c_{1j}}, \cdots, \frac{\partial u_j}{\partial c_{mj}} \right)^\top. \tag{2.5}$$

That is, $\tilde{\nabla} u_j$ equals the same vector of constants for all the agents, just as in Theorem 2.5. The converse to this theorem follows by an application of the usual separating theorem, as in Duffie (2001, Chapter 1). ‖

Note, if $\ell_1 = 1$ and $\ell_j = \lambda_j$ for $j = 2, \cdots, n$, then, $\psi_i = \phi_i$ $(i = 1, \cdots, m)$ and so the first order conditions in Theorem 2.5 and 2.6 would lead to the same allocation. More generally, we have:

THEOREM 2.7 (Centralization of competitive equilibrium through Pareto weightings). *The outcome of any competitive equilibrium can be obtained, through a central planner who maximizes the program in [2.P2], with system of social weights equal to $\ell_j = 1/\kappa_j$, where κ_j is the marginal utility of income for agent j.*

So agents with high marginal utility of income for a given price vector, will receive little social weight in the centralized planner allocation procedure. This result is particularly useful when it comes to study financial markets in economies with heterogeneous agents. Theorem 2.7 is also a point of reference, where to move from, when it comes to study asset prices in a world of incomplete markets. Chapter 9 contains several instances relying on these old but useful insights.

PROOF OF THEOREM 2.7. In the competitive equilibrium,

$$\nabla u_j = \kappa_j p, \quad p \equiv (p_1, \cdots, p_m), \tag{2.6}$$

where κ_j are the Lagrange multipliers for the agents budget constraint, so that κ_j is the agent j marginal utility of income:

$$\kappa_j = \frac{\partial}{\partial m_j} u_j \left(\hat{c}_{1j}(p, w_{1j}, \cdots, w_{mj}), \cdots, \hat{c}_{mj}(p, w_{1j}, \cdots, w_{mj}) \right), \quad m_j \equiv \sum_{i=1}^{m} p_i w_{ij}.$$

By comparing the competitive equilibrium solution in Eq. (2.6) with the Pareto optimality property of the equilibrium in Eq. (2.5), we deduce that, a competitive equilibrium $(\bar{c}, p)$ can be implemented, by a social planner acting as in Theorem 2.6, when $\ell_j = 1/\kappa_j$. Then, it also follows that, necessarily, $\psi = p$, by the aggregate resource constraint, $\sum_{j=1}^{n} c^j \leq w$, which, intuitively, has to hold both in the competitive and the centralized economy. Indeed, we have:

$$w_i = \sum_{j=1}^{n} \hat{f}_{ij}(\kappa_j p), \quad w_i = \sum_{j=1}^{m} f_{ij}(\kappa_j \psi), \quad i = 1, \cdots, m, \tag{2.7}$$

where $\hat{f}_{ij}$ and f_{ij} are the inverse functions for consumption, as implied by the private allocation in Eq. (2.6) and the social, in Eq. (2.5). The first of Eqs. (2.7) determines the general equilibrium price vector, p_* say. The second of Eqs. (2.7) is the aggregate constraint faced by the central planner with $\ell_j = 1/\kappa_j$, and clearly, this constraint is satisfied by a Lagrange multiplier ψ_*, say, which exactly matches p_*, viz $\psi_* = p_*$, in which case $f_{ij} = \hat{f}_{ij}$ by construction. Moreover, this is the unique solution for ψ as f_{ij} is monotonically decreasing. ‖

2.3 The role of financial securities in markets with uncertainty

2.3.1 Commodity markets

The theory of general equilibrium provides a convenient means to define subfields in economics in contexts with frictionless markets. Consider the following definition.

> "A commodity is characterized by its physical properties, the date and the place at which it will be available."
>
> Gerard Debreu (1959, Chapter 2)

For example, by freezing time and physical properties, we have a theory of international commerce, and by freezing places and physical properties, we have theory of finance. However, the previous definition does not include the notion of uncertainty, which plays such an important role in financial markets. To cope with uncertainty, Debreu (1959, Chapter 7) extends the previous definition by highlighting that a commodity should still be described through a list of physical properties, but proposing that the structure of dates and places be replaced by an event structure. The following example illustrates the difference between two contracts underlying the delivery of corn arising under conditions of certainty (case A) and uncertainty (case B):

A *An agent will deliver 5000 tons of corn of a specified type to another agent, who will accept the delivery at date t and in place ℓ.*

B *An agent will deliver 5000 tons of corn of a specified type to another agent, who will accept the delivery in place ℓ and in the event s_t at time t. If s_t does not occur at time t, no delivery will take place.*

In both cases, the contract is paid at the time it is agreed. There are instances where we can actually use the model of the previous section to deal with contracts like that in case B. Consider, for example, a two-period economy, and suppose that in the second period, s_n mutually exhaustive and exclusive states of nature may occur. We can, then, recover the model of the previous section, once we replace m (the number of commodities described by physical properties, dates and places) with m^*, where $m^* = s_n \cdot m$. With m^* replacing m, the competitive equilibrium in this economy is defined as the competitive equilibrium in the economy of the previous section.

The assumption underlying this analogy is that markets exist, in which commodities are traded for each state of nature. These "contingent" markets are *complete*: at the "dawn of history," all markets open for all goods in each future state of the world, and all transactions are executed then; in the future, there will only be "settlements" by which counterparties will deliver and accept the quantities in accordance with the signed contracts. Because this economy is formally the same as that without uncertainty, resource allocation is Pareto-optimal. However, the assumption that there exist $s_n \cdot m$ contingent markets is strong. Next, we show how financial securities help mitigate this assumption.

2.3.2 Financial securities and rational expectations

The assumption that markets only open at the "dawn of history" is clearly unrealistic. In reality, markets open all the time. Indeed, Hicks (1939) had already developed a conceptual framework in which markets open in each period—a *temporary equilibrium theory*. In this framework, agents may well have incorrect expectations, and could transfer value by trading securities in each period.

Arrow (1953) formalizes the Hicksian idea that markets open all the time in a context with uncertainty. What role financial securities play in this context? Consider the following interpretation. Rather than signing contracts for delivery of *commodities* that are contingent on the realization of events, agents may agree on contracts generating *payoffs* that are contingent on the realization of events—i.e. financial securities. The payoffs delivered by the assets in the various states of the world can then be collected to finance state-contingent consumption plans. In other words, it is now financial assets to be traded at the dawn of history, and real markets will, instead, open in the various future states of the world, where the asset payoffs will help implement a state-contingent consumption plan. This economy is formally the same as Debreu's but the model emphasizes the presence of financial assets. Equally importantly, the model leads to a new and nuanced role of the equilibrium concept, the foundations to subsequent work on *rational expectations equilibrium* (Radner, 1968): at the dawn of history, the agents perfectly anticipate the equilibrium price of each good in each state of the world.

Let us illustrate. Consider a two-period economy. An asset $\mathcal{A}_i$ is a contract that promises to pay a payoff $v_i(s)$ in some state $s \in \mathcal{S}$ in the second period, where $\mathcal{S}$ denotes the set of all possible future events. We assume there exist X assets, and that markets re-open for each commodity in each state of nature in the second period. Assets and commodities are linked as follows. At time-zero, I purchase θ_i units of the asset $\mathcal{A}_i$. If state of nature s occurs in the second period, I will utilize the payoff $v_i(s)$ and finance net transactions on the commodity markets re-opening in the second period, viz

$$p(s) \cdot e(s) = \sum_{i=1}^{X} \theta_i v_i(s), \quad \forall s \in \mathcal{S}, \tag{2.8}$$

where $p(s)$ and $e(s)$ denote the commodity price and excess demand vectors, which are both contingent on the realization of state s. In this context, "rational expectations" means that at time-zero, I know what the price vector will be in state s, i.e., I know the vector function $p(s)$.

Thus, financial assets transfer value across states of nature. Security markets shrink this uncertain economy to one similar to that of the previous section, with one added feature: there are no such complicated markets to open at time zero for each state of nature in the future. There are, simply, commodity and asset markets in the first period, and commodity markets that re-open in the second period. This economy is more realistic than in the previous section although perfectly isomorphic to the previous once Eq. (2.8) holds true. However, Eq. (2.8) does not hold, in general. It would, should the number of assets be equal to the number of events in $\mathcal{S}$, i.e. $X = s_n$. We say security markets are *complete* in this case.

What is the fair value of these securities? Section 2.4 explains the basic theory of security evaluation, which relies on both a precise notion of arbitrage opportunties and how absence of arbitrage links to contexts with maximizing agents, and the existence of prices for a specific class of securities, known as Arrow-Debreu securities, and their relation to the notion of complete markets. The remainder of this section aims to develop in detail the notion of Arrow-Debreu securities, and a number of introductory examples illustrating their importance. We begin with

two examples, which aim to draw attention to what we really mean by absence of arbitrage and how this notion can help differentiate between "gambles" and "securities."

2.3.3 Securities prices as shadow values

The equivalence between Arrow's and Debreu's economies leads to the following conceptual difficulty. When markets are complete, financial assets are valued *as if* they were shadow values in an otherwise identical (Debreu's) economy in which they are not needed! This difficulty clearly shows how important it is that market imperfections could be introduced to deal with a more realistic description of the role of financial securities. These imperfections will be dealt with several times in these lectures.

2.3.4 Gambles and securities

Many explanations of asset prices in these lectures rely on the assumption that markets are arbitrage free, that is: two portfolios promising the same returns are worth the same and in the absence if any other frictions. This section buils up on ideas put forward in the Introduction while relying on a principle: securities are distinct from gambles, in that they could be priced by requiring absence of arbitrage.

Securities are contracts that are traded at prices reflecting a number of factors such as supply and demand, bargaining power, the market microstructure and, of course, absence of arbitrage, as explained in a moment. Instead, gambles are risks that are typically not traded (like in a casino), but there are nuances to be made below. Naturally, a gamble has a value that also depends on many factors, such as the bargaining power of the parties entering into play, and their risk-attitudes.

If we were capable to slice a gamble into elementary risks that are actually traded, we would be able to price the gamble through no-arbitrage: the fair value of the gamble would equal that of a portfolio comprising one unit of each of the risks the gamble is split into (see the examples of the previous section). In fact, such a gamble would not be like those in a casino anymore, but a "derivative" priced through the price of the underlying risks. Note, then, that this gamble is not traded (only its constituent risks are), although, then, it may be "replicated."

All in all, there are securities that can, and securities that cannot, be replicated, through the set of existing assets. Those that can, are priced without regard to anything but the prices of the assets that add up to the replicating portfolio. Arbitrage is the possibility to profit from price inconsistencies, by trading the risk constituents. Absence of arbitrage imposes discipline on the set of economically viable security prices. Likewise, there are gambles that can actually be replicated, as soon as their constituent risks are traded. What these gambles have in common with replicable securities, is that their price is free from anything relating to factors such as risk-aversion, or supply and demand, say, being then determined by no-arbitrage. Gambles that cannot be sliced in this way have a value that likely depends on the gamblers' risk-aversion or the bookmaker's bargaining power. Securities that cannot be replicated have this trait in common with unreplicable gambles: their value is not tied down to anything that is already traded, and subject to a number of factors that we may only speculate about. Still, the price of traded securities cannot be anything. To avoid arbitrage, security prices satisfy a fundamental economic restriction, known as the *martingale restriction*, which is the focus of much of this chapter.

2.3.4.1 Short-run bets and asymptotic arbitrage

What is the price of a coins game? If we could only toss the coin many times, then, by the Law of Large Numbers, the value of this gamble would approach 50% of a stake as the number of draws grows, to avoid an "asymptotic arbitrage," i.e. the possibility to gain from betting for tails, if the price for the tails is less than 50% of the stake over such a large number of trials—or selling tails, whenever this price is higher than 50%. Note that by allowing for the possibility to "sell tails," we are meaning that this gamble is, in fact, a tradeable asset.

Yet assume, and realistically, that the number of trials is small, in which case the value of this gamble might well deviate from 50% of a stake, reflecting for instance preferences, bargaining power, or in general, supply and demand. The previous asymptotic arbitrage could not take place because tossing a coin only once, say, is far from guaranteeing an almost sure 50% frequency of tails outcomes.

This example shows that the value of an unreplicable gamble might well depend on preferences and other factors—in short, by supply and demand. In fact, in the next section, we show that factors such as bettors' preferences are not only necessary to value a gamble. These preferences would sometimes even need to be restricted for us to conclude about the logical consistency of the very same gamble.

2.3.4.2 St Petersburg paradox

Consider the following, celebrated gamble, leading to the so-called *St Petersburg paradox*. We toss a coin a number of times, and the gamble ends as soon as a tail (say) arises as an outcome, with a payoff doubling all the time, being equal to £2^n if the gamble stops at n. The probability to receive £2^n at the n-th trial is, naturally, the probability that $n-1$ heads occur over the first $n-1$ trials and one tail in one trial, which in total is 2^{-n}, given the independence of the trials. Therefore, the "fair value" of this gamble is,

$$\sum_{n=1}^{\infty} (2^n) \times 2^{-n} = \infty.$$

This gamble is obviously not trivial, as there is a risk anyway, in that the payoff, although positive, is not certain. Because the payoff is positive, we have to pay something to enter this gamble, although it is unlikely that anyone would be willing to pay a large amount of money for it. Why? A moment reflection suggests the payoff profile is quite unattractive. For example, there are almost 95% chances to obtain less than £ 16, which illustrates how implausible it seems that anyone would be willing to spend a large amount of money to enter this gamble.

Bernoulli (1738) proposes to solve this puzzle by replacing outcomes with concave utilities of outcomes—concave utilities would dampen the occurrence of very positive returns. For example, a decision maker with log-utility, would perceive an expected utility from this gamble, equal to:

$$\sum_{n=1}^{\infty} (\ln 2^n) \times 2^{-n} = \ln 4.$$

Therefore, he would not be willing to pay more than £ 4, to enter this gamble. (This amount of money, £4, is the certainty equivalent of the gamble, CE, say, i.e. $\ln \mathrm{CE} = E(\ln \tilde{x})$, where $\tilde{x}$ is the random outcome from the gamble.) As anticipated in the previous section, gambles might be quite subtle. Not only does their value depend on preferences. Factors such as risk-aversion are critical to the very existence of such a game. In the St Petersburg paradox example, the game cannot exist were the bettors risk-seeking or risk-neutral.

2.3.4.3 World Cup pricing, and arbitrage pricing

The risks underlying the gambles of the previous examples cannot be evaluated in a preference-free fashion. As for the coins game, we would need a large number of draws to conclude its fair value is 50% of a stake. As for the coins game leading to the St Petersburg paradox, we would actually need risk-aversion to keep the value of the gamble finite. We now develop a third example where, again, the value of the gamble cannot be determined without making reference to supply and demand although then we could price some of the elementary risks that make up this gamble, in a preference-free fashion, for the very simple reason that these risks are actually traded.

Consider, then, the odds posted by a bookmaker, displayed in Table 2.1, against four hypothetical teams competing for the World Cup. In bookmaking language, odds set at a-b against a certain event mean that on a $£b$ stake, the bettor receives $£a$ if the given event occurs, and loses his stake of $£b$, otherwise. Therefore, odds of 5-1 against team B imply that the bookmaker stands ready to pay £500 for £100 stake, should B win the World Cup. Let p be the probability of the event that a given team wins, under which the bettor's expected gain is zero, $p : a \times p + (-b) \times (1-p) = 0$, such that $\frac{a}{b}$ equals the odds ratio, $\frac{1-p}{p}$, with b being fixed to one in this example. Also shown in Table 2.1 are these "implied break-even probabilities," calculated as $p_j = \frac{b_j}{b_j+a_j}$, where a_j-b_j are the odds against team j, and p_j are the corresponding break-even probability, with $b_j \equiv 1$.

Team	odds against	implied prob.
A	2-1	33.33%
B	5-1	16.67%
C	6-1	14.28%
D	9-1	10.00%

TABLE 2.1. Odds against teams in a World Cup.

As in the coins game of Section 2.3.3.1, the breakeven probabilities do not necessarily lead to the fair value of the gamble. Indeed, bookmakers typically make profits through what is known as the "overound"—quoted probabilities that exceed 100%. Overounds are obviously not arbitrage opportunities, because a given World Cup is quite an unique event, with its own weather conditions, teams components and many other factors, such that the chance of any team to win the competition has only a weak linkage to the objective probability of winning over an hypothetically large number of matches. Once again, the price and, then, the quoted odds, depend on both the bettors' risk-appetite and the bookmakers' bargaining power.

While the single odds depend on supply and demand, we can price other gambles which cover events relating to those the bookmakers are making a market for, and independently of supply and demand. For example, what are the odds against either team A or team B winning the World Cup? It is a "derivative gamble," because its value depends on that of the constituent events. Let us convert the odds into tickets, meaning that the cost of a ticket for a payoff equal to $£1$ for a a-b odds is equal to $c : £(1-c) \times p + £(-c) \times (1-p) = 0$; that is, $c = £\frac{b}{a+b}$, meaning that upon paying c for the ticket, we are entitled to $£1$ if the event occurs and $£0$ otherwise. For example, the unit ticket for team A winning is worth $£0.33333$, and that for team B is $£0.16667$. It easy to see, then, that the unit ticket value for the composite event needs

to be £0.33333 + £0.16667 = £0.5. If it was lower, say, £0.4, we could sell the two bets for A and B for £0.5, and use £0.4 to ensure that we could honour the payoff that A or B will win. If the unit ticket value for the composite event was, instead, £0.6, we could sell the composite bet and, then, bet for A and B separately, to honour the payoff relating to the composite event.

Pricing derivatives in a preference-free format as in this example relies on the fact that markets are *complete*, in that we could replicate the composite risk by betting on each single team. The next sections extend this notion to various situation of interest that include financial security transactions. We begin with introducing the simplest securities, those that only pay off in specific states of nature, similarly as the single bets of the World Cup of this section.

2.3.5 Arrow-Debreu securities

We now develop a very first evaluation model relying on the notion of a special type of quite simple securities, those that pay off one unit of numéraire in state s, if state s will prevail in the future, and zero otherwise. We usually refer this set of assets to as *Arrow-Debreu* securities, in honour of Kenneth Arrow and Gérard Debreu. Arrow-Debreu securities are perhaps the most elementary we might imagine, and constitute the bricks of the entire asset evaluation framework in financial economics, a bit like atoms in Physics, so to speak. Naturally, Arrow-Debreu securities do not exist in reality, for the very simple reason that it is impossible to determine states at such a pure level, in practice. States are simply part of a model's assumptions.

While Arrow-Debreu securities are model-dependent, they can be used to price assets—after all, asset pricing is model-dependent by nature. This section develops a very intuitive link between asset prices and Arrow-Debreu security prices, and explores how asset prices can, then, be re-cast in terms of a certain probability under which they equate their expected payoffs, in an actuarial sense. We conclude with a derivation of the price of these securities based on a simple equilibrium model.

2.3.5.1 Pricing

Consider a two-period economy where at time $t = 1$, three mutually exclusive states of the world may occur. We consider three Arrow-Debreu securities: the i-th, AD_i, pays off £1 in state i, and zero otherwise, as in the next figure.

	$t=1$	AD_1	AD_2	AD_3
	state 1	£1	0	0
$t=0$ ⟶	state 2	0	£1	0
	state 3	0	0	£1

Let ϕ_j be the price of the asset AD_j, and consider a portfolio that has one unit of each of the three assets. This portfolio yields one, for sure, at time $t = 1$. Therefore, its price must be that of a pure discount bond, or,

$$\frac{1}{1+r} = \sum_{j=1}^{3} \phi_j, \tag{2.9}$$

where r denotes the riskless interest rate. More generally, consider an asset $\mathcal{A}_0$, which pays off D_i in state i, as in the next figure. Naturally, the values of D_i are known at zero—although of

course we do not know which D_i will be drawn at time $t = 1$. Also shown in this figure are the payoffs of three assets, $\mathcal{A}_i$, $i = 1, 2, 3$, which are rescaled versions of Arrow-Debreu securities. Let us explain.

	$t=1$	$\mathcal{A}_0$	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$
	state 1	$£D_1$	$£D_1$	0	0
$t=0$ $\nearrow$ $\longrightarrow$ $\searrow$	state 2	$£D_2$	0	$£D_2$	0
	state 3	$£D_3$	0	0	$£D_3$

By definition, paying ϕ_1 yields 1 if state 1 occurs, and zero otherwise. Therefore, paying $\phi_1 \times D_1$ yields $1 \times D_1$ if state 1 occurs, and zero otherwise. We can, then, slice the risks related to asset $\mathcal{A}_0$ into three, by considering the three elementary assets $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$. Each of these assets costs $\phi_j D_j$ at $t = 0$. Moreover, a portfolio that has one unit of each of these three assets needs to have the same value of $\mathcal{A}_0$, say S,

$$S = \sum_{j=1}^{3} \phi_j D_j. \tag{2.10}$$

We can say more about S. Consider Eq. (2.9), which we rewrite as, $\sum_{j=1}^{3} Q_j = 1$, where $Q_j = (1 + r)\, \phi_j$. Thus, Q_j is a probability distribution. Replacing Q_j into Eq. (2.10) reveals an important property of Q_j,

$$S = \frac{1}{1+r} \sum_{j=1}^{3} Q_j D_j = \frac{1}{1+r} E^Q(\tilde{D}).$$

The price of $\mathcal{A}_0$ can be expressed as the expectation of the future payoff, $\tilde{D}$, discounted. It is as if we evaluated assets through actuarial methods. We shall refer to Q_j as risk-neutral probability for this reason, and we shall discuss the properties of this benchmark probability in detail in the next sections.

To sumup, Arrow-Debreu securities would allow us to price any asset, through Eq. (2.10). In Section 2.5, we shall actually show that there are no arbitrage opportunities if and only if there exists a vector $\phi : S = \phi^\top V$, where S is a m-dimensional vector of security prices, V is a $m \times d$ matrix of security payoffs, and d is the number of states. In this general context, the vector ϕ carries the natural interpretation of one containing Arrow-Debreu security prices. Note, also, that the security we are evaluating in this specific example relies on the assumption that available for trading are an appropriate set of Arrow-Debreu securities. Alternatively, we may assume that only a selected number of Arrow-Debreu securities are available for trading, in which case markets become "incomplete," in a sense to be made precise in Section 2.3.6. In this case, it is unlikely the given security could be given an unique as in Eq. (2.10) as we shall explain.

Note, finally, an interesting property of Arrow-Debreu securities. Consider a general context with a finite number of states, i.e. not necessarily equal to three, and define the random return of the i-th asset as $\tilde{R}_i = \frac{\tilde{x}_i}{\phi_i}$, where $\tilde{x}_i = 1$ in state i, and zero otherwise. Given the objective

probability distribution of the states, P_j, we have that each pair of asset returns is negatively correlated, with,

$$E(\tilde{R}_i) = \frac{P_i}{\phi_i}, \quad var(\tilde{R}_i) = \frac{P_i\left(1-P_i\right)}{\phi_i^2}, \quad corr(\tilde{R}_i, \tilde{R}_j) = -\sqrt{\frac{P_i P_j}{\left(1-P_i\right)\left(1-P_j\right)}}.$$

It is natural to expect these prices to be negatively correlated, as they pay off in mutually exclusive states of the world! As we decrease the number of states, the correlation obviously increases in absolute value, with the extreme case $corr(\tilde{R}_i, \tilde{R}_j) = -1$, arising when there are only two states, such that $P_i = 1 - P_j$. It is also natural, as we increase the number of states, there is a progressively higher number of states where any two Arrow-Debreu securities do not pay off, which brings the correlation of their returns up. This correlation is still negative, however.

2.3.5.2 Tying down Arrow-Debreu securities to consumption: an example

Consider the previous two-period economy, where now there is a single agent, who maximizes,

$$\max_{c_0, c_1}\left[u\left(c_0\right) + \beta E\left(u\left(c_1\right)\right)\right], \quad \text{s.t. (i) } c_0 + \theta S = \omega_0 \text{ and } \quad \text{(ii) } c_1 = \tilde{\omega}_1 + \theta\tilde{D},$$

where c_j is consumption at time $t = j$, ω_0 is endowment at time $t = 0$, $\tilde{\omega}_1$ is the random endowment at time 1, u is a concave utility function satisfying all the regularity we need, E is the expectation taken under the physical probability, and remaining notation is as before. We then have two constraints. The first (i), says that endowment is split into time $t = 0$ consumption and investment, and the second (ii) says that time $t = 1$ consumption is financed by $\tilde{\omega}_1$ and the payoff provided by the asset.

Replacing the two constraints into the utility functions, first order conditions are obtained by differentiating the resulting objective function with respect to θ, and lead to,

$$S = \frac{\beta E(u'(\tilde{\omega}_1 + \theta\tilde{D})\tilde{D})}{u'\left(\omega_0 - \theta S\right)} = \frac{\beta E(u'(\tilde{\omega}_1)\tilde{D})}{u'\left(\omega_0\right)} = \sum_{j=1}^{3} \frac{\beta u'\left(\omega_{1,j}\right)}{u'\left(\omega_0\right)} D_j P_j, \tag{2.11}$$

where $\omega_{1,j}$ is the endowment in state j, and P_j is the probability of state j. The second equality follows by an equilibrium condition, namely that the asset is in zero net supply, such that our Robinson Crusoe has to consume his endowment in each state.

Comparing Eq. (2.10) and (2.11) allows to identify Arrow-Debreu security prices, as follows,

$$\phi_j = \frac{\beta u'\left(\omega_{1,j}\right)}{u'\left(\omega_0\right)} P_j.$$

By concavity of the utility function, ϕ_j is high when $\omega_{1,j}$ is low. In other words, the Arrow-Debreu price for the "bad states" of nature are high or, risk-neutral probabilities assign high values to bad states. We shall come back to the intepretation of this result many times in this and subsequent chapters.

2.4 Arbitrage and replication: Introduction

We illustrate how absence of arbitrage leads asset prices to satisfy some restrictions. At this juncture, we cannot avoid explain that the concept of arbitrage was first relied upon in economics in a very famous contribution by Modigliani and Miller (1958). In this classic, Modigliani

and Miller showed that the value of a firm is independent of its capital structure. Chapter 4 discusses this important result in detail and develops explanations for the value of the firm for which this irrelevance result ceases to hold. This section aims to develop introductory examples that illustrate absence of arbitrage at work.

2.4.1 Rain and sunshine

We consider a simple introductory example to illsutrate how absence of arbitrage leads to a unique price of redundant securities. Consider an economy with complete markets, in which all uncertainty regards the event whether "tomorrow it will rain." Mr Law is a decision maker who would like to implement the following state-contingent plan. If tomorrow will be sunny, he will need $c_s > 0$ units of numéraire to purchases sunglasses; if tomorrow it will rain, he will need $c_r > 0$ to purchase an umbrella. Mr Law has access to a financial market where m assets are traded. He builds up a portfolio θ aimed to deliver the payments he will need tomorrow:

$$\begin{cases} \sum_{i=1}^{m} \theta_i S_i (1 + x_i(r)) = c_r \\ \sum_{i=1}^{m} \theta_i S_i (1 + x_i(s)) = c_s \end{cases} \tag{2.12}$$

where S_i is the price of the i-th asset, θ_i is the number of assets to put in the portfolio, and $x_i(r)$ and $x_i(s)$ are the net returns of asset i in the two states of nature, which are part of the information set available to Mr Law. Finally, and remarkably, we are not making any assumption regarding Mr Law's preferences. We only know that he needs c_i in state i.

Eqs. (2.12) form a system of two equations with m unknowns $(\theta_1, \cdots, \theta_m)$. If $m < 2$, no perfect hedging strategies are available to Mr Law—that is, Eqs. (2.12) cannot be solved to obtain the desired pair $(c_i)_{i=r,s}$. We say markets are incomplete in this case. More generally, consider an economy with s_n states of nature, ande define markets to be complete if and only if Mr Law has access to s_n assets, which allow him to achieve any consumption plan, independently of his preferences. Preferences would only play a role, if any, when it comes to pick up a particular consumption plan amongst all possible..

Let us, then, define the following "payoff matrix,"

$$V = \begin{bmatrix} S_1(1 + x_1(s_1)) & & S_m(1 + x_m(s_1)) \\ & \ddots & \\ S_1(1 + x_1(s_n)) & & S_m(1 + x_m(s_n)) \end{bmatrix},$$

where $x_i(s_j)$ is the return performed by the i-th asset in the state s_j. To implement any state contingent consumption plan $c \in \mathbb{R}^{s_n}$, Mr Law needs to solve the following system,

$$c = V \cdot \theta,$$

where $\theta \in \mathbb{R}^m$, the portfolio. If $\text{rank}(X) = s_n = m$, the previous system has a unique solution, given by $\hat{\theta} = V^{-1}c$. Consider, for example, the previous case where $s_n = 2$, and take $m = 2$, for any additional assets would be redundant here. Then, we have,

$$\begin{cases} \hat{\theta}_1 = \dfrac{(1 + x_2(r))c_s - (1 + x_2(s))c_r}{S_1 \left[(1 + x_1(s))(1 + x_2(r)) - (1 + x_1(r))(1 + x_2(s))\right]} \\ \hat{\theta}_2 = \dfrac{(1 + x_1(s))c_r - (1 + x_1(r))c_s}{S_2 \left[(1 + x_1(s))(1 + x_2(r)) - (1 + x_1(r))(1 + x_2(s))\right]} \end{cases}$$

Finally, assume that the second asset is safe, or that it yields the same return in the two states of nature: $x_2(r) = x_2(s) \equiv r$. Let $x_s = x_1(s)$ and $x_r = x_1(r)$. Then, the pair $(\hat{\theta}_1, \hat{\theta}_2)$ can be rewritten as,

$$\hat{\theta}_1 = \frac{c_s - c_r}{S_1(x_s - x_r)}, \quad \hat{\theta}_2 = \frac{(1 + x_s)c_r - (1 + x_r)c_s}{S_2(1 + r)(x_s - x_r)}.$$

As is clear, we are dealing with an issue relating to the *replication* of random variables. Our random variable is a state contingent consumption plan $(c_i)_{i=r,s}$, where c_r and c_p are known, which we want to replicate for hedging purposes—Mr Law will need to buy either a pair of sun-glasses or an umbrella, tomorrow.

In this example, any two-state variable can be generated by investing into two assets with independent payoffs. Our next step is to understand whether there are implications for the price of an asset that would exactly deliver the same random variable $(c_i)_{i=r,s}$. Let H be the current price of such an hypothetical asset, $\mathcal{A}$ say. We claim that

$$H = w \equiv \hat{\theta}_1 S_1 + \hat{\theta}_2 S_2, \tag{2.13}$$

for the financial market to be free of *arbitrage opportunities*, to be defined in a moment. Indeed, if $w < H$, we can buy $\hat{\theta}$ and sell at the same time the third asset $\mathcal{A}$. The result is a sure profit, or an arbitrage opportunity, equal to $H - w$, for $\hat{\theta}$ generates c_r if tomorrow it will rain, and c_s if tomorrow it will not rain. In both cases, the portfolio $\hat{\theta}$ generates the payments we need to honour the sale of $\mathcal{A}$. Likewise, we can show that the inequality $w > H$ would lead to an arbitrage. Hence, Eq. (2.13) must hold true.

We are left with the calculation of the right hand side of Eq. (2.13). We have:

$$H = \frac{1}{1 + r}\left[Qc_s + (1 - Q)c_r\right], \quad Q = \frac{x_r - r}{x_s - x_r}. \tag{2.14}$$

Eq. (2.14) is an evaluation formula for the asset $\mathcal{A}$, and says that the price, H, can be expressed as the present value of the expected payoffs promised by $\mathcal{A}$ under a new probability, Q, which for obvious reasons we term *risk-neutral probability*. Note that, remarkably, we are able to price the asset $\mathcal{A}$ without any reference to agents' preferences. The reasons underlying this *preference-free* result link to the fact that we can replicate the asset $\mathcal{A}$, through $\hat{\theta}$. Eq. (2.14) does not obviously require that any agent is using this portfolio. For example, Mr Law might be so poor that he could not even implement $\hat{\theta}$. The point is, rather, that the portfolio $\hat{\theta}$ could be used to implement an arbitrage opportunity, arising as soon as Eq. (2.13) does not hold. In this case, any penniless agent could implement the arbitrage described above.

The next step is to extend Eq. (2.14) to a dynamic setting. Suppose an additional day is available for trading, with the same description as before: the day after tomorrow, the asset $\mathcal{A}$ payoffs are c_{ss} if it will be sunny, provided the previous day was sunny; c_{rs} if it will be sunny, provided the previous day was raining, etc. By the same arguments leading to Eq. (2.14), we have that:

$$H = \frac{1}{(1 + r)^2}\left[Q^2 c_{ss} + Q(1 - Q)c_{sr} + (1 - Q)Qc_{rs} + Q^2 c_{rr}\right].$$

Finally, by extending the same reasoning to T trading days,

$$H = \frac{1}{(1 + r)^T} E^Q(c_T), \tag{2.15}$$

where E^Q denotes the expectation taken under the probability Q.

The key assumption underlying Eq. (2.15) is that markets are complete at each trading day. True, at the beginning of the trading period, Mr Law faces 2^T mutually exclusive states of nature for T. If he did not have access to markets in each period, he would need 2^T securities to replicate the asset $\mathcal{A}$. Yet in this example, 2 assets *and* T trading periods for these assets are needed, to replicate and, then, price $\mathcal{A}$. To emphasize this dynamic feature, we say the that the structure of assets and transaction dates make the markets *dynamically complete*. Dynamically complete markets allow to implement dynamic trading strategies that replicate the value of the asset $\mathcal{A}$, period by period. As a result, the asset $\mathcal{A}$ is priced, without any need to assume anything about the agents' preferences. The next section further clarifies these issues, and the existence of the risk-neutral probability, Q.

2.4.2 Replication and pricing: the role of complete markets

What are the origins of the preference-free formula in Eq. (2.14)? Consider a general two-period model where the assets deliver a payoff matrix, V, and denote as usual with ϕ the vector of Arrow-Debreu prices, ϕ. We know that the initial price vector is, $S = \phi^\top V$. In a setting with complete markets, we can extract a unique vector of state prices from the previous relation,

$$\hat{\phi}^\top = SV^{-1}. \tag{2.16}$$

Next, we want to replicate any asset (e.g., a derivative) with final payoff c at the second period, through a portfolio comprising the initial assets. Denote the initial value of this portfolio with w and the value in the second period with w',

$$w \equiv S\theta, \quad w' \equiv V\theta. \tag{2.17}$$

We want to use this portfolio to replicate the payoff of the derivative in the second period, i.e.,

$$c = w' = V\theta \Longrightarrow \hat{\theta} = V^{-1}c.$$

We plug $\hat{\theta} = V^{-1}c$ into w of the first of Eqs. (2.17) and obtain the initial value of the replicating portfolio,

$$\hat{w} \equiv S\hat{\theta} = SV^{-1}c = \hat{\phi}^\top c,$$

where the last equality follows by Eq. (2.16). As usual, two portfolios that yield the same thing must be worth the same, in the presence of frictionless markets. Therefore, the price of the derivative, H, is,

$$H = \hat{w} = \hat{\phi}^\top c = \frac{1}{1+r}E^Q(\tilde{c}).$$

Done. We could price the given asset in this preference-free fashion because markets are complete in that the asset payoff can be replicated through securities that are actually traded (the complete market assumption). These traded securities "convey" information about Arrow-Debreu securities, ϕ, to the derivative, so to speak. Chapters 11 through 13 explain how this "extract-and-plug-in procedure" is used to evaluate complex derivative instruments in the market practice.

In the next sections, we study Arrow-Debreu securities from a more theoretical perspective, connecting them more deeply with the notion of absence of arbitrage, and relying on general equilibrium model without production, and with possibly incomplete markets. We now introduce a fundamental pillar in financial economics: a firm value is independent of its capital structure in markets without arbitrage and in absence of frictions.

2.5 No-arbitrage: theory

This section formalizes the notion of absence of arbitrage opportunities and link it the competitive equilibrium in Section 2.2. We consider markets with one single commodity, with Appendix 2 dealing briefly with the intricacies arising in the multi-commodities case.

2.5.1 Lands of Cockaigne

Let $v_i(\omega_s)$ be the payoff of asset i in state ω_s, $i = 1, \cdots, m$ and $s = 1, \cdots, d$. Consider the payoff matrix:

$$V \equiv \begin{bmatrix} v_1(\omega_1) & & v_m(\omega_1) \\ & \ddots & \\ v_1(\omega_d) & & v_m(\omega_d) \end{bmatrix}.$$

Let $v_{si} \equiv v_i(\omega_s)$, $v_{s,\cdot} \equiv [v_{s1}, \cdots, v_{sm}]$, $v_{\cdot,i} \equiv [v_{1i}, , \cdots, v_{di}]^\top$. We assume that $\text{rank}(V) = m \leq d$.

The budget constraint of each agent is:

$$\begin{cases} c_0 - w_0 = -S\theta = -\sum_{i=1}^{m} S_i\theta_i \\ c_s - w_s = v_{s\cdot}\theta = \sum_{i=1}^{m} v_{si}\theta_i, \ s = 1, \cdots, d \end{cases}$$

Let $x^1 = [x_1, \cdots, x_d]^\top$. The second constraint can be written as:

$$c^1 - w^1 = V\theta.$$

We define an arbitrage opportunity as a portfolio that has a negative value at the first period, and a positive value in at least one state of world in the second period, or a positive value in all states of the world in the second period and a nonpositive value in the first period. Let us introduce the following pieces of notation.

NOTATION. [In progress ...] Given a vector $x \in \mathbb{R}^m$, $x > 0$ means that at least one component of x is strictly positive while the other components of x are nonnegative. $x \gg 0$ means that all components of x are strictly positive. $V\theta > 0$ means $[V\theta]_j \geq 0$, $j = 1, \cdots, d$, with at least one j for which $[V\theta]_j > 0$. $V\theta \geq 0$ means that $[V\theta]_j \geq 0$, $j = 1, \cdots, d$, i.e. it allows for $[V\theta]_j = 0$, $j = 1, \cdots, d$.

[**Insert here further notes**]

DEFINITION 2.9. *An arbitrage opportunity is a strategy θ that yields either $V\theta \geq 0$ with an initial investment $S\theta < 0$, or a strategy θ that produces $V\theta > 0$ with an initial investment $S\theta \leq 0$.*

An arbitrage opportunity cannot exist in a competitive equilibrium, for the agents' programs would not be well-defined in this case. Precisely, consider the $(d+1) \times m$ matrix,

$$W = \begin{bmatrix} -S \\ V \end{bmatrix},$$

the vector subspace of $\mathbb{R}^{d+1}$,

$$\langle W\rangle = \left\{z \in \mathbb{R}^{d+1} : z = W\theta,\ \theta \in \mathbb{R}^m\right\},$$

and the null space of $\langle W\rangle$,

$$\langle W\rangle^{\perp} = \left\{x \in \mathbb{R}^{d+1} : xW = \mathbf{0}_m\right\}.$$

The interpretation of the vector subspace $\langle W\rangle$ is that of the excess demand space in all the states of nature, generated by the *income transfers across states* induced by the portfolio choice θ. Naturally, $\langle W\rangle^{\perp}$ and $\langle W\rangle$ are orthogonal, as $\langle W\rangle^{\perp} = \left\{x \in \mathbb{R}^{d+1} : xz = 0,\ z \in \langle W\rangle\right\}$. (We shall return to the interpretation of $\langle W\rangle^{\perp}$ below; see Eq. (2.20)). The assumption there are no arbitrage opportunities is equivalent to the following condition,

$$\langle W\rangle \bigcap \mathbb{R}^{d+1}_{+} = \{0\}. \tag{2.18}$$

The interpertation of Eq. (2.18) is that in the absence of arbitrage opportunities, there should be no portfolios generating income transfers that are (i) non-negative and (ii) strictly positive in at least one state, i.e. $\nexists\theta : W\theta > 0$. Hence, $\langle W\rangle$ and the positive orthant $\mathbb{R}^{d+1}_{+}$ cannot intersect.

2.5.2 Enforced asset prices

The no-arbitrage condition in Eq. (2.18) tells us that the prices of all the assets satisfy a joint restriction:

THEOREM 2.10. *There are no arbitrage opportunities if and only if there exists a* $\phi \in \mathbb{R}^d_{++}$: $S = \phi^{\top}V$. *If* $m = d$, ϕ *is unique, and if* $m < d$, $\dim\left(\phi \in \mathbb{R}^d_{++} : S = \phi^{\top}V\right) = d - m$.

PROOF. In the appendix.

Theorem 2.10 provides foundations to the main architecture underlying asset evaluation in frictionless markets. Pre-multiply the second constraint by $\phi^{\top}$, obtaining,

$$\phi^{\top}(c^1 - w^1) = \phi^{\top}V\theta = S\theta = -(c_0 - w_0),$$

where the second equality follows by Theorem 2.10, and the third equality is the first period budget constraint. Critically, then, Theorem 2.10 shows that in the absence of arbitrage, each agent faces the following budget constraint,

$$0 = c_0 - w_0 + \phi^{\top}\left(c^1 - w^1\right) = c_0 - w_0 + \sum_{s=1}^{d}\phi_s\left(c_s - w_s\right),\ \text{with}\ \left(c^1 - w^1\right) \in \langle V\rangle, \tag{2.19}$$

where $\langle V\rangle$ is the subspace generated by the payoffs stemming from the portfolio choices,

$$\langle V\rangle = \left\{e \in \mathbb{R}^d : e = V\theta,\ \theta \in \mathbb{R}^m\right\}.$$

Eq. (2.19) suggests to interpret ϕ as the price vector of commodities available for consumption over the various states of nature, with first-period consumption being the numéraire, in the spirit of the Debreu (1959, Chapter 7) quote given in Section 2.3.1. Moreover, Theorem 2.10 tells us

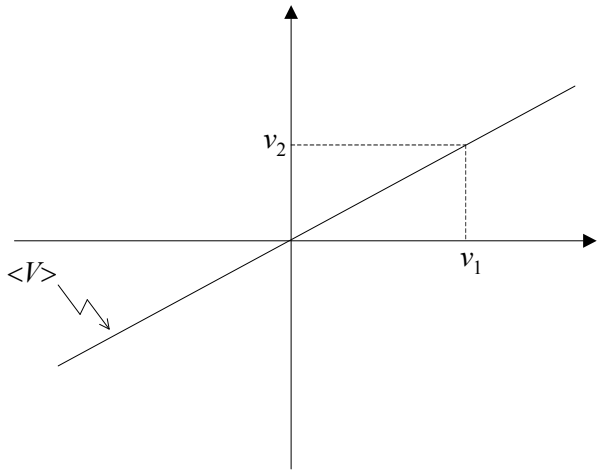


FIGURE 2.2. Incomplete markets, $d = 2$, $m = 1$.

that because $S = \phi^\top V$, the vector ϕ is, then, the *Arrow-Debreu state price* vector, generalizing the heuristic notions introduced in Section 2.3.4. Finally, we can interpret the subspace $\langle W\rangle^\perp$ as follows. Let $\phi_1 \equiv [1\ \phi]^\top$. Then, by Eq. (2.19),

$$\phi_1 \in \langle W\rangle^\perp, \tag{2.20}$$

In particular, no-arbitrage implies that $\langle W\rangle^\perp \bigcap \mathbb{R}^{d+1} = \{0\}$.

Notwithstanding these interesting conceptual links, it is misleading to interpret the budget constraint in (2.19) is that arising within the static Arrow-Debreu type model of Section 2.2, as this case would only obtain when $m = d$, in which case $\langle V\rangle = \mathbb{R}^d$ in (2.19). This case, which according to Theorem 2.0 arises when markets are complete, also implies the remarkable property that there exists a unique ϕ that is compatible with the asset prices we observe. If markets are incomplete, $m < d$, the situation quite differs, as the agents have then access to a "smaller" subspace of excess demands in the second period, $\langle V\rangle \subseteq \mathbb{R}^d$.

Consider, for example, the case $d = 2$ and $m = 1$. In this case, $\langle V\rangle = \{e \in \mathbb{R}^2 : e = V\theta,\ \theta \in \mathbb{R}\}$, with $V = V_1$, where $V_1 = \binom{v_1}{v_2}$ say, and $\dim\langle V\rangle = 1$, as illustrated by Figure 2.2. Next, suppose we open a new market for a second financial asset with payoffs given by: $V_2 = \binom{v_3}{v_4}$. Then, $m = 2$, $V = \left(\begin{smallmatrix} v_1 & v_3 \\ v_2 & v_4 \end{smallmatrix}\right)$, and $\langle V\rangle = \left\{e \in \mathbb{R}^2 : e = \binom{\theta_1 v_1 + \theta_2 v_3}{\theta_1 v_2 + \theta_2 v_4},\ \theta \in \mathbb{R}^2\right\}$, i.e. $\langle V\rangle = \mathbb{R}^2$. As a result, we can now generate any excess demand in $\mathbb{R}^2$, just as in the Arrow-Debreu economy of Section 2.2. To generate any excess demand, we multiply the payoff vector V_1 by θ_1 and the payoff vector V_2 by θ_2. For example, suppose we wish to generate the payoff the payoff vector V_4 in Figure 2.3. Then, we choose some $\theta_1 > 1$ and $\theta_2 < 1$. (The exact values of θ_1 and θ_2 are obtained by solving a linear system.) In Figure 2.3, the payoff vector V_3 is obtained with $\theta_1 = \theta_2 = 1$.

We are in a position to state a fundamental result regarding the "viability of the model." Define the second period consumption $c_j^1 \equiv [c_{1j}, \cdots, c_{dj}]^\top$, where c_{sj} is the second-period consumption in state s, and let,

$$\left(\hat{c}_{0j}, \hat{c}_j^1\right) \in \arg\max_{c_{0j}, c_j^1} \left[u_j\left(c_{0j}\right) + \beta_j E(\nu_j(c_j^1))\right], \ \text{ subject to } \left\{ \begin{array}{rcl} c_{0j} - w_{0j} & = & -S\theta_j \\ c_j^1 - w_j^1 & = & V\theta_j \end{array} \right. \tag{2.P3}$$

where u_j and ν_j are utility functions, both satisfying Assumption 2.1. Naturally, we could use more general formulations of utilities than that in [2.P3], and in fact we shall in more advanced

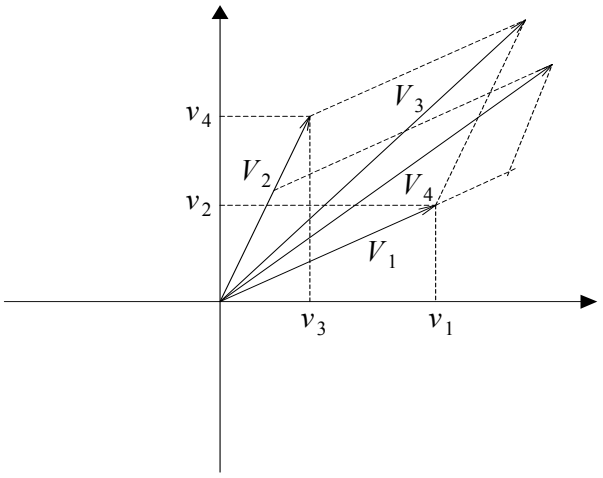


FIGURE 2.3. Complete markets, $\langle V \rangle = \mathbb{R}^2$.

parts of these lectures. For the sake of this introductory chapter, we only consider additive utility.

We have:

THEOREM 2.10. *The program [2.P3] has a solution if and only if there are no arbitrage opportunities.*

PROOF. Let us suppose on the contrary that the program [2.P3] has a solution $\hat{c}_{0j}, \hat{c}_j^1, \hat{\theta}_j$, but that there exists a $\theta : W\theta > 0$. The program constraint is, with straightforward notation, $\hat{c}_j = w_j + W\hat{\theta}_j$. Then, we may define a portfolio $\theta_j = \hat{\theta}_j + \theta$, such that $c_j = w_j + W(\hat{\theta}_j + \theta) = \hat{c}_j + W\theta > \hat{c}_j$, which contradicts the optimality of $\hat{c}_j$. For the converse, note that the absence of arbitrage opportunities implies that $\exists \phi \in \mathbb{R}^d_{++} : S = \phi^\top V$, which leads to the budget constraint in (2.19), for a given ϕ. This budget constraint is clearly a closed subset of the compact budget constraint B_j in [2.P1] (in fact, it is B_j restricted to $\langle V \rangle$). Therefore, it is a compact set and, hence, the program [2.P3] has a solution, as a continuous function attains its maximum on a compact set. ‖

2.6 Equivalent martingales, and equilibrium

We provide the definition of an equilibrium with financial markets, when the financial assets are in zero net supply.

DEFINITION 2.11. *An equilibrium is given by allocations and prices* $\{(\hat{c}_{0j})_{j=1}^n, ((\hat{c}_{sj})_{j=1}^n)_{s=1}^d, (\hat{S}_i)_{i=1}^m \in R_+^n \times R_+^{nd} \times R_+^d\}$*, where the allocations are solutions to program [2.P3], and satisfy:*

$$0 = \sum_{j=1}^n (\hat{c}_{0j} - w_{0j})\,; \quad \text{and for } s = 1, \cdots, d, \quad 0 = \sum_{j=1}^n (\hat{c}_{sj} - w_{sj}) \text{ and } 0 = \sum_{j=1}^n \theta_{ij} \quad (i = 1, \cdots, d)\,.$$

We now express demand functions in terms of the stochastic discount factor, and then look for an equilibrium by looking for the stochastic discount factor that clears the commodity markets.

By Walras' Law, clearing in the commodity market implies clearing in the financial market. Indeed, by aggregating the agent's constraints in the second period,

$$\sum_{j=1}^{n}\left(c_j^1 - w_j^1\right) = V\sum_{j=1}^{n}\theta_{ij}.$$

For simplicity, we also assume that $u_j'(x) > 0$, $u_j''(x) < 0$ $\forall x > 0$ and $\lim_{x\to 0} u_j'(x) = \infty$, $\lim_{x\to\infty} u_j'(x) = 0$ and that ν_j satisfies the same properties.

2.6.1 Rational expectations

Rational expectations is the assumption that agents beliefs are *endogenous*, in that they are entirely consistent with the probability distribution of the equilibrium outcome. In other words, agents *correctly* anticipate the equilibrium price in each state of nature. Consider, for example, the agents' expectations of an equilibrium price. In a rational expectations setting, the equilibrium price is a function of the state; accordingly, the equilibrium distribution of the price specifies the likelihood of any price realization in correspondence of each state.[1] Strictly speaking, the agents' price expectations are rational when they are taken with respect to such equilibrium probability laws. Endogenous beliefs "reinforce" the equilibrium, so to speak. That is, beliefs are endogenous, in that they are determined by the equilibrium of the model, but then, the equilibrium of the model is determined by the agents' actions, which ultimately depend on their beliefs.

[In progress: mention the literature. Radner (1968), Green, Lucas]

Consider the following example, a model with asymmetric information (see Chapters 4 and 10). Uninformed agents base their portfolio decisions while only relying on the asset price they observe in equilibrium. The equilibrium asset price depends on a noisy signal of the information possessed by some informed investors about the asset payoff. The asset payoff is d and the noisy signal is s, and the relevant piece of statistical inference for the uninformed investor is the evaluation of the asset payoff given the price he observes. In this market, "rational expectations" means that uninformed agents correctly anticipate the functional form of the asset price: they (correctly) conjecture that the asset price is a function of the state of the world, say $\sigma(s, Z)$, where Z is (per-capita) supply (a random variable). Accordingly, the uninformed agents estimate of the asset payoff is $E\left(d\middle|\, p\left(\sigma\left(s,Z\right)\right) = p\right)$, where $p\left(\sigma\left(\cdot,\cdot\right)\right)$ is the rational expectation asset price, a function of the state of the world that is relevant for the asset evaluation, i.e., $\sigma(\cdot,\cdot)$. This estimate is based on the equilibrium probability laws regarding the asset price.

Based on this inference, each uninformed investor forms a portfolio $x_U\left(p\left(\sigma\left(\cdot,,\cdot\right)\right)\right)$ (say), the units of the assets he purchases. Instead, each informed investor's portfolio choice is a function $x_I(d)$, and in equilibrium,

$$Z = \lambda x_I(d) + (1-\lambda)\, x_U\left(p\left(\sigma\left(s,Z\right)\right)\right),$$

where λ is the proportion of informed agents. A rational expectation equilibrium is an asset price function that satisfies the previous equation.

Chapters 4 and 10 contains detailed explanations regarding the importance of the source of randomness, Z. Intuitively, the presence of a supply shock guarantees the existence of an

[1] To illustrate, assume that the state is a dividend d that is uniformly distributed on $(0,1)$, and that the equilibrium price is $p = 1 + d$: in this case, the equilibrium distribution for p is a uniform law on $(1,2)$.

equilibrium in this market with asymmetric information, because it makes it impossible for the uninformed investors to reverse-engineer the information available to the informed investors, d. That is, prices only partially reveal the market fundamentals, d. For this reason, this equilibrium concept is commonly referred to as *noisy* rational expectation equilibrium. We now proceed with the simpler models of this chapter, in which information is disseminated symmetrically across agents.

2.6.2 Stochastic discount factors

Theorem 2.10 states that in the absence of arbitrage opportunities,

$$S_i = \phi^\top v_{\cdot,i} = \sum_{s=1}^{d} \phi_s v_{s,i}, \quad i = 1, \cdots, m. \tag{2.21}$$

We assume the first asset is safe, in that $v_{s,1} = 1 \; \forall s$, such that

$$S_1 \equiv \frac{1}{1+r} = \sum_{s=1}^{d} \phi_s. \tag{2.22}$$

Eq. (2.22) validates the economic interpretation of the state prices in Eq. (2.19). Because the states of nature are exhaustive and mutually exclusive, ϕ_s is the price to be paid today to obtain one unit of numéraire, tomorrow, in state s. Eq. (2.22) then confirms that the prices of all these rights sum up to the price of a pure discount bond—i.e. the asset yielding one unit of numéraire, tomorrow, for sure.

Eq. (2.22) can be elaborated to provide us with a second interpretation of the state prices in Theorem 2.10. Define,

$$Q_s \equiv (1+r)\phi_s,$$

which satisfies, by construction,

$$\sum_{s=1}^{d} Q_s = 1.$$

Therefore, we can interpret $Q \equiv (Q_s)_{s=1}^{d}$ as a probability distribution. Moreover, by replacing Q in Eq. (12.17) leaves,

$$S_i = \frac{1}{1+r} \sum_{s=1}^{d} Q_s v_{s,i} = \frac{1}{1+r} E^Q \left(v_{\cdot,i} \right), \quad i = 1, \cdots, m. \tag{2.23}$$

Eq. (2.23) confirms Eq. (2.14), obtained in the introductory example of Section 2.4. It says that the price of any asset is the expectation of its future payoffs, taken under the probability Q, discounted at the risk-free interest rate r. For this reason, we usually refer to the probability Q as the *risk-neutral* probability. Eq. (2.23) can be extended to a dynamic context, as we shall see in later chapters. Intuitively, consider an asset that distributes dividends in every period, let S_t be its price at time t, and D_t the dividend paid off at time t. Then, the "payoff" it promises for the next period is $S_{t+1} + D_{t+1}$. Then, and consistently with Eq. (2.23), $S_t = (1+r)^{-1} E_t^Q \left(S_{t+1} + D_{t+1} \right)$ where E_t^Q denotes the risk-neutral expectation given the information at time t, or, by rearranging terms,

$$E_t^Q \left(\frac{S_{t+1} + D_{t+1} - S_t}{S_t} \right) = r. \tag{2.24}$$

That is, the expected return on the asset under Q equals the safe interest rate, r. In a dynamic context, the risk-neutral probability Q is also referred to as the *risk-neutral martingale measure*, or *equivalent martingale measure*, for the following reason. Define a money market account as an asset with value evolving over time as $M_t \equiv (1+r)^t$. Then, Eq. (2.24) can be rewritten as $S_t/M_t = E_t^Q\left[\left(S_{t+1}+D_{t+1}\right)/M_{t+1}\right]$. This shows that if $D_{t+1}=0$ for some t, then, the discounted process S_t/M_t is a martingale under Q.

Next, let us replace Q into the budget constraint in (2.19), to obtain, for $(c^1-w^1)\in\langle V\rangle$,

$$0 = c_0 - w_0 + \sum_{s=1}^{d}\phi_s\left(c_s-w_s\right) = c_0 - w_0 + \frac{1}{1+r}\sum_{s=1}^{d}Q_s\left(c_s-w_s\right) = c_0 - w_0 + \frac{1}{1+r}E^Q\left(c^1-w^1\right). \tag{2.25}$$

For reasons developed below, it is also useful to derive an alternative representation of the budget constraint, in terms of the objective probability P, say. Accordingly, we introduce the state-dependent ratio ζ_s that indicates how far Q_s and P_s are:

$$\zeta_s = \frac{Q_s}{P_s}, \quad s=1,\cdots,d.$$

We assume ζ_s is strictly positive and bounded—mathematically, it is tantamount to say that Q and P are *equivalent* measures, in that they assign the same weight to the null sets. Define the *stochastic discount factor*, $m=(m_s)_{s=1}^d$, defined as,

$$m_s \equiv \frac{1}{1+r}\zeta_s = \frac{1}{1+r}\frac{Q_s}{P_s} = \frac{\phi_s}{P_S}. \tag{2.26}$$

We have,

$$\frac{1}{1+r}E^Q\left(c^1-w^1\right) = \sum_{s=1}^{d}\frac{1}{1+r}Q_s\left(c_s-w_s\right) = \sum_{s=1}^{d}\underbrace{\frac{1}{1+r}\zeta_s}_{=m_s}\left(c_s-w_s\right)P_s = E\left[m\cdot\left(c^1-w^1\right)\right].$$

Hence, we can rewrite Eq. (2.25) as,

$$0 = c_0 - w_0 + E\left[m\cdot\left(c^1-w^1\right)\right], \quad \left(c^1-w^1\right)\in\langle V\rangle.$$

Similarly, by replacing the stochastic discount factor m into Eq. (2.23) we obtain,

$$S_i = \frac{1}{1+r}E^Q\left(v_{\cdot,i}\right) = E\left(m\cdot v_{\cdot,i}\right), \quad i=1,\cdots,m. \tag{2.27}$$

We now derive optimality conditions and, then, solve for the equilibrium in this economy.

2.6.3 Equilibrium

In the absence of arbitrage, any agent j faces the following program:

$$\max_{(c_0,c^1)}\left[u_j(c_{0j})+\beta_j\cdot E(\nu_j(c_j^1))\right] \quad \text{subject to } 0 = c_{0j}-w_{0j}+E\left[m\cdot\left(c_j^1-w_j^1\right)\right], \quad \left(c^1-w^1\right)\in\langle V\rangle. \tag{2.P4}$$

This way to formalize the agents problem and constraints is quite convenient and, it helps understand the nature of incomplete markets, as in the cases illustrated through Figures 2.2

and 2.3. In this section, we will further illustrate how useful the representation of the program in [2.P4] is, while studying incomplete markets in general, through the so-called min-max stochastic discounting approach. Magill and Quinzii (1996) contain an extensive analysis of how this representation helps studying general equilibrium with incomplete markets in quite general models. Truth be said, the representation in [2.P4] is not the exclusive way to study decision problems arising in incomplete markets. In some cases, such as those covered by some portions of Chapter 9 (see Section 9.5), it seems easier to make reference to the initial program in [2.P3]. The choice of which program to consider, [2.P3] or [2.P4], quite depends on the problem at hand.

2.6.3.1 Complete markets and risk sharing

In the complete markets case, $\langle V \rangle = \mathbb{R}^d$, so that the first order conditions to the program [2.P4] are,

$$u_j'(\hat{c}_{0j}) = \lambda_j, \quad \beta_j \nu_j'(\hat{c}_{sj}) = \lambda_j m_s, \quad s = 1, \cdots, d, \tag{2.28}$$

where λ_j is a Lagrange multiplier. So, really, the properties of this model are the same as those of the static model in Section 2.2. Formally, the complete markets economy in this section is the same as the static economy in Section 2.2, once we set $m = d$, where m is the dimension of the commodity space, in Section 2.2, and $p_s = \phi_s$, where p_s is the price of the s-th commodity in Section 2.2, with $p_1 = 1$ (the numéraire), and ϕ_s is the Arrow-Debreu state price in the unified budget constraint of Eq. (2.25).

These facts formalize the reasoning made at the beginning of Section 2.3 (see Sections 2.3.1 and 2.3.2): when markets are complete, an economy with uncertainty can be understood through a static one. Complex markets with heterogeneous agents have potentially interesting asset pricing implications, and still, apparently, seem so hopelessly difficult to analyze; however, we shall see that they can be still "centralized" through a dedicated design of Pareto's weights, as formalized in Theorem 2.7.

These properties are robust to dynamic extensions, as explained in Chapter 5 and Part II (Chapter 9), provided markets are dynamically complete—a property explained in the next two chapters. However, the assumption seems unrealistic that agents can trade Arrow-Debreu securities for all states of the world: one reason financial innovation is in practice so pervasive is that markets are incomplete. Yet, as discussed in Chapter 9, markets have been centralized in economies with incomplete markets, relying on "stochastic Pareto weights."

We now derive the equilibrium implications of the first-order conditions in the simple case of an economy with a single agent, where the following stochastic discount factor is:

$$m_s = \beta \frac{\nu'(w_s)}{u'(w_0)}.$$

The economic interpretation of m_s is the following. In the autarchic case,

$$-\frac{dc_0}{dc_s}\bigg|_{c_0=w_0, c_s=w_s} = \beta \frac{\nu'(w_s)}{u'(w_0)} P_s = m_s P_s = \phi_s$$

is the present consumption the agent is willing to give up to at $t = 0$, in order to obtain additional consumption at time $t = 1$, in state s. In other words, ϕ_s is the price, in terms of the present consumption numéraire, of one additional unit of consumption at time $t = 1$ and state s. So it is a state price that makes the agent happy to consume his own endowment, without

any incentives left to trade in the financial markets. The risk-neutral probability is,

$$Q_s = \zeta_s P_s = (1+r)\, m_s P_s = (1+r)\, \beta \frac{\nu'(w_s)}{u'(w_0)} P_s.$$

By the first order conditions, and the expression for the evaluation of a pure discount bond, $\frac{1}{1+r} = E(m)$, we have that $1 = \sum_{s=1}^{d} Q_s$. Moreover,

$$\frac{Q_s}{P_s} = m_s\,(1+r) = m_s \beta^{-1} \frac{u'(w_0)}{E\left[\nu'(w_s)\right]} = \frac{\nu'(w_s)}{E\left[\nu'(w_s)\right]}.$$

The case with heterogenous agents is similar provided markets are complete. By the optimality conditions applying to each agent (see Eq. (2.28)), and Eq. (2.26), the marginal rate of substitution for each agents are:

$$\beta_j \frac{\nu_j'\left(\hat{c}_{sj}\right)}{u_j'\left(\hat{c}_{0j}\right)} P_s = \phi_s, \quad s = 1, \cdots, d, \quad j = 1, \cdots, n. \tag{2.29}$$

That is, in equilibrium agents *do* have the same marginal rate of substitution when markets are complete. It is so because the vector ϕ of state prices and, then, m_s is unique if and only if markets are complete (see Theorem 2.10). The marginal rates of substitution are, then, independent of j, and the equilibrium allocation is a Pareto optimum as a result, by the discussion at the beginning of this section, and Theorem 2.5.

The fact that agents have the same marginal rates of substitution in each state of the world is known as *risk-sharing*. It means that, given an initial endowment distribution, the market mechanism is capable to shift risks around through a system of complete security markets. Risks are borne more by the agents most willing to take them.

Suppose for example that two agents have the same utility and discount rate but that the first agent is less risk-averse than the second, i.e. the CRRAs satisfy: $\gamma_1 < \gamma_2$. Then, $\mathrm{Gr}_{s1} = (\mathrm{Gr}_{s2})^{\gamma_2/\gamma_1}$, where Gr_{si} is consumption growth for the i-th agent in state s. In good times, when $\mathrm{Gr}_{s2} > 1$, the more risk-averse agent experiences a lower consumption growth rate ex-post, $\mathrm{Gr}_{s2} < \mathrm{Gr}_{s1}$. However, in bad times, when $\mathrm{Gr}_{s2} < 1$, the more risk-averse agent experiences a higher consumption growth rate ex-post, $\mathrm{Gr}_{s2} > \mathrm{Gr}_{s1}$, as illustrated by Figure 2.4 in the case where $\gamma_2/\gamma_1 = 3$. In other words, capital markets, when complete, operate in such a way to have the more risk-averse agent face a less volatile consumption growth.

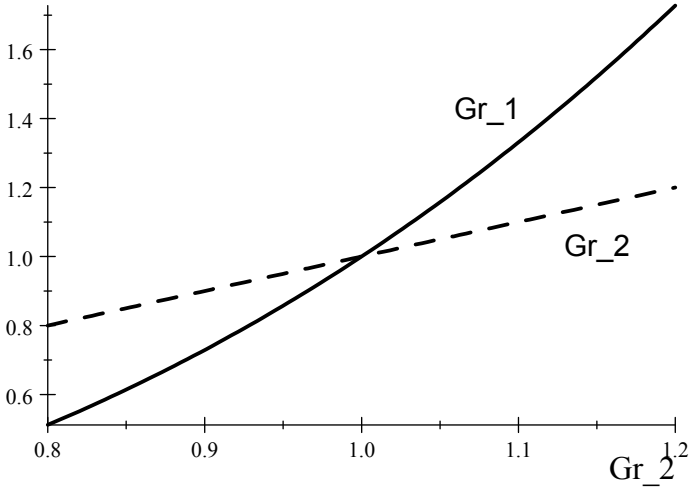


FIGURE 2.4. Equilibrium consumption growth rates of two agents with different risk-aversion. The dashed line depicts the consumption growth of the more risk averse agent, and the solid is the consumption growth of the less risk-averse. The ratio of the two CRRA is 3.

Risk-sharing carries another meaning, that of "mutuality" (Wilson, 1968). Suppose that the distribution of endowments across the population is heterogeneous, in that there are some states of the world in which some agents are better off than others. A system of complete markets would allow the agents to insure each other in a way that they would only bear the "macroeconomic risk," not the idiosyncratic risk. Let us illustrate. By Eq. (2.29),

$$\frac{\beta_j \nu_j' \left(\hat{c}_{sj}\right)}{u_j' \left(\hat{c}_{0j}\right)} = \frac{\beta_\ell \nu_\ell' \left(\hat{c}_{s\ell}\right)}{u_\ell' \left(\hat{c}_{0\ell}\right)}, \quad \text{for all agents } j, \ell, \text{ and state } s. \tag{2.30}$$

Suppose, then, that the aggregate endowment is higher in some state s than in some other s'. Then, we claim that each agent has a strictly higher consumption in state s than in s'. In particular, this means that in absence of aggregate risk (aggregate endowment being the same in each state), the agents bear no risk. Indeed, consider two distinct states s and s' and assume that the aggregate endowment satisfies $\omega_s > \omega_{s'}$. Then, there must exist an individual j^* such that $\hat{c}_{sj^*} > \hat{c}_{s'j^*}$ and, hence, $\nu_{j^*}' \left(\hat{c}_{sj^*}\right) < \nu_{j^*}' (\hat{c}_{s'j^*})$. By Eq. (2.30), then, $\nu_\ell' (\hat{c}_{s\ell}) < \nu_\ell' (\hat{c}_{s'\ell})$ and, hence, $\hat{c}_{s\ell} > \hat{c}_{s'\ell}$ for any agent ℓ. Finally, note that because the utilities $\nu_j (\cdot)$ are state-independent, the equilibrium distribution of the aggregate endowment does not depend on s.

The previous conclusions imply that the equilibrium allocations are state-independent functions of the aggregate endowment,

$$\hat{c}_{sj} = a_j \left(\omega_s\right),$$

for some strictly increasing functions $a_j (\cdot)$. Whence, the mutuality result mentioned above: the equilibrium allocations do not depend on the state of the economy if the aggregate endowment ω_s does not vary across states s.[2] The functions $a_j (\cdot)$ are known as Pareto sharing rules. Huang and Litzenberger (1988, Chapter 5) explain in detail further properties of these functions. In particular, $a_j (\cdot)$ are linear if and only if the utility functions take the following form, $\nu_i (c) = (\alpha_i + \gamma c)^{-\frac{1}{\gamma}}$, for some constants a_i and γ.

2.6.3.2 Incomplete markets

Agents' marginal rates of substitution cannot be equal if markets are incomplete—except perhaps on a negligible set of endowment distribution. The best outcome in this case, is a set of equilibria known as *constrained Pareto optima*, i.e. constrained by ... the states of nature. It might actually turn out that no constrained Pareto optima could even exist in multiperiod economies with incomplete markets. [Elaborate, in progress]

When market are incomplete, the state price vector ϕ is not unique. That is, suppose that $\phi^\top$ is an equilibrium state price. Then, all the elements of

$$\Phi = \{\phi' \in \mathbb{R}^d_{++} : (\phi' - \phi)^\top V = 0\} \tag{2.31}$$

[2]To illustrate, consider an economy with two agents (A and B) and no-aggregate risk. Pareto allocations can then be characterized within an Edgeworth's box where the axes indicate consumption in the two states s' and s. Note that the 45° line of the box is the set in which the two agents are perfecty insured across states. Moreover, the agents' indifference curves are tangent along the contract curve (i.e. the set of Pareto allocations), meaning that $\frac{dc_{s'A}}{dc_{sA}} = \frac{dc_{s'B}}{dc_{sB}}$. Because $-\frac{dc_{s'j}}{dc_{sj}} = \frac{\nu_j'(\hat{c}_{sj})P_s}{\nu_j'(\hat{c}_{s'j})P_{s'}} = \frac{\phi_{s'}}{\phi_s}$, this tangency condition is met only on the 45° line: Pareto allocations lead to mutuality.

are also equilibrium state prices. That is, there exist many equilibrium state prices consistent with absence of arbitrage. Or, in other words, there are many equilibrium state prices consistent with the same observable asset price vector S, for $\phi'^{\top}V = \phi^{\top}V = S$.

How do we proceed in this case? Introduce the following budget constraint:

$$\mathcal{C} = \left\{ c \in \mathbb{R}^{d+1}_{++} : 0 = c_0 - w_0 + \phi^{\top}\left(c^1 - w^1\right), \quad \left(c^1 - w^1\right) \in \langle V \rangle, \quad \forall \phi \in \mathbb{R}^d_{++} : S = \phi^{\top} V \right\}. \tag{2.32}$$

Because Φ in Eq. (2.31) has many elements as discussed, the budget constraint $\mathcal{C}$ does then include many constraints to account for in a context with incomplete markets: the "martingale methods" in the previous sections do not apply anymore.

Yet let $\text{Val}\,(P_I)$ be the value of the following program in the incomplete markets at hand:

$$\max_{c \in \mathcal{C}} \left[u_j\left(c_{0j}\right) + \beta_j E(\nu_j(c^1_j)) \right]. \tag{2.P$_{\text{I}}$}$$

Consider, next, the following constraint:

$$\mathcal{C}_\phi = \left\{ \begin{array}{c} c \in \mathbb{R}^{d+1}_{++} : 0 = c_0 - w_0 + \phi^{\top}\left(c^1 - w^1\right), \quad \left(c^1 - w^1\right) \in \mathbb{R}^d, \\ \text{for some given } \phi \in \mathbb{R}^d_{++} : S = \phi^{\top} V \end{array} \right\},$$

and let $\text{Val}\,(P_\phi)$ be the value of the program in some abstract complete markets case:

$$\max_{c \in \mathcal{C}_\phi} \left[u_j\left(c_{0j}\right) + \beta_j E(\nu_j(c^1_j)) \right]. \tag{2.P$_\phi$}$$

Clearly, we have, $\text{Val}\,(P_I) \leq \text{Val}\,(P_\phi)$ for all ϕ, for the constraint in the incomplete markets case, $\mathcal{C}$, is more stringent than that in any complete market setting, $\mathcal{C}_\phi$: the solution to the program in the incomplete markets case [2.P$_{\text{I}}$], must satisfy the budget constraints in $\mathcal{C}$, formed using *all* of the possible Arrow-Debreu state prices (including the Arrow-Debreu state price ϕ given in $\mathcal{C}_\phi$), as the constraint in Eq. (2.32) shows. Moreover, $(c^1 - w^1) \in \langle V \rangle$. These remarks suggest to define the following "min-max" Arrow-Debreu state price:

$$\phi^* = \arg\min_{\phi \in \Phi} \text{Val}\,(P_\phi).$$

The natural issue is to ascertain whether

$$\text{Val}\,(P_I) = \text{Val}\,(P_{\phi^*}). \tag{2.33}$$

This is indeed the case, under regularity conditions. For the characterization of ϕ^*, suppose there exists $\hat{\phi} : \text{Val}\,(P_I) = \text{Val}(P_{\hat{\phi}})$. Then, $\hat{\phi} = \phi^*$. Indeed, suppose the contrary, i.e. there exists $\phi' : \text{Val}(P_{\phi'}) < \text{Val}(P_{\hat{\phi}})$. Then, we would have,

$$\text{Val}\,(P_I) \leq \text{Val}(P_{\phi'}) < \text{Val}(P_{\hat{\phi}}) = \text{Val}\,(P_I),$$

a contradiction. Note, again, this is a characterization result of ϕ^*, not an existence proof. But as mentioned earlier, Eq. (2.33) holds true, as shown in a dynamic setting by He and Pearson (1991). Chapter 5 provides general guidance on an even more general approach to solving problems of this kind, arising in a broader context with market imperfections, where incomplete markets arise as a special case.

A final note. The min-max state price does not give rise to a "pricing probability." Rather, it is a device to solve programs arising in the context of incomplete markets. Consider, for

example, the standard case in which agents have separable utilities. The counterparts to the first order conditions in Eqs. (2.28) would be $u_j'(\hat{c}_{0j}) = \lambda_j$ and $\beta_j \nu_j'(\hat{c}_{sj}) = \lambda_j m_{s,j}^*$, where $m_{s,j}^*$ is the min-max stochastic discount factor for agent j—we have as many min-max stochastic discount factors as agents. An equilibrium would then have to be found by imposing market clearing. Naturally, a single-agent case is much easier, with the pricing probability being then the min-max probability, i.e. that resulting from the min-max state price ϕ^*.

2.6.3.3 Financial innovation

Introduction of derivatives to complete the markets.

2.6.3.4 Computation of the equilibrium

The first order conditions satisfied by any agent's program are:

$$\hat{c}_{0j} = I_j(\lambda_j), \quad \hat{c}_{sj} = H_j\left(\beta_j^{-1}\lambda_j m_s\right), \tag{2.34}$$

where I_j and H_j denote the inverse functions of u_j' and ν_j'. By the assumptions we made on u_j and ν_j', I_j and H_j inherit the same properties of u_j' and ν_j'. By replacing these functions into the constraint,

$$0 = \hat{c}_{0j} - w_{0j} + E\left[m \cdot (\hat{c}_j^1 - w_j^1)\right] = I_j(\lambda_j) - w_{0j} + \sum_{s=1}^{d} P_s\left[m_s \cdot \left(H_j(\beta_j^{-1}\lambda_j m_s) - w_{sj}\right)\right].$$

Define the function,

$$z_j(\lambda_j) \equiv I_j(\lambda_j) + E\left[m H_j(\beta_j^{-1}\lambda_j m)\right] = w_{0j} + E\left(m \cdot w_j^1\right).$$

We see that $\lim_{x\to 0} z(x) = \infty$, $\lim_{x\to\infty} z(x) = 0$ and $z'(x) < 0$. Therefore, there exists a unique solution for λ_j:

$$\lambda_j \equiv \Lambda_j\left[w_{0j} + E(m \cdot w_j^1)\right],$$

where $\Lambda(\cdot)$ denotes the inverse function of z. By replacing back into Eqs. (2.34), we obtain:

$$\hat{c}_{0j} = I_j\left(\Lambda_j\left(w_{0j} + E(m \cdot w_j^1)\right)\right), \quad \hat{c}_{sj} = H_j\left(\beta_j^{-1} m_s \Lambda_j\left(w_{0j} + E(m \cdot w_j^1)\right)\right).$$

To determine the general equilibrium, we need to pin down the stochastic discounting factor, m. We have d unknowns (m_s, $s = 1, \cdots, d$), and $d+1$ equilibrium conditions (holding in the $d+1$ markets). By Walras' law, only d of these are independent. Consider the equilibrium conditions in the d markets at the second period:

$$g_s\left(m_s; (m_{s'})_{s'\neq s}\right) \equiv \sum_{j=1}^{n} H_j\left(\beta_j^{-1} m_s \Lambda_j\left(w_{0j} + E(m \cdot w_j^1)\right)\right) = \sum_{j=1}^{n} w_{sj} \equiv w_s, \ s = 1, \cdots, d.$$

These conditions determine the kernel $(m_s)_{s=1}^d$ which leads to compute prices and equilibrium allocations. Finally, once the optimal $\hat{c}_s$ are computed, for $s = 0, 1, \cdots, d$, the portoflio $\hat{\theta}$ generated them can be inferred through $\hat{\theta} = V^{-1}(\hat{c}^1 - w^1)$.

2.7 Consumption-CAPM

We refer a theory to be "consumption-based," if the equilibrium expected returns, or the "risk-premiums," are determined through optimal consumption choices. Under certain conditions, studied in this section, one of its prediction is that these risk-premiums are high for securities that pay high returns when consumption is high (i.e. when we don't need high returns) and low returns when consumption is low (i.e. when we need high returns). In fact, while this statement is quite often used to explain, in a nutshell, the quintessence of the Consumption-CAPM, it might be misleading in economies where assets are in zero net supply, such as those we studied in previous sections.

Consider, for example, Eq. (2.27). It is an asset pricing equation (2.27), which states that for every asset delivering a gross return,

$$1 = E(m \cdot \tilde{R}), \tag{2.35}$$

where m is some pricing kernel. Let us elaborate on Eq. (2.35), so as to obtain a representation of the expected return on any asset. Naturally, for a riskless asset, $1 = E(m \cdot R)$, which combined with Eq. (2.35) leaves $E[m \cdot (\tilde{R} - R)] = 0$, and by rearranging terms,

$$E(\tilde{R}) = R - \frac{cov(\nu'(w^+), \tilde{R})}{E\left[\nu'(w^+)\right]}. \tag{2.36}$$

Eq. (2.36) can be rewritten as,

$$E(\tilde{R}) - R = -\frac{cov(m, \tilde{R})}{E(m)} = -R \cdot cov(m, \tilde{R}). \tag{2.37}$$

We also know from previous sections, that in economies with a single agent, or in economies with complete markets, this pricing kernel is given by:

$$m_s = \beta \frac{\nu'(w_s)}{u'(w_0)}.$$

That is, the pricing kernel directly links to optimal consumption choice—we are dealing with a consumption-based CAPM. However, the risk-premium, $-R \cdot cov(m, \tilde{R})$, does merely depend on how the pricing kernel co-varies with the asset returns, and not directly on the cyclical properties of dividends. For example, were endowments be constant in the second period, the risk-premium would be zero! The next section aims to clarify how we should set the Consumption-CAPM as an appropriate context to think about the cyclical properties of asset returns.

2.7.1 Risk-neutral pricing and macroeconomic risks

Up to now, we have considered assets in zero net supply. We now generalize previous findings to the relevant case where some assets could be in positive supply, which is a case applying many times in these lectures. For simplicity, consider an economy which only has an asset in positive supply equal to θ_0, which is the endowment of a representative agent. The budget constraints of the agent over his two period of life are:

$$c_1 + S\theta = S\theta_0 + w \quad \text{and} \quad \tilde{c}_2 = \tilde{d}\theta + \tilde{w}, \tag{2.38}$$

where S and θ are the risky asset price and demand, w is the initial endowment, c_1 is first period consumption, $\tilde{w}$ is the random endowment for the good in the second period, $\tilde{d}$ is the random dividend promised by the asset in the second period, and $\tilde{c}_2$ is second-period consumption. We assume that $\tilde{d} > 0$.

Note that this model is one with incomplete markets, because the agent can invest using only one asset, and yet there are two sources of risk: the asset dividend *and* endowment. Based on this model, Chapter 9 explains the extent to which incomplete markets might help rationalize the equity premium we observe, in practice, within a consumption-based perspective.

In equilibrium, $\theta = \theta_0$, such that the asset price equation is,

$$S = \frac{E\left(\beta v'(\tilde{d}\theta_0 + \tilde{w})\tilde{d}\right)}{u'(w)}. \tag{2.39}$$

The previous sections have studied the special case where the asset is in zero-net supply, $\theta_0 = 0$, as mentioned. Eq. (2.39) shows that in general, and assuming $\tilde{d} > 0$, the asset price S is decreasing with asset supply, θ_0. It is not a mere supply-demand effect but, rather a "risk-premium effect, as we now explain.

We know that the state price for state s is, consistently with results in Section 2.6.3,

$$\phi_s = \frac{\beta v'(d_s\theta_0 + w_s)}{u'(w)} P_s. \tag{2.40}$$

It is, as usual, the present consumption we are willing to give up, today, to obtain additional consumption tomorrow, in state s. Due to decreasing marginal utility, we have that consumption demand in state s, $\hat{c}_s$, is decreasing in ϕ_s, $\hat{c} : \phi_s = \frac{\beta v'(\hat{c}_s)}{u'(\hat{c}_0)}$, for $s = 1, \cdots, d$, such that $\hat{c}_s$ is decreasing in ϕ_s. Therefore, in equilibrium, ϕ_s is decreasing in the good supply in state s, $d_s\theta_0 + w_s$.

The previous "scarcity effect," by which a shrinkage in supply for state s determines an increase in the state price ϕ_s, is quite simple yet powerful. As we know, we price assets through Arrow-Debreu assets, by assigning high weights, i.e. by utilizing high Arrow-Debreu prices, to the bad states of nature—a scarcity effect. Therefore, an increase in the asset supply θ_0, being uniform across all states of nature, mitigates the previous scarcity effect and, then, lowers the entire set of Arrow-Debreu prices, thereby reducing the asset price, S. In other words, the asset price decreases because consumption become cheaper in each state of the world, due to this scarcity channel, which requires less demand for savings.

We can approach this problem from a different angle. Note that we can rewrite Eq. (2.39) as follows,

$$S = \frac{E(\tilde{d})}{R(\theta_0)} + \frac{\beta}{u'(w)} cov\left(v'(\tilde{d}\theta_0 + \tilde{w}), \tilde{d}\right), \quad R(\theta_0) = \frac{u'(w)}{E\left(\beta v'(\tilde{d}\theta_0 + \tilde{w})\right)} \tag{2.41}$$

The first term is an actuarial evaluation of the asset. The second term is a risk-premium, which comes as a discount to the initial actuarial evaluation, given the assumptions of decreasing marginal utility, and $cov(\tilde{w}, \tilde{d}) > 0$. The previous equation reveals that an increase in θ_0 entails an heavier discounting effect, as the interest rate, $R(\theta_0)$, increases with θ_0, as a result of a decreased demand for savings. The second term is negative, as explained—the asset pays off exactly when it is not neeeded. However, it becomes thinner and thinner as θ_0 increases,

reflecting the fact that as θ_0 increases, and given the assumption that $\tilde{d} > 0$, the agent bears less and less risk: even over poor realization of the states would guarantee handsome overall returns, when the asset supply is large. As we know from the discussion of Eq. (2.39), however, the discounting effect dominates, with asset prices falling as θ_0 increases.

Note that the previous reasoning does not apply, once we assume the asset is in zero net supply, $\theta_0 = \theta = 0$, because as Eq. (2.38) makes clear, capital markets do not affect equilibrium consumption anyway. Instead, in this case, only a "pure endowment scarcity" channel leads to low Arrow-Debreu prices for the bad states of the world: those states where endowments are lower command higher Arrow-Debreu prices, at least provided that $cov(\tilde{w}, \tilde{d}) > 0$, such that the second term in Eq. (2.41) is negative.[3]

All in all, Arrow-Debreu prices are independent of the asset returns, when the asset is in zero net supply. When, instead, $\theta_0 > 0$ and, still, $cov(\tilde{w}, \tilde{d}) > 0$, one additional scarcity channel is activated, which consists in a drop in consumption generated by a drop in the dividend. Suppose, for example, that we enter bad times, when w_s falls as a result for example of a job loss. Not only, then, does w_s fall, the dividend also likely falls, and exactly when you would need it to compensate for the fall in w_s. It is in this sense, which we may say that asset investing might make consumption even more volatile.

Part II reviews the literature on the "equity premium puzzle," the challenges that consumption-based explanations of the equity premium have to be consistent with the observed equity premium.

2.7.2 The beta relation

Suppose there is a $\tilde{R}_m$ such that

$$\tilde{R}_m = -\gamma^{-1} \cdot \nu'(w_s), \quad \text{all } s.$$

In this case,

$$E(\tilde{R}) = R + \frac{\gamma \cdot cov(\tilde{R}_m, \tilde{R})}{E[\nu'(w^+)]} \quad \text{and} \quad E(\tilde{R}_m) = R + \frac{\gamma \cdot var(\tilde{R}_m)}{E[\nu'(w^+)]}.$$

These relations can be combined to yield,

$$E(\tilde{R}) - R = \beta \cdot [E(\tilde{R}_m) - R], \quad \beta \equiv \frac{cov(\tilde{R}_m, \tilde{R})}{var(\tilde{R}_m)}.$$

2.7.3 CCAPM & CAPM

Let $\tilde{R}^p$ be the portfolio return which is the most highly correlated with the pricing kernel m. We have,

$$E(\tilde{R}^p) - R = -R \cdot cov(m, \tilde{R}^p). \tag{2.42}$$

Using Eqs. (2.37) and (2.42),

$$\frac{E(\tilde{R}) - R}{E(\tilde{R}^p) - R} = \frac{cov(m, \tilde{R})}{cov(m, \tilde{R}^p)},$$

[3]Note, finally, a special instance of this model, arising when $\theta_0 = 0$ and $\tilde{w} = \bar{w}$, in which case consumption is not random in the second period, which means that the agent does not bear any risk. The price is, then, simply, the discounted expected dividend, $S = \frac{E(\tilde{d})}{R}$.

and by rearranging terms,

$$E(\tilde{R}) - R = \frac{\beta_{\tilde{R},m}}{\beta_{\tilde{R}^p,m}}[E(\tilde{R}^p) - R] \quad [\textbf{CCAPM}].$$

If $\tilde{R}^p$ is perfectly correlated with m, i.e. if there exists $\gamma : \tilde{R}^p = -\gamma m$, then

$$\beta_{\tilde{R},m} = -\gamma \frac{cov(\tilde{R}^p, \tilde{R})}{var(\tilde{R}^p)} \quad \text{and} \quad \beta_{\tilde{R}^p,m} = -\gamma$$

and then

$$E(\tilde{R}) - R = \beta_{\tilde{R},\tilde{R}^p}[E(\tilde{R}^p) - R] \quad [\textbf{CAPM}].$$

This is not the only way the CAPM obtains. Chapter 7 explains that the CAPM also obtains through the so-called "maximum correlation portfolio," which is the portfolio that is the most highly correlated with the pricing kernel m.

2.8 Infinite horizon

We consider d states of the nature and $m = d$ Arrow securities. We write a unified budget constraint, as in the valuation equilibria approach of Debreu (1954). We have,

$$\begin{cases} p_0 \left(c_0 - w_0\right) = -S^{(0)}\theta^{(0)} = -\sum_{i=1}^{m} S_i^{(0)}\theta_i^{(0)} \\ p_s^1 \left(c_s^1 - w_s^1\right) = \theta_s^{(0)}, \ s = 1, \cdots, d \end{cases}$$

or,

$$p_0(c_0 - w_0) + \sum_{i=1}^{m} S_i^{(0)} \left[p_i^1 \left(c_i^1 - w_i^1\right)\right] = 0.$$

The previous relation holds in a two-period economy. In a multiperiod economy, in the second period (as in the following periods) agents save indefinitively for the future. In the appendix, we show that,

$$0 = E\left[\sum_{t=0}^{\infty} m_{0,t} \cdot p^t \left(c^t - w^t\right)\right], \tag{2.43}$$

where $m_{0,t}$ are the state prices. From the perspective of time 0, at time t there exist d^t states of nature and, thus, d^t possible prices.

2.9 Further topics on incomplete markets

2.9.1 Nominal assets and real indeterminacy of the equilibrium

The equilibrium is a set of prices $(\hat{p}, \hat{S}) \in \mathbb{R}_{++}^{m\cdot(d+1)} \times \mathbb{R}_{++}^{a}$ such that:

$$0 = \sum_{j=1}^{n} e_{0j}(\hat{p}, \hat{S}), \qquad 0 = \sum_{j=1}^{n} e_{1j}(\hat{p}, \hat{S}), \qquad 0 = \sum_{j=1}^{n} \theta_j(\hat{p}, \hat{S}),$$

where the previous functions are the results of optimal plans of the agents. This system has $m \cdot (d+1) + a$ equations and $m \cdot (d+1) + a$ unknowns, where $a \leq d$. Let us aggregate the constraints of the agents,

$$p_0 \sum_{j=1}^{n} e_{0j} = -S \sum_{j=1}^{n} \theta_j, \qquad p_1 \square \sum_{j=1}^{n} e_{1j} = B \sum_{j=1}^{n} \theta_j.$$

Suppose the financial markets clearing condition is satisfied, i.e. $\sum_{j=1}^{n} \theta_j = 0$. Then,

$$\begin{cases} 0 = p_0 \sum\limits_{j=1}^{n} e_{0j} \equiv p_0 e_0 = \sum\limits_{\ell=1}^{m} p_0^{(\ell)} e_0^{(\ell)} \\ \mathbf{0}_d = p_1 \square \sum\limits_{j=1}^{n} e_{1j} \equiv p_1 \square e_1 = \left[\sum\limits_{\ell=1}^{m} p_1^{(\ell)}(\omega_1) e_1^{(\ell)}(\omega_1), \cdots, \sum\limits_{\ell=1}^{m} p_1^{(\ell)}(\omega_d) e_1^{(\ell)}(\omega_d) \right]^\top \end{cases}$$

Therefore, there is one redundant equation for each state of nature, or $d+1$ redundant equations, in total. As a result, the equilibrium has less independent equations $(m \cdot (d+1) - 1)$ than unknowns $(m \cdot (d+1) + d)$, i.e., an indeterminacy degree equal to $d+1$. This result does not rely on whether markets are complete or not. In a sense, it is even not an indeterminacy result when markets are complete, as we may always assume agents would organize the exchanges at the beginning. In this case, onle the suitably normalized Arrow-Debreu state prices would matter for agents.

The previous indeterminacy can be reduced to $d-1$, as we may use two additional homogeneity relations. To pin down these relations, let us consider the budget constaint of each agent j,

$$p_0 e_{0j} = -S\theta_j, \quad p_1 \square e_{1j} = B\theta_j.$$

The first-period constraint is still the same if we multiply the spot price vector p_0 and the financial price vector S by a positive constant, λ (say). In other words, if $(\hat{p}_0, \hat{p}_1, \hat{S})$ is an equilibrium, then, $(\lambda \hat{p}_0, \hat{p}_1, \lambda \hat{S})$ is also an equilibrium, which delivers a first homogeneity relation. To derive the second homogeneity relation, we multiply the spot prices of the second period by a positive constant, λ and increase at the same time the first period agents' purchasing power, by dividing each asset price by the same constant, as follows:

$$p_0 e_{0j} = -\frac{S}{\lambda} \lambda \theta_j, \quad \lambda p_1 \square e_{1j} = B \lambda \theta_j.$$

Therefore, if $(\hat{p}_0, \hat{p}_1, \hat{S})$ is an equilibrium, then, $\left(\hat{p}_0, \lambda \hat{p}_1, \frac{\hat{S}}{\lambda}\right)$ is also an equilibrium.

2.9.2 Nonneutrality of money

The previous indeterminacy arises because financial contracts are *nominal*, i.e. the asset payoffs are expressed in terms of some *unité de compte* that, among other things, we did not make precise. Such an indeterminacy vanishes if we were to consider *real* contracts, i.e. contracts with payoffs expressed in terms of the goods. To show this, note that in the presence of real contracts, the agents' constraints are

$$\begin{cases} p_0 e_{0j} = -S\theta_j \\ p_1(\omega_s) e_{1j}(\omega_s) = p_1(\omega_s) A_s \theta_j, \ s = 1, \cdots, d \end{cases}$$

where $A_s = [A_s^1, \cdots, A_s^a]$ is the $m \times a$ matrix of the real payoffs. The previous constraint now reveals how to "recover" $d+1$ homogeneity relations. For each strictly positive vector $\lambda = [\lambda_0, \lambda_1 \cdots, \lambda_d]$, we have that if $[\hat{p}_0, S, p_1(\omega_1), \cdots, p_1(\omega_s), \cdots, p_1(\omega_d)]$ is an equilibrium, then, $[\lambda_0\hat{p}_0, \lambda_0 S, p_1(\omega_1), \cdots, p_1(\omega_s), \cdots, p_1(\omega_d)]$ is also an equilibrium, and so is $[\hat{p}_0, S, p_1(\omega_1), \cdots, \lambda_s p_1(\omega_s), \cdots, p_1(\omega_d)]$, for λ_s, $s = 1, \cdots, d$.

As is clear, the distinction between nominal and real assets has a precise meaning, when one considers a multi-commodity economy. Even in this case, however, such a distinctions is not very interesting without a suitable introduction of a *unité de compte*. These considerations led Magill and Quinzii (1992) to solve the indeterminacy while still remaining in a framework with nominal assets. They simply propose to introduce money as a mean of exchange. The indeterminacy can then be resolved by "fixing" the prices via the $d+1$ equations defining the money market equilibrium in all states of nature:

$$M_s = p_s \cdot \sum_{j=1}^{n} w_{sj}, \ s = 0, 1, \cdots, d.$$

Magill and Quinzii showed that the monetary policy $(M_s)_{s=0}^d$ is generically nonneutral.

2.10 Appendix 1

In this appendix we prove that the program [2.P1] has a unique maximum. Indeed, suppose on the contrary that we have two maxima:

$$\bar{c} = (\bar{c}_{1j}, \cdots, \bar{c}_{mj}) \quad \text{and} \quad \overline{\bar{c}} = \left(\overline{\bar{c}}_{1j}, \cdots, \overline{\bar{c}}_{mj}\right).$$

These two maxima would satisfy $u_j(\bar{c}) = u_j(\overline{\bar{c}})$, with $\sum_{i=1}^m p_i \bar{c}_{ij} = \sum_{i=1}^m p_i \overline{\bar{c}}_{ij} = R_j$. To check that this claim is correct, suppose on the contrary that $\sum_{i=1}^m p_i \overline{\bar{c}}_{ij} < R_j$. Then, the consumption bundle,

$$\overline{\bar{c}} = \left(\overline{\bar{c}}_{1j} + \varepsilon, \cdots, \overline{\bar{c}}_{mj}\right), \quad \varepsilon > 0,$$

would be preferred to $\overline{\bar{c}}$, by Assumption 2.1, and, at the same time, it would hold that, for sufficiently small ε,

$$\sum_{i=1}^m p_i \overline{\bar{c}}_{ij} = \varepsilon p_1 + \sum_{i=1}^m p_i \overline{\bar{c}}_{ij} < R_j.$$

[Indeed, we have, $A \equiv \sum_{i=1}^m p_i \overline{\bar{c}}_{ij}$. $A < R_j \Rightarrow \exists \varepsilon > 0 : A + \varepsilon p_1 < R_j$. E.g., $\varepsilon p_1 = R_j - A - \eta$, $\eta > 0$. The condition is then: $\exists \eta > 0 : R_j - A > \eta$.] Hence, $\overline{\bar{c}}$ would be a solution to [2.P1], thereby contradicting the optimality of $\overline{\bar{c}}$. Therefore, the existence of two optima would imply a full use of resources. Next, consider a point y lying between $\bar{c}$ and $\overline{\bar{c}}$, viz $y = \alpha \bar{c} + (1-\alpha)\overline{\bar{c}}$, $\alpha \in (0,1)$. By Assumption 2.1,

$$u_j(y) = u_j\left(\alpha \bar{c} + (1-\alpha)\overline{\bar{c}}\right) > u_j(\bar{c}) = u_j(\overline{\bar{c}}).$$

Moreover,

$$\sum_{i=1}^m p_i y_i = \sum_{i=1}^m p_i \left(\alpha \bar{c}_{ij} + (1-\alpha)\overline{\bar{c}}_{ij}\right) = \alpha \sum_{i=1}^m p_i \bar{c}_{ij} + \sum_{i=1}^m \overline{\bar{c}}_{ij} - \alpha \sum_{i=1}^m \overline{\bar{c}}_{ij} = \alpha R_j + R_j - \alpha R_j = R_j.$$

Hence, $y \in B_j(p)$ and is also strictly preferred to $\bar{c}$ and $\overline{\bar{c}}$, which means that $\bar{c}$ and $\overline{\bar{c}}$ are not optima, as initially conjectured. This establishes uniqueness of the solution to [2.P1].

2.11 Appendix 2: Proofs of selected results

We first provide a useful result, a well-known theorem on the separation of two convex sets. We use this theorem to deal with the proof of the second welfare theorem (Theorem 2.4) and the existence of state prices tying up all asset prices (Theorem 2.10). A final proof we provide in this appendix is that of Eq. (2.43).

Minkowski's separation theorem. *Let A and B be two non-empty convex subsets of $\mathbb{R}^d$. If A is closed, B is compact and $A\bigcap B=\emptyset$, then there exists a $\phi\in\mathbb{R}^d$ and two real numbers d_1,d_2 such that:*

$$a^\top\phi\leq d_1<d_2\leq b^\top\phi,\quad \forall a\in A,\ \forall b\in B.$$

We are now ready to prove Theorems 2.4 and 2.10.

Proof of Theorem 2.4. Let $\bar{c}$ be a Pareto optimum and $\tilde{B}_j=\left\{c^j:u_j(c^j)>u_j(\bar{c}^j)\right\}$. Let us consider the two sets $\tilde{B}=\bigcup_{j=1}^n\tilde{B}_j$ and $A=\left\{(c^j)_{j=1}^n:c^j\geq 0\,\forall j,\ \sum_{j=1}^n c^j=w\right\}$. A is the set of all possible combinations of feasible allocations. By the definition of a Pareto optimum, there are no elements in A that are simultaneously in $\tilde{B}$, or $A\bigcap\tilde{B}=\emptyset$. In particular, this is true for all compact subsets B of $\tilde{B}$, or $A\bigcap B=\emptyset$. Because A is closed, then, by the Minkowski's separating theorem, there exists a $p\in\mathbb{R}^m$ and two distinct numbers d_1, d_2 such that

$$p^\top a\leq d_1<d_2\leq p^\top b,\quad \forall a\in A,\ \forall b\in B.$$

This means that for all allocations $\left(c^j\right)_{j=1}^n$ preferred to $\bar{c}$, we have:

$$p^\top\sum_{j=1}^n w^j<p^\top\sum_{j=1}^n c^j,$$

or, by replacing $\sum_{j=1}^n w^j$ with $\sum_{j=1}^n\bar{c}^j$,

$$p^\top\sum_{j=1}^n\bar{c}^j<p^\top\sum_{j=1}^n c^j. \tag{2A.1}$$

Next we show that $p>0$. Let $\bar{c}_i=\sum_{j=1}^n\bar{c}_{ij}$, $i=1,\cdots,m$, and partition $\bar{c}=(\bar{c}_1,\cdots,\bar{c}_m)$. Let us apply the inequality in (2A.1) to $\bar{c}\in A$ and, for $\mu>0$, to $c=(\bar{c}_1+\mu,\cdots,\bar{c}_m)\in B$. We have $p_1\mu>0$, or $p_1>0$. By reiterating the argument, $p_i>0$ for all i. Finally, we choose $c^j=\bar{c}^j+\mathbf{1}_m\frac{\epsilon}{n}$, $j=2,\cdots,n$, $\epsilon>0$ in (2A.1), $p^\top\bar{c}^1<p^\top c^1+p^\top\mathbf{1}_m\epsilon$ or,

$$p^\top\bar{c}^1<p^\top c^1,$$

for ϵ sufficiently small. This means that $u_1(c^1)>u_1(\bar{c}^1)\Rightarrow p^\top c^1>p^\top\bar{c}^1$. This means that $\bar{c}^1=\arg\max_{c^1}u_1(c^1)$ s.t. $p^\top c^1=p^\top\bar{c}^1$. By symmetry, $\bar{c}^j=\arg\max_{c^j}u_j(c^j)$ s.t. $p^\top c^j=p^\top\bar{c}^j$ for all j. ‖

Proof of Theorem 2.10. The condition in (2.18) holds for any compact subset of $\mathbb{R}_+^{d+1}$, and therefore it holds when it is restricted to the unit simplex in $\mathbb{R}_+^{d+1}$,

$$\langle W\rangle\bigcap\mathcal{S}^d=\{0\}.$$

By the Minkowski's separation theorem, $\exists\tilde{\phi}\in\mathbb{R}^{d+1}:w^\top\tilde{\phi}\leq d_1<d_2\leq\sigma^\top\tilde{\phi}$, $w\in\langle W\rangle$, $\sigma\in\mathcal{S}^d$. By walking through the simplex boundaries, one finds that $d_1<\tilde{\phi}_s$, $s=1,\cdots,d$. On the other hand,

$0 \in \langle W \rangle$, which reveals that $d_1 \geq 0$, and $\tilde{\phi} \in \mathbb{R}^{d+1}_{++}$. Next we show that $w^\top \tilde{\phi} = 0$. Assume the contrary, i.e. $\exists w_* \in \langle W \rangle$ that satisfies at the same time $w_*^\top \tilde{\phi} \neq 0$. In this case, there would be a real number ϵ with $\text{sign}(\epsilon) = \text{sign}(w_*^\top \tilde{\phi})$ such that $\epsilon w_* \in \langle W \rangle$ and $\epsilon w_*^\top \tilde{\phi} > d_2$, a contradiction. Therefore, we have $0 = \tilde{\phi}^\top W\theta = (\tilde{\phi}^\top(-S\ V)^\top)\theta = (-\tilde{\phi}_0 S + \tilde{\phi}_{(d)}^\top V)\theta$, $\forall \theta \in \mathbb{R}^m$, where $\tilde{\phi}_{(d)}$ contains the last d components of $\tilde{\phi}$. Whence $S = \phi^\top V$, where $\phi^\top = \left(\frac{\tilde{\phi}_1}{\tilde{\phi}_0}, \cdots, \frac{\tilde{\phi}_d}{\tilde{\phi}_0}\right)$.

The proof of the converse is immediate (hint: multiply by θ): **shown in further notes**.

The proof of the second part is the following one. We have that "each point of $\mathbb{R}^{d+1}$ is equal to each point of $\langle W \rangle$ plus each point of $\langle W \rangle^\perp$," or $\dim \langle W \rangle + \dim \langle W \rangle^\perp = d+1$. Since $\dim \langle W \rangle = \text{rank}(W)$, $\dim \langle W \rangle^\perp = d+1-\dim \langle W \rangle$, and since $S = \phi^\top V$ in the absence of arbitrage opportunities, $\dim \langle W \rangle = \dim \langle V \rangle = m$, whence:

$$\dim \langle W \rangle^\perp = d - m + 1.$$

In other terms, before we showed that $\exists \tilde{\phi} : \tilde{\phi}^\top W = 0$, or $\tilde{\phi}^\top \in \langle W \rangle^\perp$. Whence $\dim \langle W \rangle^\perp \geq 1$ in the absence of arbitrage opportunities. The previous relation provides more information. Specifically, $\dim \langle W \rangle^\perp = 1$ if and only if $d = m$. In this case, $\dim\{\tilde{\phi} \in \mathbb{R}^{d+1}_+ : \tilde{\phi}^\top W = 0\} = 1$, which means that the relation $-\tilde{\phi}_0 S + \tilde{\phi}_d^\top V = 0$ also holds true for $\tilde{\phi}^* = \tilde{\phi} \cdot \lambda$, for every positive scalar λ, but there are no other possible candidates. Therefore, $\phi^\top = \left(\frac{\tilde{\phi}_1}{\tilde{\phi}_0}, \cdots, \frac{\tilde{\phi}_d}{\tilde{\phi}_0}\right)$ is such that $\phi = \phi(\lambda)$, and then it is unique. By a similar reasoning, $\dim\{\tilde{\phi} \in \mathbb{R}^{d+1}_+ : \tilde{\phi}^\top W = 0\} = d-m+1 \Rightarrow \dim\left\{\phi \in \mathbb{R}^d_{++} : S = \phi^\top V\right\} = d-m$. ‖

Proof of Eq. (2.43). Let $S^{(2)(\ell)}_{s',s}$ be the price at $t = 2$ in state s' if the state in $t = 1$ is s, for the Arrow-Debreu security promising 1 unit of numéraire in state ℓ at $t = 3$. Let $S^{(2)}_{s',s} = [S^{(2)(1)}_{s',s}, \cdots, S^{(2)(m)}_{s',s}]$. Let $\theta_i^{(1)(s)}$ be the quantity purchased at $t = 1$ in state i of Arrow-Debreu securities promising 1 unit of numéraire if s at $t = 2$. Let $p^2_{s,i}$ be the price of the good at $t = 2$ in state s if the previous state at $t = 1$ was i. The budget constraint is

$$\begin{cases} p_0\left(c_0 - w_0\right) = -S^{(0)}\theta^{(0)} = -\sum_{i=1}^m S^{(0)(i)}\theta^{(0)(i)} \\ p^1_s\left(c^1_s - w^1_s\right) = \theta^{(0)(s)} - S^{(1)}_s\theta^{(1)}_s = \theta^{(0)(s)} - \sum_{i=1}^m S^{(1)(i)}_s\theta^{(1)(i)}_s,\ s = 1, \cdots, d. \end{cases}$$

where $S^{(1)(i)}_s$ is the price to be paid at time 1 and in state s, for an Arrow-Debreu security yielding 1 unit of numéraire in state i at time 2. The previous two equations can be combined to leave,

$$p_0\left(c_0 - w_0\right) = -\textstyle\sum_i S^{(0)(i)}\left[p^1_i\left(c^1_i - w^1_i\right) + S^{(1)}_i\theta^{(1)}_i\right],$$

or,

$$\begin{aligned} 0 &= p_0\left(c_0 - w_0\right) + \textstyle\sum_i S^{(0)(i)}p^1_i\left(c^1_i - w^1_i\right) + \sum_i S^{(0)(i)}S^{(1)}_i\theta^{(1)}_i \\ &= p_0\left(c_0 - w_0\right) + \textstyle\sum_i S^{(0)(i)}p^1_i\left(c^1_i - w^1_i\right) + \sum_i S^{(0)(i)}\sum_i S^{(1)(j)}_i\theta^{(1)(j)}_i \\ &= p_0\left(c_0 - w_0\right) + \textstyle\sum_i S^{(0)(i)}p^1_i\left(c^1_i - w^1_i\right) + \sum_i\sum_j S^{(0)(i)}S^{(1)(j)}_i\theta^{(1)(j)}_i. \end{aligned} \tag{2A.2}$$

At time 2,

$$p^2_{s,i}\left(c^2_{s,i} - w^2_{s,i}\right) = \theta^{(1)(s)}_i - S^{(2)}_{s,i}\theta^{(2)}_{s,i} = \theta^{(1)(s)}_i - \sum_{\ell=1}^m S^{(2)(\ell)}_{s,i}\theta^{(2)(\ell)}_{s,i},\ s = 1, \cdots, d, \tag{2A.3}$$

where $S^{(2)}_{s,i}$ denotes the price vector to be paid at $t=2$ in state s if the state at $t=1$ is i, for the Arrow-Debreu securities expiring at $t=3$, with remaining notation being straightforward.

Plugging Eq. (2A.3) into Eq. (2A.2) leaves:

$$\begin{aligned}0 &= p_0\left(c_0-w_0\right)+\textstyle\sum_i S^{(0)(i)}p_i^1\left(c_i^1-w_i^1\right)+\sum_i\sum_j S^{(0)(i)}S_i^{(1)(j)}\left[p_{j,i}^2\left(c_{j,i}^2-w_{j,i}^2\right)+S_{j,i}^{(2)}\theta_{j,i}^{(2)}\right]\\ &= p_0\left(c_0-w_0\right)+\textstyle\sum_i S^{(0)(i)}p_i^1\left(c_i^1-w_i^1\right)+\sum_i\sum_j S^{(0)(i)}S_i^{(1)(j)}p_{j,i}^2(c_{j,i}^2-w_{j,i}^2)\\ &+\textstyle\sum_i\sum_j\sum_\ell S^{(0)(i)}S_i^{(1)(j)}S_{j,i}^{(2)(\ell)}\theta_{j,i}^{(2)(\ell)}. \qquad (2A.4)\end{aligned}$$

In the absence of arbitrage, $\exists\phi_{t+1,s'}\in\mathbb{R}^d_{++}$, which is the state prices vector for $t+1$ if the state in t is s', such that:

$$S^{(t)(\ell)}_{s',s}=\phi'_{t+1,s'}\cdot e_\ell,\ \ell=1,\cdots,m,$$

where $e_\ell\in\mathbb{R}^d_+$ and has all zeros except in the ℓ-th component which is 1. Next, we restate the previous relation in terms of the kernel $m_{t+1,s'}=(m^{(\ell)}_{t+1,s'})^d_{\ell=1}$ and the probability distribution $P_{t+1,s'}=(P^{(\ell)}_{t+1,s'})^d_{\ell=1}$ of the events in $t+1$ when the state in t is s':

$$S^{(t)(\ell)}_{s',s}=m^{(\ell)}_{t+1,s'}\cdot P^{(\ell)}_{t+1,s'},\quad \ell=1,\cdots,m. \qquad (2A.5)$$

Eq. (2.43) in the main text follows by replacing Eq. (2A.5) into Eq. (2A.4), and by imposing the transversality condition:

$$\sum_{\ell_1=1}^m\sum_{\ell_2=1}^m\sum_{\ell_3=1}^m\sum_{\ell_4=1}^m\cdots\sum_{\ell_t=1}^m\cdots S^{(0)(\ell_1)}S_{\ell_1}^{(1)(\ell_2)}S_{\ell_2,\ell_1}^{(2)(\ell_3)}S_{\ell_3,\ell_2}^{(3)(\ell_4)}\cdots S_{\ell_{t-1},\ell_{t-2}}^{(t-1)(\ell_t)}\cdots\underset{t\to\infty}{\to}0,$$

we get eq. (2.43). ‖

2.12 Appendix 3: The multicommodity case

The multicommodity case is interesting, but at the same time is extremely delicate to deal with when markets are incomplete. While standard regularity conditions ensure the existence of an equilibrium in the static and complete markets case, only "generic" existence results are available for the incoplete markets cases. Hart (1974) built up well-chosen examples in which there exist sets of endowments distributions for which no equilibrium can exist. However, Duffie and Shafer (1985) showed that such sets have zero measure, which justifies the terminology of "generic" existence.

Here we only provide a derivation of the contraints. m_t commodities are traded in period t ($t = 0, 1$). The states of nature in the second period are d, and the number of traded assets is a. The first period budget constraint is:

$$p_0 e_{0j} = -S\theta_j,\ e_{0j} \equiv c_{0j} - w_{0j}$$

where $p_0 = (p_0^{(1)}, \cdots, p_0^{(m_1)})$ is the first period price vector, $e_{0j} = (e_{0j}^{(1)}, \cdots, e_{0j}^{(m_1)})'$ is the first period excess demands vector, $S = (S_1, \cdots, S_a)$ is the financial asset price vector, and $\theta_j = (\theta_{1j}, \cdots, \theta_{aj})'$ is the vector of assets quantities that agent j buys at the first period.

The second period budget constraint is,

$$\underset{d\times d\cdot m_2}{E_1}\ p_1' = B \cdot \theta_j$$

where

$$\underset{d\times d\cdot m_2}{E_1} = \begin{bmatrix} \underset{1\times m_2}{e_1(\omega_1)} & \underset{1\times m_2}{0} & \cdots & \underset{1\times m_2}{0} \\ \underset{1\times m_2}{0} & \underset{1\times m_2}{e_1(\omega_2)} & \cdots & \underset{1\times m_2}{0} \\ & & & \\ \underset{1\times m_2}{0} & \underset{1\times m_2}{0} & \cdots & \underset{1\times m_2}{e_1(\omega_d)} \end{bmatrix}$$

is the matrix of excess demands, $p_1 = \underset{m_2\times 1}{(p_1(\omega_1)}, \cdots, \underset{m_2\times 1}{p_1(\omega_d))}$ is the matrix of spot prices, and

$$\underset{d\times a}{B} = \begin{bmatrix} v_1(\omega_1) & & v_a(\omega_1) \\ & \ddots & \\ v_1(\omega_d) & & v_a(\omega_d) \end{bmatrix}$$

is the payoffs matrix. We can rewrite the second period constraint as $p_1 \square e_{1j} = B \cdot \theta_j$, where e_{1j} is defined similarly as e_{0j}, and $p_1 \square e_{1j} \equiv (p_1(\omega_1)e_{1j}(\omega_1), \cdots, p_1(\omega_d)e_{1j}(\omega_d))'$. The budget constraints are then,

$$p_0 e_{0j} = -S\theta_j, \quad p_1 \square e_{1j} = B\theta_j.$$

Now suppose that markets are complete, i.e., $a = d$ and B can be inverted. The second constraint is then: $\theta_j = B^{-1} p_1 \square e_{1j}$. Consider without loss of generality Arrow securities, or $B = I$. We have $\theta_j = p_1 \square e_{1j}$, and by replacing into the first constraint,

$$\begin{aligned} 0 &= p_0 e_{0j} + S\theta_j \\ &= p_0 e_{0j} + S p_1 \square e_{1j} \\ &= p_0 e_{0j} + S \cdot (p_1(\omega_1)e_{1j}(\omega_1), \cdots, p_1(\omega_d)e_{1j}(\omega_d))' \\ &= p_0 e_{0j} + \sum_{i=1}^{d} S_i \cdot p_1(\omega_i) e_{1j}(\omega_i) \\ &= \sum_{h=1}^{m_1} p_0^{(h)} e_{0j}^{(h)} + \sum_{i=1}^{d} S_i \cdot \sum_{\ell=1}^{m_2} p_1^{(\ell)}(\omega_i) e_{1j}^{\ell}(\omega_i) \\ &= \sum_{h=1}^{m_1} p_0^{(h)} e_{0j}^{(h)} + \sum_{i=1}^{d} \sum_{\ell=1}^{m_2} \hat{p}_1^{(\ell)}(\omega_i) e_{1j}^{\ell}(\omega_i) \end{aligned}$$

where $\tilde{p}_1^{(\ell)}(\omega_i) \equiv S_i \cdot p_1^{(\ell)}(\omega_i)$. The price to be paid today for the obtention of a good ℓ in state i is equal to the price of an Arrow asset written for state i multiplied by the spot price $\tilde{p}_1^{(\ell)}(\omega_i)$ of this good in this state; here the Arrow-Debreu state price is $\tilde{p}_1^{(\ell)}(\omega_i)$. The general equilibrium can be analyzed by making reference to such state prices. From now on, we simplify and set $m_1 = m_2 \equiv m$. Then we are left with determining $m(d+1)$ equilibrium prices, i.e. $p_0 = (p_0^{(1)}, \cdots, p_0^{(m)})$, $\tilde{p}_1(\omega_1) = (\tilde{p}_1^{(1)}(\omega_1), \cdots, \tilde{p}_1^{(m)}(\omega_1))$, $\cdots$, $\tilde{p}_1(\omega_d) = (\tilde{p}_1^{(1)}(\omega_d), \cdots, \tilde{p}_1^{(m)}(\omega_d))$. By exactly the same arguments of the previous chapter, there exists one degree of indeterminacy. Therefore, there are only $m(d+1)-1$ relations that can determine the $m(d+1)$ prices. (Price normalization can be done by letting one of the first period commodities be the numéraire.) On the other hand, in the initial economy we have to determine $m(d+1)+d$ prices $(\hat{p}, \hat{S}) \in \mathbb{R}_{++}^{m\cdot(d+1)} \times \mathbb{R}_{++}^{d}$ which are the solution to the system:

$$\sum_{j=1}^{n} e_{0j}(\hat{p}, \hat{S}) = 0, \quad \sum_{j=1}^{n} e_{1j}(\hat{p}, \hat{S}) = 0, \quad \sum_{j=1}^{n} \theta_j(\hat{p}, \hat{S}) = 0,$$

where the previous functions are obtained as solutions to the agents' programs. When we solve for Arrow-Debreu prices, in a second step we have to determine $m(d+1)+d$ prices starting from the knowledge of $m(d+1)-1$ relations defining the Arrow-Debreu prices, which implies a price indeterminacy of the initial economy equal to $d+1$. In fact, it is possible to show that the degree of indeterminacy is only $d-1$.

References

Arrow, K. J. (1953): "Le rôle des valeurs boursières pour la répartitition la meilleure des risques." *Econométrie* 41-48. CNRS, Paris. Translated and reprinted in 1964: "The Role of Securities in the Optimal Allocation of Risk-Bearing." *Review of Economic Studies* 31, 91-96.

Bernoulli, D. (1738): "Specimen Theoriae Novae de Mensura Sortis." *Commentarii Academiae Scientiarum Imperialis Petropolitanae* V, 175-192. Reprinted in English in 1954: "Exposition of a New Theory on the Measurement of Risk." *Econometrica* 22, 23-36

Debreu, G. (1954): "Valuation Equilibrium and Pareto Optimum." *Proceedings of the National Academy of Sciences* 40, 588-592.

Debreu, G. (1959): *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. New Haven: Yale University Press.

Duffie, D. (2001): *Dynamic Asset Pricing Theory*. Princeton: Princeton University Press.

Duffie, D. and W. Shafer (1985): "Equilibrium in Incomplete Markets: I. A Basic Model of Generic Existence." *Journal of Mathematical Economics* 13 285-300.

Hart, O. (1974): "On the Existence of Equilibrium in a Securities Model." *Journal of Economic Theory* 9, 293-311.

He, H. and N. Pearson (1991): "Consumption and Portfolio Policies with Incomplete Markets and Short-Sales Constraints: The Infinite Dimensional Case." *Journal of Economic Theory* 54, 259-304.

Hicks, J. (1939): *Value and Capital*. Oxford: Oxford University Press.

Huang, C-F. and R.H. Litzenberger (1988): *Foundations for Financial Economics*. New York: North-Holland.

Magill, M. and M. Quinzii (1996): *Theory of Incomplete Markets*. Cambridge: MIT Press.

Modigliani, F. and M. Miller (1958): "The Cost of Capital, Corporation Finance and the Theory of Investment." *American Economic Review* 48, 261-297.

Radner, R. (1968): "Competitive Equilibrium under Uncertainty." *Econometrica* 36, 31-58.

Wilson, R. (1968): "The Theory of Syndicates." *Econometrica* 36, 119-132.

3
Infinite horizon economies

3.1 Introduction

This chapter extends the analysis of the previous two. Consumption is still an important determinant of asset prices. At the same time, this chapter analyzes asset prices in multiperiod economies. We consider simple economies, in which agents either live forever and have access to a set of complete markets, or belong to overlapping generations. We consider models without and with production, without and with money. We aim to develop fundamental tools applied or extended in subsequent chapters while dealing with financial frictions, bubbles or sunspots.

[In progress]

3.2 Consumption-based asset evaluation

3.2.1 Recursive plans: introduction

We consider a simple, benchmark case, arising in the absence of any risks for a decision maker. Consider an agent endowed with initial wealth equal to w_0, who solves the following problem:

$$\begin{aligned} &V(w_0) \equiv \max_{(c_t)_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t) \\ &\text{s.t. } \; w_{t+1} = (w_t - c_t) R_{t+1}, \quad (R_t)_{t=0}^{\infty} \text{ given} \end{aligned} \tag{3.P1}$$

This problem can be reformulated in a recursive format:

$$V(w_t) = \max_{c_t} \left[u(c_t) + \beta V(w_{t+1}) \right] \quad \text{s.t.} \quad w_{t+1} = (w_t - c_t) R_{t+1}. \tag{3.1}$$

By replacing the wealth constraint into the maximand, it is easily checked that the first-order condition for c leads to, $u'(c_t) = \beta V'(w_{t+1}) R_{t+1}$. Therefore, the consumption policy is a function of both wealth and the interest rate, and is denoted with $c(w_t)$ for simplicity. The value function and the first-order condition, then, can be written as:

$$V(w_t) = u(c(w_t)) + \beta V\left((w_t - c(w_t)) R_{t+1} \right), \quad u'(c(w_t)) = \beta V'\left((w_t - c(w_t)) R_{t+1} \right) R_{t+1}.$$

By differentiating the value function, and using the first-order condition, leaves the envelope condition:

$$V'(w_t) = u'(c(w_t))\,c'(w_t) + \beta V'((w_t - c(w_t))R_{t+1})(1 - c'(w_t))R_{t+1} = u'(c(w_t)). \quad (3.2)$$

Therefore, $V'(w_{t+1}) = u'(c(w_{t+1}))$ too, and by substituting back into the first-order condition,

$$\beta \frac{u'(c(w_{t+1}))}{u'(c(w_t))} = \frac{1}{R_{t+1}}. \quad (3.3)$$

The economic intuition underlying Eq. (3.3) is the same as that we saw in the two-period economy analyzed in Chapter 2. Eq. (3.3) says that along an optimal consumption path, the present consumption I give up at t to obtain additional consumption at $t+1$ has to equal the price at t of a pure discount bond. That is, the bond price is the relative price of consumption tomorrow relative to consumption today.

We can achieve the same conclusion relying on an alternative approach, based on Lagrange multipliers. This approach is useful when dealing with more intricate issues relating to production economies or economies with financial frictions, as we shall see in this and further chapters. So consider the constraint in program [3.P1]. Savings at time t are $\text{sav}_t \equiv w_t - c_t$. Using this definition, the constraint in [3.P1] is: $c_{t+1} + \text{sav}_{t+1} = R_{t+1}\text{sav}_t$, with $\text{sav}_{-1} = w_0$, given. Let λ_t be a sequence of Lagrange multipliers associated to these constraints. Consider the program,

$$\mathcal{L}(\text{sav}_{-1}) \equiv \max_{(c_t,\text{sav}_t)_{t=0}^{\infty}} \sum_{t=0}^{\infty} \left[\beta^t u(c_t) - \lambda_t (c_t + \text{sav}_t - R_t\text{sav}_{t-1})\right],$$

where λ_t is a sequence of Lagrange multipliers. The first-order condition for consumption c_t is, $\beta^t u'(c_t) = \lambda_t$, and the first-order condition for savings sav_t leads to: $\lambda_t = \lambda_{t+1}R_{t+1}$. Putting all together yields precisely Eq. (3.3). Finally, note that the same program can be cast, and solved, in a recursive format,

$$\mathcal{L}(\text{sav}_{t-1}) = \max_{c_t,\text{sav}_t,\lambda_t} \left[u(c_t) - \lambda_t (c_t + \text{sav}_t - R_t\text{sav}_{t-1}) + \beta\mathcal{L}(\text{sav}_t)\right].$$

The first-order condition for consumption and savings are $u'(c_t) = \lambda_t$ and $\lambda_t = \beta\mathcal{L}'(\text{sav}_t)$, respectively. By replacing the first-order condition for λ_t, i.e. the budget constraint, and differentiating $\mathcal{L}(\text{sav}_{t-1})$, leaves $\mathcal{L}'(\text{sav}_{t-1}) = \beta\mathcal{L}'(\text{sav}_t)R_t$. These conditions lead to Eq. (3.3).

As a simple example, consider the case of a logarithmic utility function, $u(c) = \ln c$. Let us guess that the value function is $V(w_t) \equiv V(w_t; R_t) = a_t + b\ln w_t$. Substituting this guess into the envelope condition in (3.2) leaves $c(w) = b^{-1}w$. By Eq. (3.3), then, $w_{t+1} = \beta w_t R_{t+1}$. Comparing the R.H.S. of this equation with the R.H.S. of the constraint in [3.P1], leaves $c(w_t) = (1-\beta)w_t$; in other terms, $b = (1-\beta)^{-1}$.[1]

Next, we introduce uncertainty and develop heuristic details regarding asset evaluation in this context.

[1] To pin down the coefficient series a_t, use the definition of the value function, $V(w_t; R_t) \equiv u(c(w_t)) + \beta V(w_{t+1}; R_{t+1})$. By plugging $V(w, R_t) = a_t + b\log w$ and $c(w) = (1-\beta)^{-1}w$ into this definition leaves, $a_t = \ln(1-\beta) + \beta a_{t+1} + \frac{\beta}{1-\beta}\ln(\beta R_{t+1})$. If R is constant, a_t is also constant, and equal to $(\ln(1-\beta) + \frac{\beta}{1-\beta}\ln(\beta R))/(1-\beta)$.

3.2.2 Asset pricing: the marginalist argument

Consider the following thought experiment. At time t, I give up to a small quantity of consumption equal to Δc_t. The reduction of utility at t then equals $u'(c_t)\Delta c_t$. But by investing Δc_t in a safe asset, I can consume $\Delta c_{t+1} = R_{t+1}\Delta c_t$ more at $t+1$. These additional consumption units lead to an expected utility gain equal to $\beta E_t\left(u'\left(c_{t+1}\right)\Delta c_{t+1}\right)$, where E_t denote the expectation conditional on time-t information. If c_t and c_{t+1} are part of an optimal consumption plan, I should be left with no incentives to implement these intertemporal consumption transfers. Therefore, along an optimal consumption plan, any reductions and gains in the welfare of the type considered above need to be identical:

$$u'(c_t) = \beta E_t\left(u'(c_{t+1})R_{t+1}\right).$$

This relation generalizes Eq. (3.3). Next, suppose that at time t, Δc_t can be invested in a risky asset whose price is S_t. I can buy $\Delta c_t / S_t$ units of this asset. Come time $t+1$, I could sell the asset for S_{t+1}, pocket its dividend D_{t+1} if any, and finance additional units of consumption equal to $\Delta c_{t+1} = \left(\Delta c_t / S_t\right) \cdot \left(S_{t+1} + D_{t+1}\right)$. The reduction in the current utility is $u'(c_t)\Delta c_t$. The boost in the expected utility at time $t+1$ is $\beta E_t\left(u'(c_{t+1})\Delta c_{t+1}\right)$. Again, if I am on an optimal consumption plan, there should not be incentives left to implement these intertemporal transfers. Therefore, the celebrated Lucas (1978) asset pricing equation holds:

$$u'(c_t) = \beta E_t\left[u'(c_{t+1})\frac{S_{t+1} + D_{t+1}}{S_t}\right]. \tag{3.4}$$

Section 3.2.4 derives Eq. (3.4) through dynamic programming methods, which are essential, once we wish to work through more complex models such as those including financial frictions. The next section, instead, elaborates on the optimality condition in Eq. (3.3) and develops key concept in both financial economics and macroeconomics: the intertemporal elasticity of substitution.

3.2.3 Intertemporal elasticity of substitution

The elasticity of substitution between two consumption goods, c_A and c_B, is defined as the ratio,

$$\mathrm{ES}\left(c_A, c_B\right) = -\frac{\partial\left(\frac{c_B}{c_A}\right)}{\partial\left(\frac{p_B}{p_A}\right)} \div \frac{\left(\frac{c_B}{c_A}\right)}{\left(\frac{p_B}{p_A}\right)},$$

where p_A and p_B are the prices of the two goods. It measures the percentage change in the relative consumption choice of two goods after a percentage change in their relative prices.

Similarly, define $\mathrm{EIS}\left(c_t, c_s\right) \equiv \mathrm{ES}\left(c_t, c_s\right)$ as the elasticity of *intertemporal* substitution of consumption c_t and c_s at two points in time t and $s \geq t$. By the first order conditions, $\beta\frac{u'(c_B)}{u'(c_A)} = \frac{p_B}{p_A}$,

$$\mathrm{EIS}\left(c_t, c_s\right) = -\frac{d\left(c_s / c_t\right)}{d\left(u'\left(c_s\right) / u'\left(c_t\right)\right)}\frac{u'\left(c_s\right) / u'\left(c_t\right)}{c_s / c_t} = -\frac{d\left(c_s / c_t\right)}{\left(c_s / c_t\right)} \div \frac{db}{b},$$

where

$$b = \frac{\beta u'\left(c_s\right)}{u'\left(c_t\right)} \equiv R^{-1}$$

denotes the price of a zero-coupon bond; accordingly, R denotes the gross interest rate from t to s. Note that we are assuming no uncertainty in these basic derivations.

The elasticity $\text{EIS}(c_t, c_s)$ is a measure of the percentage increase in the desired consumption tomorrow relative to today, after a percentage decrease of the price of consumption tomorrow relative to today. Intuitively, high values of $\text{EIS}(c_t, c_s)$ describe a situation where the agent is quite sensitive about consuming at t and s: even a small increase in the interest rate R from t to s and, hence, a small percentage drop in b, can induce him to a substantial relative increase of consumption in the future.

As $s \to t$, $\text{EIS}(c_t, c_s)$ collapses to the inverse of the elasticity of marginal utility with respect to consumption or, simply, the relative risk-aversion

$$\begin{aligned}\frac{1}{\text{EIS}(c_t)} \equiv \lim_{s\to t}\frac{1}{\text{EIS}(c_t, c_s)} &= -\lim_{s\to t}\frac{c_s/c_t}{u'(c_s)/u'(c_t)}\frac{d\left(u'(c_s)/u'(c_t)\right)}{d\left(c_s/c_t\right)} \\ &= -\lim_{s\to t}\frac{d\left(1 + c_t\frac{u''(c_t)}{u'(c_t)}\left(\frac{c_s}{c_t} - 1\right)\right)}{d\left(\frac{c_s}{c_t}\right)} \\ &= -c_t\frac{u''(c_t)}{u'(c_t)},\end{aligned}$$

where the second equality follows by a first-order Taylor's expansion of the marginal utility of consumption at time s, $u'(c_s) = u'(c_t) + u''(c_t)(c_s - c_t) + O((c_s - c_t)^2)$. The expression, $\text{EIS}(c_t)$, is called "instantaneous elasticity of intertemporal substitution."

For example, in the CRRA case, and again in the deterministic case, we have that along an optimal consumption path, $\frac{c_{t+1}}{c_t} = (\beta R)^{1/\gamma}$, where γ is the CRRA: as R increases, it becomes more attractive to save and postpone consumption. In a stylized equilibrium with a representative agent, $\ln R = -\ln\beta + \gamma g$, where g denotes the growth rate of the economy. When g is high, more consumption will be available in the future, which mitigates the incentives to save, driving the interest rate R up.

An agent with a low EIS has a quite inelastic demand for bonds. Intuitively, when the price of consumption in the future relative to today, b, drops, desired consumption tomorrow relative to today increases. But for an agent with a low EIS, the desired relative increase in future consumption is quite limited, and so is his demand for bonds—the instruments that allow him to allocate intertemporal consumption.

3.2.4 Lucas' model

3.2.4.1 The optimality condition

We consider markets for m "trees," and assume that the only source of risk stems from the dividends related to these trees: $D = (D_1, \cdots, D_m)$. We assume D is a Markov process and denote its conditional distribution function with $P(D_{t+1}|D_t)$. A representative agent solves the following program:

$$\begin{aligned}&V(\theta_t, D_t) = \max_{(c_{t+i},\theta_{t+i})_{i=0}^{\infty}} E_t\left[\sum_{i=0}^{\infty}\beta^i u(c_{t+i})\middle|\mathcal{F}_t\right] \\ &\text{s.t. } c_t + S_t\theta_{t+1} = (S_t + D_t)\theta_t\end{aligned} \qquad [3.P2]$$

where $\mathcal{F}_t$ denotes the information set as of time t, $\theta_{t+1} \in \mathbb{R}^m$ is $\mathcal{F}_t$-measurable, that is, θ_{t+1} needs to be chosen at time t. We can solve the program [3.P2], using the same recursive approach in

Section 3.2.1, once due account is made of uncertainty. The Bellman's equation is:

$$V(\theta_t, D_t) = \max_{c_t, \theta_{t+1}} E\left[u(c_t) + \beta V(\theta_{t+1}, D_{t+1}) \middle| \mathcal{F}_t\right] \quad \text{s.t.} \quad c_t + S_t\theta_{t+1} = (S_t + D_t)\,\theta_t. \tag{3.5}$$

Similarly as we did for Eq. (3.1), let us replace the budget constraint into the maximand. The following first-order condition holds for θ_i:

$$0 = E\left[-u'\left((S_t + D_t)\,\theta_t - S_t\theta_{t+1}\right) S_{i,t} + \beta V_{1i}\left(\theta_{t+1}, D_{t+1}\right) \middle| \mathcal{F}_t\right], \tag{3.6}$$

where the subscript in the value function on the right hand side denotes a partial derivative: $V_{1i}(\theta, D) = \partial V(\theta, D)/\partial\theta_i$. The optimal policy, θ_{t+1} is a function of the current state, (θ_t, D_t), say $\theta_{t+1} = \mathcal{T}(\theta_t, D_t)$. By differentiating the value function with respect to θ_i, and using the previous first-order condition, leaves:

$$\begin{aligned} V_{1i}(\theta_t, D_t) &= E_t\left[u'(c_t)\left(S_{i,t} + D_{i,t} - \sum_{j=1}^{m} S_{j,t}\mathcal{T}_{1j}^{i}(\theta_t, D_t)\right) + \beta \sum_{j=1}^{m} V_{1i}\left(\theta_{t+1}, D_{t+1}\right)\mathcal{T}_{1j}^{i}(\theta_t, D_t)\right] \\ &= u'(c_t)\left(S_{i,t} + D_{i,t}\right), \end{aligned}$$

where for brevity, we use E_t to denote the expectation operator conditional upon $\mathcal{F}_t$, and we have defined $\mathcal{T}_{1j}^{i}(\theta_t, D_t) = \partial\mathcal{T}_i(\theta, D)/\partial\theta_j$ and $\mathcal{T}_i$ is the i-th component of the vector $\mathcal{T}$. Substituting this result into Eq. (3.6) yields precisely the Lucas equation (3.4), holding for each asset i:

$$u'(c_t) = \beta E_t\left[u'(c_{t+1})\,\frac{S_{i,t+1} + D_{i,t+1}}{S_{i,t}}\right]. \tag{3.7}$$

It is easy to show to extend these conditions to the case where a representative agent can also invest into a locally riskless asset, that is, an asset that expires over the next period. The budget constraint in Eq. (3.5) is, in this case: $c_t + S_t\theta_{t+1} + S_{0,t}\theta_{0,t+1} = (S_t + D_t)\,\theta_t + \theta_{0,t}$, where $\theta_{0,t}$ denotes the amount of the locally riskless asset, $S_{0,t} \equiv e^{-r_t}$, and r_t is the riskless interest rate, and the Lucas equation for the r_t would be $e^{-r_t} = \beta E_t\left[u'(c_{t+1})/u'(c_t)\right]$.

3.2.4.2 Rational expectations equilibrium

The asset market clears when for each t, $\theta_t = \mathbf{1}_m$ and $\theta_{0,t} = 0$. By the budget constraint, then, the market for goods also clears, $c_t = \sum_{i=1}^{m} D_{it} \equiv \bar{D}_t$. A *rational expectation equilibrium* is a sequence of asset prices $(S_t)_{t=0}^{\infty}$ such that the optimality condition in Eq. (3.7) holds, the markets clear, $c_t = \bar{D}_t$, and each asset price is a function of the state, $S_{i,t} = S_i(D_t)$ say. All in all,

$$u'(\bar{D}_t)S_i(D_t) = \beta \int u'(\bar{D}_{t+1})\left(S_i(\bar{D}_{t+1}) + D_{i,t+1}\right) dP\left(D_{t+1} \middle| D_t\right). \tag{3.8}$$

This is a functional equation in $S_i(\cdot)$. Let us focus, first, on the IID case: $P\left(D_{t+1} \middle| D_t\right) = P\left(D_{t+1}\right)$.

IID shocks

Eq. (3.8) simplifies to:

$$u'(\bar{D}_t)S_i(D_t) = \beta \int u'(\bar{D}_{t+1})\left(S_i(\bar{D}_{t+1}) + D_{i,t+1}\right) dP\left(D_{t+1}\right). \tag{3.9}$$

Note that the right hand side of this equation is independent of D. Therefore, $u'(\bar{D}_t)S_i(D_t)$ equals some constant κ_i (say), which we can easily find by substituting it back into the previous equation, leaving:

$$\kappa_i = \frac{\beta}{1-\beta}\int u'(\bar{D}_{t+1})D_{i,t+1}dP(D_{t+1}).$$

Therefore, the solution for $S_i(D)$ is:

$$S_i(D_t) = \frac{\kappa_i}{u'(\bar{D}_t)}.$$

Note, the elasticity of the price to dividend equals $-\frac{u''(\bar{D})}{u'(\bar{D})}D_i$, which collapses to relative risk-aversion, once we assume only one tree exists. For example, if relative risk-aversion is constant and equal to γ,

$$S(D_t) = \kappa \cdot D_t^{\gamma}, \quad \kappa \equiv \frac{\beta}{1-\beta}\int D^{1-\gamma}dP(D).$$

Figure 3.1 depicts the behavior of the asset price function $S(D)$, under the assumption that κ is not increasing in γ. The asset price collapses to the constant, $\beta(1-\beta)^{-1}E(D)$, in the special case where the representative agent is risk-neutral, $\gamma = 0$.

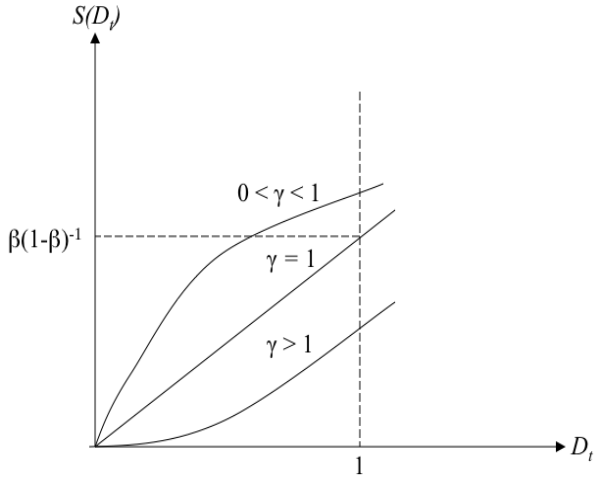


FIGURE 3.1. The asset pricing function $S(D_t)$ in the IID case and constant relative risk-aversion, equal to γ.

Dependent shocks

Define $g_i(D) \equiv u'(\bar{D})S_i(D)$ and $h_i(D) \equiv \beta\int u'(\bar{D}_{t+1})D_{i,t+1}dP(D_{t+1}|D)$. In terms of these new functions, Eq. (3.8) is:

$$g_i(D) = h_i(D) + \beta\int g_i(D_{t+1})\,dP(D_{t+1}|D).$$

It is a functional equation in g_i, which we can show it admits a unique solution, under the conditions contained in the celebrated Blackwell's theorem below:

THEOREM 3.1. *Let $\mathcal{B}(X)$ the Banach space of continuous bounded real functions on $X \subseteq \mathbb{R}^n$ endowed with the norm $\|f\| = \sup_X |f|$, $f \in \mathcal{B}(X)$. Introduce an operator $T : \mathcal{B}(X) \mapsto \mathcal{B}(X)$ with the following properties*:

(i) T *is monotone*: $\forall x \in X$ *and* $f_1, f_2 \in \mathcal{B}(X)$, $f_1(x) \leq f_2(x) \Longleftrightarrow T[f_1](x) \leq T[f_2](x)$;
(ii) $\forall x \in X$ *and* $c \geq 0$, $\exists \beta \in (0,1) : T[f+c](x) \leq T[f](x) + \beta c$.
Then, T *is a* β*-contraction and,* $\forall f_0 \in \mathcal{B}(X)$*, it has a unique fixed point* $\lim_{\tau \to \infty} T^{\tau}[f_0] = f = T[f]$.

So let us introduce the following operator:

$$T[g_i](D) = h_i(D) + \beta \int g_i(D')\, dP(D'|D).$$

The existence of g_i and, hence, S_i, relies on the existence of a fixed point of $T : g_i = T[g_i]$. It is easily checked that conditions (i) and (ii) in Theorem 3.1 hold here. To establish that $T : \mathcal{B}(D) \mapsto \mathcal{B}(D)$ as well, it is sufficient to show that $h_i \in \mathcal{B}(D)$. A sufficient condition given by Lucas (1978) is that u is bounded, and bounded away by a constant $\bar{u}$.[2] Note, a log-utility agent would not satisfy this condition, yet, this case can be easily solved in the case of a single tree, as shown next.

Suppose, then, that $u(c) = \ln c$, and that there is one single asset, such that Eq. (3.8) collapses to

$$\frac{S(D_t)}{D_t} = \beta \int \left(\frac{S(D_{t+1})}{D_{t+1}} + 1 \right) dP(D_{t+1}|D_t).$$

The solution to this equation is a constant price-dividend ratio,

$$\frac{S(D_t)}{D_t} = \frac{\beta}{1-\beta}.$$

Note that this result does not depend on any distribution assumption on the dividend process. However, in the general CRRA case, it cannot be said more, not even in the single asset case. Indeed, by Eq. (3.8),

$$\frac{S(D_t)}{D_t} = \beta \int \left(\frac{D_{t+1}}{D_t} \right)^{1-\gamma} \left(\frac{S(D_{t+1})}{D_{t+1}} + 1 \right) dP(D_{t+1}|D_t).$$

It is easily seen that the solution to this functional equation is:

$$\frac{S(D_t)}{D_t} = \frac{\beta \int \left(\frac{D_{t+1}}{D_t} \right)^{1-\gamma} dP(D_{t+1}|D_t)}{1 - \beta \int \left(\frac{D_{t+1}}{D_t} \right)^{1-\gamma} dP(D_{t+1}|D_t)},$$

such that the price-dividend ratio is constant whenever the distribution of the consumption endowment growth rate is independent of D_t. In Chapter 7 of Part II, we develop this case in more detail, assuming a log-normal distribution for D_{t+1}/D_t.

3.2.4.3 Arrow-Debreu state prices

The model makes sharp prediction regarding Arrow-Debreu securities prices. In terms of the terminology of Chapter 2, we want to identify the *stochastic discounting factor*. By the asset pricing equation (3.9), it is

$$m_{t+1} \equiv \beta \frac{u'(D_{t+1})}{u'(D_t)}.$$

[2] In this case, concavity of u implies that for each D, $0 = u(0) \leq u(D) + u'(D)(-D) \leq \bar{u} - Du'(D)$, which implies that for each D, $Du'(D) \leq \bar{u}$ and, hence, $h_i(D) \leq \beta \bar{u}$. Then, it is possible to show that the solution is in $\mathcal{B}(D)$, which implies that $T : \mathcal{B}(D) \mapsto \mathcal{B}(D)$.

Under regularity conditions, the Radon-Nikodym derivative of the risk-neutral probability, P^*, with respect to P, is: $\frac{dP^*}{dP}\big|_{D_{t+1}|D_t} \equiv \frac{u'(D_{t+1})}{E[u'(D_{t+1}|D_t)]}$, such that the Arrow-Debreu state-price density is: $d\tilde{P}^*(D_{t+1}|D_t) = R_t^{-1} dP^*(D_{t+1}|D_t)$. It is the price to pay in state D_t to obtain one unit of the good the next period when the state is D_{t+1}. These conclusions generalize those of Chapter 2 to the infinite horizon case. Parts II and III contains many more examples of stochastic discounting factors that are useful to address empirical puzzles or applied issues arising in both equity and fixed income markets.

3.2.4.4 Multiple trees

The Lucas model is extraordinary complex as in general, the price of any asset depends on the dividends paid by all the remaining assets, as Eq. (3.8) makes clear. The model can generate "contagion," in that a shock in the fundamentals affecting some assets affects all the other asset evaluation, even when the dividends are not correlated. It is an interesting property, due to the simple circumstance that there is a representative agent who is pricing the same assets—markets are not segmented and a shock to the stochastic discounting factor, m_{t+1}, affects all the asset prices, $S_{i,t}$. We mention efforts made by the literature, discussed in deeper detail in Chapter 9 (Section 9.10): Menzly, Santos and Veronesi (2004), Cochrane, Longstaff and Santa-Clara (2008), Pavlova and Rigobon (2008), Martin (2011).

3.3 Production: foundational issues

In the economy of the previous section, asset payoffs are given exogenously. We now lay down analytical foundations regarding production-based economies, in which firms maximize their value and set dividends endogenously. In these economies, production and capital accumulation are endogenous. In this section, we review the foundational issues arising in economies with productive capital. In the next section, we develop asset pricing implications of these economies in absence of frictions. In Part II (see Chapters 8 and 9), we extend the framework in this chapter, and examine asset price implications of markets plagued with financial frictions.

3.3.1 Decentralized economy

A continuum of identical firms in $(0,1)$ have access to capital and labor markets, and the following technology: $(K,N) \mapsto Y(K,N)$,where $Y_i(K,N) > 0$, $y_{ii}(K,N) < 0$, $\lim_{K\to 0+} Y_1(K,N) = \lim_{N\to 0+} Y_2(K,N) = \infty$, $\lim_{K\to\infty} Y_1(K,N) = \lim_{N\to\infty} Y_2(K,N) = 0$, and subscripts denote partial derivatives. We assume Y is homogeneous of degree one, i.e. $Y(\lambda K, \lambda N) = \lambda Y(K,N)$ for all $\lambda > 0$. Per capita production is $y(k) \equiv Y(k,1)$, where $k \equiv K/N$ is per-capita capital, Population growth can be non-zero, i.e. N satisfies $N_t/N_{t-1} = (1+n)$. Firms purchase capital and labor at prices $R = Y_1(K,N)$ and $w = Y_2(K,N)$. We have,

$$R = y'(k), \quad w = y(k) - ky'(k).$$

The N_t consumers live forever. We assume each consumer offers inelastically one unit of labor, and that, for now, that $N_0 = 1$ and $n = 0$. The resource constraint for the consumer is:

$$c_t + s_t = R_t s_{t-1} + w_t N_t, \quad N_t \equiv 1, \quad t = 1, 2, \cdots. \tag{3.10}$$

At each time $t-1$, the consumer saves s_{t-1} units of capital, which he lends to the firm. At time t, the consumer receives the gross return on savings from the firm, $R_t s_{t-1}$, where $R_t = y'(k_t)$,

plus the wage receipts $w_t N_t$. Then, he chooses how to allocate these resources to consumption c_t and to lending s_t to the firm. At time zero,

$$c_0 + s_0 = V_0 \equiv Y_1(K_0, N_0)K_0 + w_0 N_0, \quad N_0 \equiv 1.$$

Following the approach developed in Chapter 2, we can write down a single budget constraint, obtained iterating Eq. (3.10):

$$0 = c_0 + \sum_{t=1}^{T} \frac{c_t - w_t N_t}{\prod_{i=1}^{t} R_i} + \frac{s_T}{\prod_{i=1}^{T} R_i} - V_0,$$

and imposing the transversality condition:

$$\lim_{T\to\infty} s_T \prod_{i=1}^{T} R_i^{-1} = 0, \tag{3.11}$$

so as to have:

$$\max_{(c_t)_{t=0}^{\infty}} \sum_{t=1}^{\infty} \beta^t u(c_t), \quad \text{s.t.} \quad V_0 = c_0 + \sum_{t=1}^{\infty} \frac{c_t - w_t N_t}{\prod_{i=1}^{t} R_i}. \tag{3.P3}$$

The economic interpretation of the transversality condition (3.11) is the following. The first-order conditions of the program [3.P3] are:

$$\beta^t u'(c_t) = l \frac{1}{\prod_{i=1}^{t} R_i}, \tag{3.12}$$

where l is a Lagrange multiplier. In equilibrium, current savings equal next period capital, or $k_{t+1} = s_t$. Therefore, Eq. (3.11) is:

$$\lim_{T\to\infty} \beta^T u'(c_T)\, k_{T+1} = 0. \tag{3.13}$$

That is, the value of capital is capital weighted by discounted marginal utility, and is worthless, eventually. We shall derive Eq. (3.13) below.

The first-order condition (3.12) leads to the usual optimality condition in Eq. (3.3), where this time, $R_{t+1} = y'(k_{t+1})$. In this economy, an equilibrium is any sequence $(c_t, k_t)_{t=0}^{\infty}$ satisfying

$$\begin{cases} k_{t+1} = y(k_t) - c_t \\ \beta \dfrac{u'(c_{t+1})}{u'(c_t)} = \dfrac{1}{y'(k_{t+1})} \end{cases} \tag{3.14}$$

and the transversality condition in Eq. (3.13). The first equation in this system is simply this: capital available for producing the next period, k_{t+1}, is equal to savings, $s_t \equiv y(k_t) - c_t$.

3.3.2 The social planner solution

3.3.2.1 Recursive plan

The market solution in (3.14) can be implemented by a social planner, who solves the following program:

$$\begin{aligned} V(k_0) \equiv & \max_{(c_t, k_t)_{t=0}^{\infty}} \sum_{i=0}^{\infty} \beta^i u(c_t) \\ \text{s.t.} \quad & k_{t+1} = y(k_t) - c_t, \quad k_0 \text{ given} \end{aligned} \tag{3.P4}$$

under the further transversality condition in Eq. (3.13).

The program in [3.P4] is easily solved. By replacing the constraint into the utility function, and taking derivatives with respect to k_t, leads directly to the second equation in (3.14). Alternatively, let us introduce the Lagrangian,

$$\mathcal{L}(k_0) = \max_{(c_t,k_{t+1})_{t=0}^{\infty}} \sum_{t=0}^{\infty} \left[\beta^t u(c_t) - \lambda_t \left(k_{t+1} - y(k_t) + c_t\right)\right].$$

The first-order condition with respect to consumption is $\lambda_t = \beta^t u'(c_t)$, and the condition for capital is $\lambda_{t-1} = \lambda_t y'(k_t)$. Putting these conditions together, leads to the second equation in (3.14). The same argument can be made, following a recursive approach. We have:

$$\mathcal{L}(k_t) = \max_{c_t,k_{t+1},\lambda_t} \left[u(c_t) - \lambda_t \left(k_{t+1} - y(k_t) + c_t\right) + \beta \mathcal{L}(k_{t+1})\right].$$

The first-order condition for consumption is $\lambda_t = u'(c_t)$, and that for capital is $\lambda_t = \beta \mathcal{L}'(k_{t+1})$. By replacing the first-order condition for λ_t (i.e., the constraint in program [3.P4]), and differentiating with respect to k_t, yields $\mathcal{L}'(k_t) = \beta \mathcal{L}'(k_{t+1}) y'(k_t)$. These three conditions lead, again, to the second equation in (3.14).

Finally, consider the Bellman's equation:

$$V(k_t) = \max_{c_t} \left[u(c_t) + \beta V(k_{t+1})\right], \quad \text{s.t. } k_{t+1} = y(k_t) - c_t.$$

The first-order condition leads to, $u'(c_t) = \beta V'(y(k_t) - c_t)$. Let us denote the policy with $c_t = c(k_t)$. In terms of the policy c function, the value function and the first-order conditions are:

$$V(k_t) = u(c(k_t)) + \beta V(y(k_t) - c(k_t)), \quad u'(c(k_t)) = \beta V'(y(k_t) - c(k_t)).$$

By differentiating the value function:

$$V'(k_t) = u'(c(k_t)) c'(k_t) + \beta V'(y(k_t) - c(k_t))(y'(k_t) - c'(k_t)) = u'(c(k_t)) y'(k_t).$$

By replacing back into the first-order condition, we obtain the second equation in (3.14).

3.3.2.2 Transversality condition again

Another derivation of the transversality condition relies on the following arguments. Consider the followig "truncated program,"

$$\max_{(c_t,k_{t+1})_{t=0}^{T}} \sum_{t=0}^{T} \left[\beta^t u(c_t) - \lambda_t \left(k_{t+1} - y(k_t) + c_t\right)\right], \quad \text{s.t. } \lambda_T K_{T+1} = 0.$$

The first order conditions are the usual ones. Moreover, by multiplying $\beta^T u'(c_T) = \lambda_T$ by K_{T+1} and utilizing the previous constraint leaves: $\beta^T u'(c_T) K_{T+1} = 0$. Taking the limit for T large yields Eq. (3.13).

3.3.3 Dynamics

We study the dynamics of the system in (3.14) in a small neighborhood of the stationary state, defined as the pair (c, k), solution to:

$$c = y(k) - k, \quad \beta = \frac{1}{y'(k)}.$$

A first-order expansion of each equation in (3.14) around its stationary state, yields the following linear system:

$$\begin{pmatrix} \bar{k}_{t+1} \\ \bar{c}_{t+1} \end{pmatrix} = A \begin{pmatrix} \bar{k}_t \\ \bar{c}_t \end{pmatrix}, \quad A \equiv \begin{pmatrix} y'(k) & -1 \\ -\frac{u'(c)}{u''(c)} y''(k) & 1 + \beta \frac{u'(c)}{u''(c)} y''(k) \end{pmatrix}. \tag{3.15}$$

The solution to this system is obtained with the tools reviewed in Appendix 1 of this chapter. It is:

$$\bar{k}_t = v_{11}\kappa_1 \lambda_1^t + v_{12}\kappa_2 \lambda_2^t, \quad \bar{c}_t = v_{21}\kappa_1 \lambda_1^t + v_{22}\kappa_2 \lambda_2^t, \tag{3.16}$$

where: κ_i are constants that depend on the initial state, λ_i are the eigenvalues of A, and $\binom{v_{11}}{v_{21}}$, $\binom{v_{12}}{v_{22}}$ are the eigenvectors associated with λ_i. In Appendix 1, we show that $\lambda_1 \in (0,1)$ and $\lambda_2 > 1$. The proof we provide in the appendix is important, as it illustrates precisely how the neoclassical model reviewed in this section, needs to be modified to induce indeterminacy in the dynamics of capital and consumption. A critical step in that proof relies on the assumption of diminishing returns, i.e. $y''(k) > 0$.

Let us return to the equations in (3.16). First, we need to rule out an explosive behavior of $\bar{k}_t$ and $\bar{c}_t$, for otherwise we would contradict (i) that (c, k) is a stationary point, and (ii) the optimality of the trajectories. Since $\lambda_2 > 1$, the only possibility is to "lock" the initial state $(\bar{k}_0, \bar{c}_0)$ in such a way that $\kappa_2 = 0$, which yields the following set of initial conditions: $\bar{k}_0 = v_{11}\kappa_1$ and $\bar{c}_0 = v_{21}\kappa_1$, or $\frac{\bar{c}_0}{\bar{k}_0} = \frac{v_{21}}{v_{11}}$.[3] Therefore, the set of initial points that ensure a non-explosive path must lie on the line $c_0 = c + \frac{v_{21}}{v_{11}}(k_0 - k)$. Since k is a pre-determined variable, there exists one, and only one, value of c_0, which ensures a non-explosive path of the system around its steady state, as Figure 3.2 illustrates. In this figure, k_* is defined as the solution of $1 = y'(k_*) \Leftrightarrow k_* = (y')^{-1}[1]$, and $k = (y')^{-1}[\beta^{-1}]$.

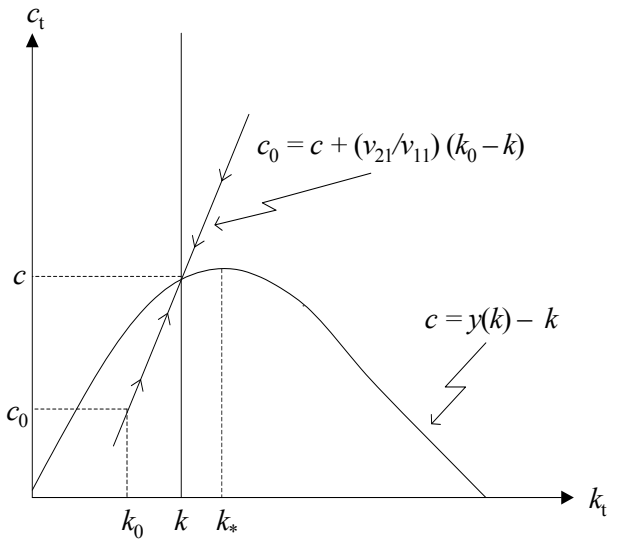


FIGURE 3.2.

The usual word of caution is in order. A linear approximation around the steady state might turn out to provide a poor approximation to the system dynamics once we are far from the steady state. Consider the following example, in which $y(k) = k^\alpha$ and $u(c) = \ln c$. It is easy to

[3] In fact, Appendix 1 shows that the converse is also true, i.e. $\frac{\bar{c}_0}{\bar{k}_0} = \frac{v_{21}}{v_{11}} \Rightarrow \kappa_2 = 0$.

show that the exact solution is:

$$c_t = (1 - \beta\alpha)\, k_t^{\alpha}, \quad k_{t+1} = \beta\alpha k_t^{\alpha}.$$

Figure 3.3 depicts the nonlinear manifold of this system (i.e., the first of the previous equations), and its linear approximation (the saddlepath), assuming that $\beta = 0.99$ and $\alpha = 0.3$; the linear saddlepath is

$$c_t = c + 0.7101\,(k_t - k), \quad \text{where: } c = (1 - \alpha\beta)\, k^{\alpha}, \quad k = (\alpha\beta)^{1/(1-\alpha)},$$

where $\bar{k}_t = \lambda_1 \bar{k}_{t-1}$, and $\lambda_1 = 0.3$.

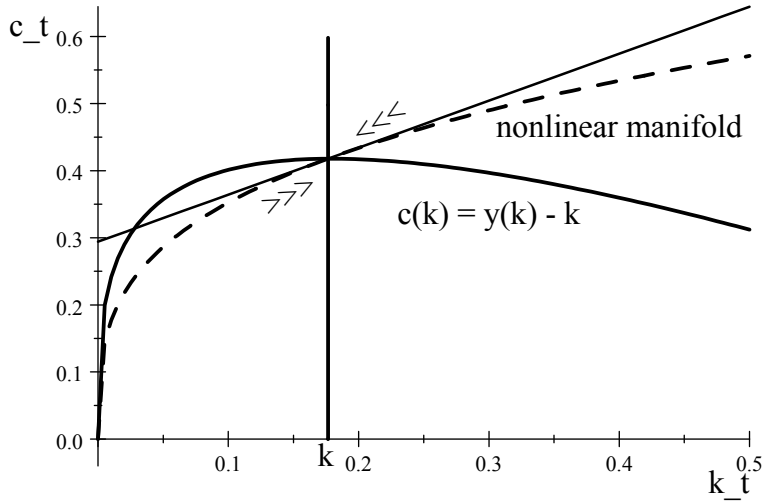


FIGURE 3.3.

3.3.4 Stochastic economies

> "Real business cycle theory is the application of general equilibrium theory to the quantitative analysis of business cycle fluctuations." Edward Prescott (1991, p. 3)
>
> "The Kydland and Prescott model is a complete markets set-up, in which equilibrium and optimal allocations are equivalent. When it was introduced, it seemed to many—myself included—to be much too narrow a framework to be useful in thinking about cyclical issues." Robert Lucas (1994, p. 184)

In its simplest version, real business cycle theory is an extension of the neoclassical model of Section 3.3.3, in which random productivity shocks are added. The engine of fluctuations, then, comes from the real sphere of the economy. This approach is in contrast with the Lucas approach of the 1970s, based on information and money, where fluctuations arise due to information delays with which agents discover the nature of a shock (real or monetary). As further reviewed in Chapter 10, the Lucas information-theoretic approach has been, instead, more successful in inspiring work on the formation of asset prices, leading to the development of market microstructure theory and, more generally, to information driven explanations of asset prices.

Despite the remarkable switch in the economic motivation, the paradigm underlying real business cycle theory is the same as the information-based approach of Lucas, as it relies on

rational expectations: macroeconomic fluctations and, then, as we shall explain, asset prices fluctuations, stem from the optimal response of the agents vis-à-vis exogeneous shocks: agents implement action plans that are state-contingent, i.e. they decide to consume, to work and to invest according to the history of shocks as well as the present shocks they observe.

3.3.4.1 Basic model

We consider an economy with complete markets and no frictions, such that its equilibrium allocations are Pareto-optimal. To characterize these allocations, we implement them through the following program of a social planner:

$$V(k_0, s_0) = \max_{(c_t)_{t=0}^{\infty}} E\left[\sum_{t=0}^{\infty} \beta^t u(c_t)\right], \tag{3.17}$$

subject to a capital accumulation constraint, with capital depreciation. Let I_t denote new investment. It is:

$$I_t = K_{t+1} - (1-\delta) K_t. \tag{3.18}$$

At time $t-1$, the available productive capital is K_t. At time t, a portion δK_t of this capital is lost, due to depreciation. Therefore, at time t, the productive system is left with $(1-\delta) K_t$ units of capital. The capital available at time t, K_{t+1}, equals the capital already in place, $(1-\delta) K_t$, plus new investments, which is exactly what Eq. (3.18) says.

Next, normalize population normalized to one, such that $K_t = k_t$. The goods market clearing condition is:

$$\tilde{y}(k_t, s_t) = c_t + I_t,$$

where $\tilde{y}(k_t, s_t)$ is the production function, which is $\mathcal{F}_t$-measurable, and s is the source of randomness—the engine for random fluctuations of the endogeneous variables. By replacing Eq. (3.18) into the equilibrium condition,

$$k_{t+1} = \tilde{y}(k_t, s_t) - c_t + (1-\delta) k_t. \tag{3.19}$$

So the planner maximizes the utility in Eq. (3.17), under the capital accumulation constraint in Eq. (3.19).

We assume that $\tilde{y}(k_t, s_t) \equiv s_t y(k_t)$, where y is as in Section 3.2, and $(s_t)_{t=0}^{\infty}$ is solution to:

$$s_{t+1} = s_t^{\rho} \epsilon_{t+1}, \tag{3.20}$$

where $\rho \in (0,1)$, and $(\epsilon_t)_{t=0}^{\infty}$ is a IID sequence with support s.t. $s_t \geq 0$. In this economy, every asset is priced as in the Lucas model of the previous section. Therefore, the gross return on savings $s.y'(k.)$ satisfies:

$$u'(c_t) = \beta E_t\left(u'(c_{t+1})\left(s_{t+1} y'(k_{t+1}) + 1 - \delta\right)\right). \tag{3.21}$$

A rational expectation equilibrium is a stochastic process $(c_t, k_t)_{t=0}^{\infty}$, satisfying Eq. (3.19), the Euler equation in (3.21), for given k_0 and s_0.

We show the existence of a saddlepoint path for the linearized version of Eqs. (3.19)-(3.20)-(3.21), which implies determinacy of the stochastic (linearized) equilibrium.[4] We study the

[4] A stochastic equilibrium is the situation where there is a stationary measure (definition: $p(+) = \int \pi(+/-) dp(-)$, where π is the transition measure) generating $(c_t, k_t)_{t=1}^{\infty}$.

behavior of $(k,c,s)_t$ in a neighborhood of $\epsilon \equiv E(\epsilon_t)$. Let (c,k,s) be consumption, capital and productivity shock, corresponding to ϵ, obtained replacing ϵ into Eqs. (3.19)-(3.20)-(3.21), and assuming no uncertainty takes place:

$$c = sy\left(k\right) - \delta k, \quad s = \epsilon^{\frac{1}{1-\rho}}, \quad \beta = \frac{1}{sy'(k) + 1 - \delta}.$$

A first-order approximation to Eqs. (3.19)-(3.20)-(3.21) around (k,c,s), leaves:

$$\hat{z}_{t+1} = \Phi \hat{z}_t + R u_{t+1}, \tag{3.22}$$

where we have defined $\hat{x}_t \equiv \frac{x_t - x}{x}$, and $\hat{z}_t = (\hat{k}_t, \hat{c}_t, \hat{s}_t)^\top$, $u_t = (u_{c,t}, u_{s,t})^\top$, $u_{c,t} = \hat{c}_t - E_{t-1}(\hat{c}_t)$, $u_{s,t} = \hat{s}_t - E_{t-1}(\hat{s}_t) = \hat{\epsilon}_t$, and, finally,

$$\Phi = \begin{pmatrix} \beta^{-1} & -\frac{c}{k} & s\frac{y(k)}{k} \\ -\frac{u'(c)}{cu''(c)} sky''(k) & 1 + \frac{\beta u'(c)}{u''(c)} sy''(k) & -\frac{\beta u'(c)}{cu''(c)} s(sy(k)y''(k) + \rho y'(k)) \\ 0 & 0 & \rho \end{pmatrix}, \quad R = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Let us consider the characteristic equation:

$$0 = \det\left(\Phi - \lambda I\right) = (\rho - \lambda)\left[\lambda^2 - \left(\beta^{-1} + 1 + \beta \frac{u'(c)}{u''(c)} sy''(k)\right)\lambda + \beta^{-1}\right].$$

A solution is $\lambda_1 = \rho$. Relying on the same arguments used in the deterministic case (see Section 3.3.3 and Appendix 1), one finds that $\lambda_2 \in (0,1)$ and $\lambda_3 > 1$.[5] Similarly as in the deterministic case, we diagonalize the system by rewriting $\Phi = P\Lambda P^{-1}$, where Λ is a diagonal matrix that has the eigenvalues of Φ on the diagonal, and P is a matrix of the eigenvectors associated to the roots of Φ. Eq. (3.22) is then:

$$\hat{y}_{t+1} = \Lambda \hat{y}_t + w_{t+1}, \tag{3.23}$$

where $\hat{y}_t \equiv P^{-1}\hat{z}_t$ and $w_t \equiv P^{-1}Ru_t$. The third equation of this system is:

$$\hat{y}_{3,t+1} = \lambda_3 \hat{y}_{3t} + w_{3,t+1}, \tag{3.24}$$

and $\hat{y}_3$ explodes unless $\hat{y}_{3t} = 0$ for all t, which is only possible when $w_{3t} = 0$ for all t.[6]

The condition that $w_{3t} = 0$ carries an interesting economic interpretation: it tells us that the only sources of uncertainty in this system can stem from shocks to the fundamentals, or that there cannot be extraneous sources of noise, or "sunspots." Let us explain. Define $\Pi \equiv P^{-1}$, such that $w_{3t} = 0$ implies that there is a linear relation between the two expectational errors (in consumption and technology), viz

$$\text{For all } t, \;\; u_{ct} = -\frac{\pi_{33}}{\pi_{32}} u_{st} \quad (\text{"no-sunspots"}). \tag{3.25}$$

[5] The linearized model in this section has state variables expressed in growth rates. However, the model can be reformulated in terms of first differences, by pre- and post- multiplying Φ by appropriate normalizing matrices. For example, if G i the 3×3 matrix that has $\frac{1}{k}$, $\frac{1}{c}$ and $\frac{1}{s}$ on its diagonal, (3.22) can be written as: $E(z_{t+1} - z) = G^{-1}\Phi G \cdot (z_t - z)$, where $z_t = (k_t, c_t, s_t)$, and we would arrive at the same conclusions. It can be easily shown that the model in this section collapses to that in Section 3.3.3, once we set $\epsilon_t = 1$, for each t, and $s_0 = 1$.

[6] In other words, Eq. (3.24) implies that $\hat{y}_{3t} = \lambda_3^{-(T-t)} E_t(\hat{y}_{3,t+T})$, and for all T. Because $\lambda_3 > 1$, this relation holds only when $\hat{y}_{3t} = 0$ for all t.

Eq. (3.25) is a "no-sunspots" condition, in that it says that expectational errors in consumption cannot be independent of those of the economic fundamentals—in this case, technological shocks. That is, the source of uncertainty in this economy is the presence of technological shocks, such that any remaining expectational errors can only be perfectly correlated to these: there are no sunspots.

Let us elaborate further on the condition that $\hat{y}_{3t} = 0$. Let $\hat{y}_t = P^{-1}\hat{z}_t \equiv \Pi\hat{z}_t$, such that

$$0 = \hat{y}_{3t} = \pi_{31}\hat{k}_t + \pi_{32}\hat{c}_t + \pi_{33}\hat{s}_t. \tag{3.26}$$

Eq. (3.26) tells us that the three state variables, $\hat{k}_t$, $\hat{c}_t$ and $\hat{s}_t$, are mutually linked in a two-dimensional plane, a saddlepath in which they exhibit a stable behavior. This stable manifold is defined as:

$$\mathcal{S} = \left\{ x \in \mathbb{R}^3 \middle| \, \pi_{3.}x = 0 \right\}, \quad \pi_{3.} = (\pi_{31}, \pi_{32}, \pi_{33}),$$

and bears the same meaning as that borne by the stable saddlepath in Figure 3.2 in the deterministic case. Precisely, $\mathcal{S}$ is the convergent subspace, with $\dim(\mathcal{S}) = 2$, i.e., the number of roots with modulus less than one. In other words, in this economy with two pre-determined variables, $\hat{k}_0$ and $\hat{s}_0$, there exists one, and only one, value of $\hat{c}_0 \in \mathcal{S}$ that ensures stability, given by $\hat{c}_0 = -\frac{\pi_{31}\hat{k}_0 + \pi_{33}\hat{s}_0}{\pi_{32}}$. This reasoning generalizes that made in the deterministic case (Section 3.3.3), and is generalized further in Appendix 1.

The solution to the linearized model can be determined by generalizing the reasoning in the deterministic case. First, by Eq. (3.23) $\hat{y}$ is:

$$\hat{y}_{it} = \lambda_i^t \hat{y}_{i0} + \zeta_{it}, \quad \zeta_{it} \equiv \sum_{j=0}^{t-1} \lambda_i^j w_{i,t-j},$$

which implies the solution for $\hat{z}$ is:

$$\hat{z}_t = P\hat{y}_t = (v_1 \; v_2 \; v_3)\hat{y}_t = \sum_{i=1}^{3} v_i \hat{y}_{it} = \sum_{i=1}^{3} v_i \hat{y}_{i0}\lambda_i^t + \sum_{i=1}^{3} v_i \zeta_{it}.$$

To pin down the components of $\hat{y}_0$, note that $\hat{z}_0 = P\hat{y}_0 \Rightarrow \hat{y}_0 = P^{-1}\hat{z}_0 \equiv \Pi\hat{z}_0$. The stability condition then requires that the state variables be in $\mathcal{S}$, or $\hat{y}_0^{(3)} = 0$, which we now use to implement the solution. We have:

$$\hat{z}_t = v_1\lambda_1^t\hat{y}_{10} + v_2\lambda_2^t\hat{y}_{20} + v_3\lambda_3^t\hat{y}_{30} + v_1\zeta_{1t} + v_2\zeta_{2t} + v_3\zeta_{3t}.$$

Moreover, the term $v_3\lambda_3^t\hat{y}_{30} + v_3\zeta_{3t}$ needs to be zero, because $\hat{y}_{30} = 0$. Finally, we have that $\zeta_{3t} = \sum_{j=0}^{t-1} \lambda_3^j w_{3,t-j}$, and since $w_{3,t} = 0$, then, then $\zeta_{3t} = 0$ as well. Therefore, the solution for $\hat{z}_t$ is:

$$\hat{z}_t = v_1\lambda_1^t\hat{y}_{10} + v_2\lambda_2^t\hat{y}_{20} + v_1\zeta_{1t} + v_2\zeta_{2t}.$$

3.3.4.2 Frictions, indeterminacy and sunspots

In the neoclassical model that we are analyzing, the equilibrium is determinate. As explained, this property arises because the number of pre-determined variables equals the dimension of the convergent subspace of the economy. If we managed to increase the dimension of the converging subspace, the equilibrium would be indeterminate, as further formalized in Appendix 1. As it turns out, indeterminacy goes hand in hand with sunspots, the expectational shocks that are extraneous to those in the economic fundamentals, as we discussed earlier, just after Eq. (3.25).

Introducing sunspots in macroeconomics has been an approach pursued in detail by Farmer in a series of articles (see Farmer, 1998, for an introductory account of this approach). The idea is quite interesting, as we know that the basic real business cycle model of this section needs many extensions in order not to be rejected, empirically, as originally shown by Watson (1993). In other words, the basic model in this section offers little room for a rich propagation mechanism, as it entirely relies on impulses, the productivity shocks, which "we hardly read about in the Wall Street Journal," as provocatively put by King and Rebelo (1999). Sunspots offer an interesting route to enrich the propagation mechanism, although their asset pricing implications in terms of the model analyzed in this section, have not been explored yet.

In a series of articles, David Cass showed that a Pareto-optimal economy cannot harbour sunspots equilibria. On the other hand, any market imperfection has the potential to be a source of sunspots. The typical example is incomplete markets. The neoclassical model analyzed in this section cannot generate sunspots, as it relies on a system of perfectly competitive markets and absence of any sort of frictions. To introduce sunspots in the economy of this section, we need to think about some deviation from optimality. Two possibilities analyzed in the literature are the presence of imperfect competition and/or externality effects. We provide an example of these effects, by working out the deterministic economy in Section 3.3.3. (Generalizations to the stochastic economy in this section are easy, although more cumbersome.)

How is it that a deterministic economy might generate "stochastic outcomes," that is, outcomes driven by shocks entirely unrelated to the fundamentals of the economy? Let us imagine this can be possible. Then, both optimal consumption and capital accumulation in Section 3.3.3 are necessarily random processes. The system in (3.15), then, must be rewritten in an expectation format,

$$E_t \begin{pmatrix} \hat{k}_{t+1} \\ \hat{c}_{t+1} \end{pmatrix} = A \begin{pmatrix} \hat{k}_t \\ \hat{c}_t \end{pmatrix}.$$

Next, let us introduce the expectational error process $u_{c,t} \equiv \hat{c}_t - E_{t-1}(\hat{c}_t)$, which we plug back into the previous system, to obtain:

$$\begin{pmatrix} \hat{k}_{t+1} \\ \hat{c}_{t+1} \end{pmatrix} = A \begin{pmatrix} \hat{k}_t \\ \hat{c}_t \end{pmatrix} + \begin{pmatrix} 0 \\ u_{c,t+1} \end{pmatrix}.$$

Naturally, we still have $\lambda_1 \in (0,1)$ and $\lambda_2 > 1$, as in Section 3.3.3. Therefore, we decompose A as $P\Lambda P^{-1}$, and have:

$$\hat{y}_{t+1} = \Lambda \hat{y}_t + P^{-1}(0 \quad u_{c,t+1})^\top.$$

Moreover, for $\hat{y}_{2t} = \lambda_2^{-T} E_t(\hat{y}_{2,t+T})$ to hold for all T, we need to have $\hat{y}_{2t} = 0$, for all t. Therefore, the second element of the vector $P^{-1}(0 \;\; u_{c,t+1})^\top$ must be zero, or, for all t,

$$0 = \pi_{22} u_{c,t} \Longleftrightarrow 0 = u_{c,t}.$$

There is no room for expectational errors and, hence, sunspots, in this model. The fact that $\lambda_2 > 1$ implies the dimension of the saddlepoint is less than the number of pre-determined variables. So a viable route to pursue here, is to look for economies such that the saddlepoint has a dimension larger than one, i.e. such that $\lambda_2 < 1$. In these economies, indeterminancy and sunspots will be two facets of the same coin. As shown in the appendix, the reasons for which $\lambda_2 > 1$ relate to the classical assumptions about the shape of the utility function u and the production function y. We now modify the production function, to see the effect on the eigenvalues of A.

[Economy with increasing returns]

[Asset pricing implications in further chapters]

3.4 Production-based asset pricing

3.4.1 Firms

For each firm, capital accumulation does satisfy the identity in Eq. (3.18), reproduced here for convenience:

$$K_{t+1} = (1-\delta) K_t + I_t. \tag{3.27}$$

The additional assumption we make, is that capital adjustment is costly: investing I_t per unit of capital already in place, K_t, entails a cost $\phi(\frac{I_t}{K_t})$, expressed in terms of the price of the final good, which we take to be the numéraire, thereby allowing the investment goods to differ from the final good the firm produces. An investment of I_t, then, leads to a cost $\phi(\frac{I_t}{K_t})K_t$, such that the profit the firm makes at time t is,

$$D(K_t, I_t) \equiv \tilde{y}(K_t, N(K_t)) - w_t N(K_t) - p_t I_t - \phi\left(\frac{I_t}{K_t}\right) K_t, \tag{3.28}$$

where $\tilde{y}(K_t, N_t)$ is the firm's production at time t, obtained with capital K_t and labor N_t, and subject to the productivity shocks described in Section 3.3.4, w_t is the real wage, $N(K)$ is the labor demand schedule, solution to the optimality condition, $\tilde{y}_N(K_t, N(K_t)) = w_t$ for all t, and p_t is the real price of the investment goods, or uninstalled capital. Finally, the adjustment-cost function satisfies $\phi \geq 0$, $\phi' \geq 0$, $\phi'' \geq 0$. In words, capital adjustment is costly when the adjustment is made fastly. Naturally, ϕ is zero in the absence of adjustment costs.

What is the *value* of the profit, from the perspective of time zero? This question can be answered, by utilizing the Arrow-Debreu state prices introduced in Chapter 2. At time t, and in state s, the profit $D_t(s)$ (say) is worth,

$$\phi_{0,t}(s) D(K_t(s), I_t(s)) = m_{0,t}(s) D_t(K_t(s), I_t(s)) P_{0,t}(s),$$

with the same notation as in Chapter 2.

3.4.1.1 The value of the firm

We assume that in each period, the firm distributes all the profits it makes, and that for a given capital K_0, it maximizes its *cum-dividend* value,

$$V_c(K_0) = \max_{(K_t, I_{t-1})_{t=1}^{\infty}} \left[D(K_0, I_0) + E\left(\sum_{t=1}^{\infty} m_{0,t} D(K_t, I_t) \right) \right],$$

subject to the capital accumulation law of Eq. (3.27).

The value of the firm at time t, $V_c(K_t)$, can be found recursively, through the Bellman's equation,

$$V_c(K_t) = \max_{I_t} \left[D(K_t, I_t) + E_t\left(m_{t+1} V_c(K_{t+1}) \right) \right],$$

where the expectation is taken with respect to the information set as of time t. The first-order conditions for I_t lead to,

$$-D_I(K_t, I_t) = E_t\left[m_{t+1} V_c'(K_{t+1}) \right]. \tag{3.29}$$

That is, along the optimal capital accumulation path, the marginal cost of new installed capital at time t, $-D_I$, must equal the expected marginal return on the investment, i.e. the expected value of the marginal contribution of capital to the value of the firm at time $t+1$, $V_c'(K_{t+1})$.

By Eq. (3.29), optimal investment is a function $I(K_t)$, and the value of the firm satisfies,

$$V_c(K_t) = D(K_t, I(K_t)) + E_t[m_{t+1} V_c(K_{t+1})].$$

Differentiating the value function in the previous equation, with respect to K_t, and using Eq. (3.29), yields the following envelope condition:

$$\begin{aligned} V_c'(K_t) &= D_K(K_t, I(K_t)) + D_I(K_t, I(K_t)) I'(K_t) + E_t[m_{t+1} V_c'(K_{t+1})((1-\delta) + I'(K_t))] \\ &= D_K(K_t, I(K_t)) - (1-\delta) D_I(K_t, I(K_t)). \end{aligned}$$

By replacing this expression for the value function back into Eq. (3.29), leaves:

$$-D_I(K_t, I(K_t)) = E_t[m_{t+1}(D_K(K_{t+1}, I(K_{t+1})) - (1-\delta) D_I(K_{t+1}, I(K_{t+1})))]. \quad (3.30)$$

Along the optimal capital accumulation path, the marginal cost of new installed capital at time t, which by Eq. (3.29) is the expected marginal return on the investment, equals the expected value of (i) the very same marginal cost at time $t+1$, corrected for capital depreciation, $(1-\delta)$, plus (ii) capital productivity, net of adjustment costs. Analytically,

$$\begin{aligned} D_K(K_t, I_t) &= \tilde{y}_K(K_t, N(K_t)) + \tilde{y}_N(K_t, N(K_t)) N'(K_t) - w_t N'(K_t) - \frac{\partial}{\partial K}\left(\phi\left(\frac{I_t}{K_t}\right) K_t\right) \\ &= \tilde{y}_K(K_t, N(K_t)) - \frac{\partial}{\partial K}\left(\phi\left(\frac{I_t}{K_t}\right) K_t\right), \\ -D_I(K_t, I(K_t)) &= p_t + \phi'\left(\frac{I_t}{K_t}\right). \end{aligned}$$

We now introduce a fundamental concept in investment theory.

3.4.1.2 q theory

The Tobin's *marginal* q is defined as the ratio of the expected marginal value of an additional unit of capital over its replacement cost:

$$\text{TQ}_t \equiv \text{Tobin's marginal q} \equiv \frac{E[m_{t+1} V_c'(K_{t+1})]}{p_t}.$$

We show that the numerator, $E[m_{t+1} V_c'(K_{t+1})]$, is, simply, the shadow price of installed capital. Consider the Lagrangian at time t,

$$V_c(K_t) = \max_{I_t, K_{t+1}, q_t} [D(K_t, I_t) - q_t(K_{t+1} - (1-\delta) K_t - I_t) + E_t(m_{t+1} V_c(K_{t+1}))], \quad (3.31)$$

which, integrated, gives rise to the value of the firm:

$$V_c(K_0) = \max_{(I_t, K_{t+1} q_t)_{t=0}^{\infty}} E\left[\sum_{t=0}^{\infty} m_{0,t}(D(K_t, I_t) - q_t(K_{t+1} - (1-\delta) K_t - I_t))\right].$$

The first-order condition for investment, I_t, is, $q_t = -D_I(K_t, I_t) = E(m_{t+1} V_c'(K_{t+1}))$, where the second equality follows by Eq. (3.29). Therefore, q_t is the expected marginal return on the

investment, that is, the shadow price of installed capital. Tobin's marginal q is, then, the ratio of the shadow price of installed capital to its replacement cost:

$$\mathrm{TQ}_t = \frac{q_t}{p_t}.$$

Replacing the condition $q_t = -D_I(K_t, I_t)$ into the valuation equation (3.30) leaves:

$$q_t = E_t\left[m_{t+1}\left(D_K\left(K_{t+1}, I_{t+1}\right) + (1-\delta)\, q_{t+1}\right)\right], \tag{3.32}$$

that is,

$$q_t = p_t + \phi'\left(\frac{I_t}{K_t}\right). \tag{3.33}$$

The shadow price of installed capital, q_t, has to equal the marginal cost of new installed capital, and is larger than the price of uninstalled capital, p_t. It is natural: to install new capital requires some (marginal) adjustment costs, which add to the "row" price of uninstalled capital, p_t. Therefore, in the presence of adjustment costs, Tobin's marginal q is larger than one.

Eq. (3.32) can be solved forward, leaving:

$$q_t = E\left[\sum_{s=1}^{\infty}(1-\delta)^{s-1} m_{0,t+s} D_K\left(K_{t+s}, I_{t+s}\right)\right].$$

The shadow price of installed capital is worth the sum of all its future marginal net productivity, discounted at the depreciation rate. Moreover, Eq. (3.33) can be inverted for I_t/K_t, to deliver:

$$\frac{I_t}{K_t} = \phi'^{-1}\left(q_t - p_t\right), \tag{3.34}$$

where ϕ'^{-1} denotes the inverse of ϕ', and is increasing, since ϕ' is increasing. Given K_t, and the fact that K_{t+1} is pre-determined, the firm evaluates q_t through Eq. (3.32), and then determines the level of new investments through Eq. (3.34). These investments are increasing in the difference between the shadow price of installed capital, q_t, and that of uninstalled capital, p_t, as originally assumed by Tobin (1969).

In the absence of adjustement costs, when $q_t = p_t$, Eq. (3.32) delivers the condition,

$$1 = E_t\left[m_{t+1}\left(\tilde{y}_K\left(K_{t+1}, N\left(K_{t+1}\right)\right) + (1-\delta)\right)\right],$$

where we have set $p_t \equiv 1$ for all t, meaning that the firm's production is just the uninstalled capital. Empirically, however, the marginal productivity of capital, $\tilde{y}_K\left(K_t, N\left(K_t\right)\right)$, is not volatile enough, to rationalize asset returns, as explained in more detail in Chapters 8 and 9. Moreover, as we argue in a moment, Tobin's marginal q can be approximated by market-to-book ratios, which are typically time-varying. Therefore, adjustment costs are important for asset pricing.

A difficulty with Tobin's marginal q is that it is quite difficult to estimate. Yet in the special case we are analyzing in this section, where firms act competitively and have access to an homogeneous production function and adjustment costs, Tobin's marginal q can be proxied by the market-to-book ratio of a given firm. Let $V\left(K_t\right)$ denote the ex-dividend value of the firm, which is its stock market value, since it nets out the dividend it pays to its holder in the current period. It is:

$$V\left(K_t\right) \equiv V_c\left(K_t\right) - D\left(K_t, I\left(K_t\right)\right) = E_t\left[m_{t+1} V_c\left(K_{t+1}\right)\right].$$

The Tobin's *average* q is defined as the ratio of the stock market value of the firm over the replacement cost of the capital:

$$\text{Tobin's average q} \equiv \frac{\text{Stock Mkt Value of the Firm}}{\text{Replacement Cost of Capital}} = \frac{V(K_t)}{p_t K_{t+1}}.$$

The next result was originally obtained by Hayashi (1982) in a continuous-time setting.

THEOREM 3.2. *Tobin's marginal q and average q coincide. That is, we have,*

$$V(K_t) = q_t K_{t+1}.$$

PROOF. By the homogeneity properties of the production function and the adjustment costs,

$$D(K_t, I_t) = D_K(K_t, I_t) K_t + D_I(K_t, I_t) I_t.$$

Therefore, the ex-dividend value of the firm is:

$$\begin{aligned} V(K_0, I_0) &= E\left[\sum_{t=1}^{\infty} m_{0,t} D(K_t, I_t)\right] \\ &= E\left[\sum_{t=1}^{\infty} m_{0,t} \left(D_K(K_t, I_t) - (1-\delta) D_I(K_t, I_t)\right) K_t\right] + E\left[\sum_{t=1}^{\infty} m_{0,t} D_I(K_t, I_t) K_{t+1}\right], \end{aligned}$$

where the second line follows by Eq. (3.27). By Eq. (3.30), and the law of iterated expectations,

$$E\left[\sum_{t=1}^{\infty} m_{0,t} \left(D_K(K_t, I_t) - (1-\delta) D_I(K_t, I_t)\right) K_t\right] = -D_I(K_0, I_0) K_1 - E\left[\sum_{t=1}^{\infty} m_{0,t} K_{t+1} D_I(K_t, I_t)\right].$$

Hence, $V(K_0, I_0) = -D_I(K_0, I_0) K_1 = q_0 K_1$. ‖

This result, in conjunction with that in Eq. (3.33), provides a simple rule of thumb for investement decisions. Consider, for example, the case of quadratic adjustment costs, where $\phi(x) = \frac{1}{2}\kappa^{-1} x^2$, for some $\kappa > 0$. Then, Eq. (3.34) is:

$$I_t = \kappa(q_t - p_t) K_t = \kappa\left(\frac{\text{Stock Mkt Value of the Firm}}{\text{Replacement Cost of Capital}} - 1\right) p_t K_t,$$

where the second equality follows by Theorem 3.2. Thus, according to q theory, we expect firms with a market value larger than the cost of reproducing their capital to grow, and firms which are not worth the cost of reproducing their capital to shrink.

3.4.2 Consumers

We now generalize the budget constraint obtained in the program [3.P3], to the uncertainty case. We claim that in this case, the relevant budget constraint is,

$$V_0 = c_0 + E\left[\sum_{t=1}^{\infty} m_{0,t} (c_t - w_t N_t)\right]. \tag{3.35}$$

We have: $c_t + S_t\theta_{t+1} = (S_t + D_t)\,\theta_t + w_t N_t$ and, then:

$$\begin{aligned} E\left[\sum_{t=1}^{\infty} m_{0,t}\,(c_t - w_t N_t)\right] &= E\left[\sum_{t=1}^{\infty} m_{0,t}\,(S_t + D_t)\,\theta_t\right] - E\left[\sum_{t=1}^{\infty} m_{0,t} S_t \theta_{t+1}\right] \\ &= E\left[\sum_{t=1}^{\infty} \underset{t-1}{E}\left(\frac{m_{0,t}}{m_{t-1,t}} m_{t-1,t}\,(S_t + D_t)\,\theta_t\right)\right] - E\left[\sum_{t=2}^{\infty} m_{0,t-1} S_{t-1}\theta_t\right] \\ &= E\left[\sum_{t=1}^{\infty} m_{0,t-1} S_{t-1}\theta_t\right] - E\left[\sum_{t=2}^{\infty} m_{0,t-1} S_{t-1}\theta_t\right] \\ &= S_0\theta_1 = V_0 - c_0. \end{aligned}$$

where the third line follows by the properties of the discount factor, $\frac{m_{0,t}}{m_{t-1,t}} = m_{0,t-1}$ and $m_t \equiv m_{t-1,t}$.

Therefore, the program consumers solve is:

$$\max_{(c_t)_{t=0}^{\infty}} E\left[\sum_{t=1}^{\infty} \beta^t u\,(c_t)\right], \quad \text{s.t. Eq. (3.35).}$$

We now have two optimality conditions, one intertemporal and another, intratemporal:

$$m_{t+1} = \beta \frac{u_1\,(c_{t+1}, N_{t+1})}{u_1\,(c_t, N_t)} \quad (\textit{inter}\text{temporal}); \quad w_t = -\frac{u_2\,(c_t, N_t)}{u_1\,(c_t, N_t)} \quad (\textit{intra}\text{temporal}).$$

3.4.3 Equilibrium

For all t,

$$\tilde{y}\,(K_t, N_t) = c_t + p_t I_t + \phi\left(\frac{I_t}{K_t}\right) K_t. \tag{3.36}$$

It is easily seen that the condition $\theta_t = 1$ in the financial market, implies that $c_t = D_t + w_t N_t$, which, upon substitution of the profits in Eq. (3.28), delivers the equilibrium condition in Eq. (3.36). Implicit in this reasoning, is the idea the adjustment costs are not paid to anyone. They represent, so to speak, capital losses incurred along the way of growth.

3.5 Money, production and asset prices in overlapping generations models

3.5.1 Introduction: endowment economies

3.5.1.1 A deterministic model

We initially assume the population is constant, and made up of one young and one old. The young agent maximizes his intertemporal utility subject to his budget constraint:

$$\max_{(c_{1t}, c_{2,t+1})} \left[u\,(c_{1t}) + \beta u\,(c_{2,t+1})\right] \quad \text{subject to} \quad \begin{cases} \text{sav}_t + c_{1t} = w_{1t} \\ c_{2,t+1} = \text{sav}_t R_{t+1} + w_{2,t+1} \end{cases} \tag{3.P5}$$

where w_{1t} and $w_{2,t+1}$ are the endowments the agent receives at his young and old age.

The agent born at time $t-1$, then, faces the constraints: $\text{sav}_{t-1} + c_{1,t-1} = w_{1,t-1}$ and $c_{2t} = \text{sav}_{t-1}R_t + w_{2t}$. By combining his second period constraint with the first period constraint of the agent born at time t,

$$\text{sav}_{t-1}R_t + w_t = \text{sav}_t + c_{1t} + c_{2t}, \quad w_t \equiv w_{1t} + w_{2t}. \tag{3.37}$$

The equilibrium in the intergenerational lending market is, naturally:

$$\text{sav}_t = 0, \tag{3.38}$$

and implies that the goods market is also in equilibrium, in that $w_t = \sum_{i=1}^{2} c_{i,t}$, and for all t. Therefore, we can analyze the model, by just analyzing the autarkic equilibrium.

As Figure 3.4 illustrates, the first-order condition for the program [3.P5] requires that the slope of the indifference curve be equal to the slope of the lifetime budget constraint, $c_{2,t+1} = -R_{t+1}c_{1,t} + R_{t+1}w_{1t} + w_{2,t+1}$, and leads to:

$$\beta \frac{u'(c_{2,t+1})}{u'(c_{1,t})} = \frac{1}{R_{t+1}}. \tag{3.39}$$

The equilibrium, then, is a sequence of gross returns R_t satisfying Eqs. (3.37), (3.38) and (3.39), or:

$$b_t \equiv \frac{1}{R_{t+1}} = \beta \frac{u'(w_{2,t+1})}{u'(w_{1t})}. \tag{3.40}$$

Note that b_t is the shadow price of a bond issued at t, i.e., an asset that pays off one unit of numéraire at $t+1$. Then, the sequence of prices, b_t, satisfying Eq. (3.40), is such that agents are happy with not being able to lend and to borrow, intergenerationally.

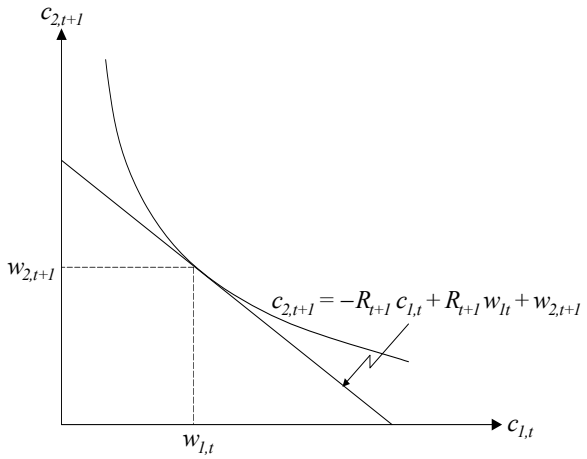


FIGURE 3.4.

The previous model is easy to extend to the case where agents are heterogeneous. The program each agent j solves is, now:

$$\max_{(c_{1j,t},c_{2j,t+1})} \left[u_j\left(c_{1j,t}\right) + \beta_j u_j\left(c_{2j,t+1}\right)\right] \quad \text{subject to} \quad \begin{cases} \text{sav}_{j,t} + c_{1j,t} = w_{1j,t} \\ c_{2j,t+1} = \text{sav}_{j,t}R_{t+1} + w_{2j,t+1} \end{cases}$$

with obvious notation. The first-order condition is, for all t and each agent j,

$$\beta_j \frac{u_j'\left(c_{2j,t+1}\right)}{u_j'\left(c_{1j,t}\right)} = \frac{1}{R_{t+1}} \equiv b_t,$$

and the equilibrium is a sequence of bond prices b_t satisfying the previous relation and the equilibrium in the intrageneration lending market:

$$\sum_{j=1}^{J} \text{sav}_{j,t} = 0, \tag{3.41}$$

where J denotes the constant number of agents in each generation.

To illustrate, suppose agents have all the same utility CRRA equal to γ, and the same discount rate, $\beta_j = \beta$. In this case,

$$\text{sav}_{j,t} = \frac{(\beta R_{t+1})^{\frac{1}{\gamma}} w_{1t} - w_{2,t+1}}{R_{t+1} + (\beta R_{t+1})^{\frac{1}{\gamma}}}.$$

The first term in the numerator reflects an income effect, while the second is a substitution effect. The coefficient $\frac{1}{\gamma}$ is the elasticity of intertemporal substitution, as explained in Section 3.2.3. For example, in the logarithmic case, $\gamma = 1$, we have

$$c_{1j,t} = \frac{1}{1+\beta}\left(w_{1j,t} + \frac{w_{2j,t+1}}{R_{t+1}}\right),\ c_{2j,t+1} = \frac{\beta}{1+\beta}\left(R_{t+1}w_{1j,t} + w_{2j,t+1}\right),\ \text{sav}_{j,t} = \frac{1}{1+\beta}\left(\beta w_{1j,t} - \frac{w_{2j,t+1}}{R_{t+1}}\right), \tag{3.42}$$

and using the equilibrium condition in Eq. (3.41),

$$b_t = \frac{1}{R_{t+1}} = \beta \frac{\sum_{j=1}^{J} w_{1j,t}}{\sum_{j=1}^{J} w_{2j,t+1}}. \tag{3.43}$$

3.5.1.2 A tree in a stochastic economy

Next, we introduce a tree that yields a stochastic dividend D_t in each period. Each agent solves the following program:

$$\max_{(c_{1t},c_{2,t+1})} \left[u\left(c_{1t}\right) + \beta E_t\left(u\left(c_{2,t+1}\right)\right)\right] \quad \text{subject to} \begin{cases} S_t\theta_t + c_{1t} = w_{1t} \\ c_{2,t+1} = (S_{t+1} + D_{t+1})\theta_t + w_{2,t+1} \end{cases} \tag{3.P6}$$

where S_t denotes the asset price, θ the units of the asset the agent chooses in his young age, E_t is the expectation taken conditional on information available at time t. The agent born at time $t-1$ faces the constraints $S_{t-1}\theta_{t-1} + c_{1,t-1} = w_{1,t-1}$ and $w_{2t} + (S_t + D_t)\theta_{t-1} = c_{2,t}$. By combining the second period constraint of the agent born at time $t-1$ with the first period constraint of the agent born at time t,

$$\left(S_t + D_t\right)\theta_{t-1} - S_t\theta_t + w_t = c_{1,t} + c_{2,t}.$$

The clearing condition in the asset market, $\theta_t = 1$, implies that the market for goods also clears, for all t: $D_t + w_{1t} + w_{2t} = c_{1,t} + c_{2,t}$. We can solve the program [3.P6] while using the constraints and substitute for $c_{1,t}$ and $c_{2,t+1}$,

$$\max_{\theta} \left[u\left(w_{1t} - S_t\theta\right) + \beta E_t\left(u\left(\left(S_{t+1} + D_{t+1}\right)\theta + w_{2,t+1}\right)\right)\right].$$

In equilibrium, $\theta_t = 1$, implying that (i) $c_{1t} = w_{1t} - S_t$ and (ii) $c_{2,t+1} = S_{t+1} + D_{t+1} + w_{2,t+1}$. Using (i) and (ii), the first-order condition for the program [3.P6] is then

$$u'\left(w_{1t} - S_t\right)S_t = \beta E_t\left[u'\left(S_{t+1} + D_{t+1} + w_{2,t+1}\right)\left(S_{t+1} + D_{t+1}\right)\right].$$

Consider, for example, the case where $u(c) = \ln c$, and set $\tilde{R}_{t+1} = (S_{t+1} + D_{t+1})/S_t$. We have:

$$\frac{1}{w_{1t} - \text{sav}_t^*} = \beta E_t \left(\frac{1}{\text{sav}_t^* \tilde{R}_{t+1} + w_{2,t+1}} \tilde{R}_{t+1} \right), \quad \text{where } \text{sav}_t^* \equiv S_t \theta_t, \ \ \theta_t = 1. \tag{3.44}$$

Note that in the previous deterministic setting,

$$\frac{1}{w_{1t} - \text{sav}_t} = \beta \frac{1}{\text{sav}_t R_{t+1} + w_{2,t+1}} R_{t+1}, \quad \text{where } \text{sav}_t \equiv 0, \tag{3.45}$$

which leads to the equilibrium bond price in Eq. (3.43). Eqs. (3.44) and (3.45) emanate from optimality conditions of a log-utility investors. Their difference is that in the tree economy, savings have to stay positive (the tree must be held by the young agent), such that in equilibrium, $\text{sav}_t^* \equiv S_t \geq 0$. Instead, in the previous economy *without* a tree, the interest rate, R_t, has to be such that savings are zero for all t, $\text{sav}_t = 0$.

Eq. (3.44) can be explicitly solved for the price of the tree, S_t, once we assume $w_{2t} = 0$ for all t. Note that in the absence of a tree, the endowments in the old age cannot be zero as the autarkic economy in this case would be such that the old generation would not consume anything. In the presence of a tree, this assumption is conceptually innocuous, as the autarkic equilibrium in this case is such that the old generation could consume the fruits of the tree (in addition to the proceedings arising from selling the tree to the young generation). Solving Eq. (3.44) for S_t when $w_{2t} = 0$, then, leads to the following price of the tree

$$S_t = \frac{\beta}{1+\beta} w_{1t}.$$

3.5.2 Diamond's model

$K_{t+1} = N_t S_t$, $S_t = S(r(K_t), w(K_t))$.

[Bubbles]

3.5.3 Money

We consider a version of the previous model with endowment (not with capital), and assume that agents can now transfer value through a piece of paper, interpreted as money. The young agent, then, maximizes his intertemporal utility, subject to a new budget constraint:

$$\max_{(c_{1t}, c_{2,t+1})} \left[u(c_{1t}) + \beta u(c_{2,t+1}) \right] \ \text{ subject to } \begin{cases} \dfrac{m_t}{p_t} + c_{1t} = w_{1t} \\ c_{2,t+1} = \dfrac{m_t}{p_{t+1}} + w_{2,t+1} \end{cases} \tag{3.P7}$$

where m_t is the amount of money he holds at time t, and p_t is the price of the consumption good as of time t.

Let

$$\text{sav}_t \equiv \frac{m_t}{p_t}, \quad R_{t+1} \equiv \frac{p_t}{p_{t+1}}. \tag{3.46}$$

Then, the budget constraint for program [3.P7] is *formally* identical to that for program [3.P5]. The difference is that in the monetary economy of this section, the young agent may wish to

transfer value over time, by saving money, earning a gross "interest rate" equal to the rate of deflation: the lower the price level the next period, the higher the purchasing power of the money he transfers from the young to the old age.

Naturally, then, by aggregating the budget constraints of the young and the old generation, Eq. (3.37) obtains although, now, sav_t and R_{t+1} are defined as in (3.46). However, sav_t is not necessarily zero in this monetary economy: money can be transferred from one generation to the next. Finally, in equilibrium, $\text{sav}_t = \frac{\bar{m}_t}{p_t}$, where $\bar{m}_t$ denotes money supply. Therefore, the real value of money (i.e., sav_t) is strictly positive once the equilibrium price p_t stays bounded over time, which can actually occur, as we shall see. It is apparent that in this context, the role of money (a medium for transferring value) is similar to that of a tree in the stochastic overlapping generations economy of Section 3.5.1.2.

Substituting the equilibrium savings $\text{sav}_t = \frac{\bar{m}_t}{p_t}$ and $R_{t+1} = \frac{p_t}{p_{t+1}}$ into Eq. (3.37), we obtain, $\bar{m}_{t-1} = \bar{m}_t + p_t (c_{1t} + c_{2t} - w_t)$, which used again in Eq. (3.37), leaves

$$\text{sav}_{t-1} R_t = \text{sav}_t - \frac{\Delta \bar{m}_t}{p_t}. \tag{3.47}$$

We need a law of movement for money creation. We assume that:[7] $\frac{\Delta \bar{m}_t}{\bar{m}_{t-1}} = \mu_t$, for some bounded sequence μ_t. Replacing this into Eq. (3.47),

$$(1 + \mu_t) \text{sav}_{t-1} R_t = \text{sav}_t. \tag{3.48}$$

Eq. (3.48) can be derived even more simply, while noting that by definition, $(1 + \mu_t) \frac{\bar{m}_{t-1}}{p_{t-1}} \frac{p_{t-1}}{p_t} = \frac{\bar{m}_t}{p_t}$. We can generalize Eq. (3.48) while assuming that population grows. Suppose that at time t, N_t individuals are born, and that $\frac{N_t}{N_{t-1}} = (1 + n)$, for some constant n. Let total money supply be defined as $M_t \equiv N_t \bar{m}_t$, and assume that for all t, $\frac{\Delta M_t}{M_{t-1}} = \mu_t$. Then, $(1 + \mu_t) N_{t-1} \frac{\bar{m}_{t-1}}{p_{t-1}} \frac{p_{t-1}}{p_t} = N_t \frac{\bar{m}_t}{p_t}$, or,

$$\frac{1 + \mu_t}{1 + n} \text{sav}(R_t) R_t = \text{sav}(R_{t+1}), \tag{3.49}$$

where we now emphasize that the real savings are a function of the interest rate, $\text{sav}_{t-1} \equiv \text{sav}(R_t)$, a property that arises while solving the program [3.P7].

Next, suppose that μ_t is independent of R, and that $\lim_{t \to \infty} \mu_t = \mu$, say, a constant. Eq. (3.49) leads to two stationary equilibria:

(a) $R = \frac{1+n}{1+\mu}$. This stationary equilibrium is known as the "Golden Rule" (once $\mu = 0$), as we shall say in Section 3.6.2). If $\mu \neq 0$, the price satisfies: $p_t = \left(\frac{1+\mu}{1+n}\right)^t p_0$. Then, we have: (i) $\frac{\bar{m}_t}{p_t} = \frac{M_t}{N_t p_t} = \frac{M_0}{N_0 p_0}$, and (ii) $\frac{\bar{m}_t}{p_{t+1}} = \frac{M_0}{N_0 p_0} \frac{1+n}{1+\mu}$. All in all, the agents' budget constraints are bounded and the real value of money is strictly positive. In this equilibrium, agents "trust" money.

(b) R_a : $\text{sav}(R_a) = 0$. This stationary equilibrium is an autarkic state. Generally, we have that $R_a < R$: prices increase more rapidly than per-capita money stocks. Analytically, $\text{sav}(R_a) = 0$ implies that $\lim_{t \to \infty} \frac{\bar{m}_t}{p_t} \to 0$, which, in turn, implies that for large t, $\frac{\bar{m}_{t+1}}{p_{t+1}} <$

[7] In this section, we assume that money transfers are made to the young generation: the money the young generation has to absorb is that from the old generation, $\bar{m}_{t-1}$, and that created by the "central bank," $\mu_t \bar{m}_{t-1}$. One might consider an alternative model in which transfers are made to old.

$\frac{\bar{m}_t}{p_t} \iff \frac{\bar{m}_{t+1}}{\bar{m}_t} = \frac{M_{t+1}/M_t}{N_{t+1}/N_t} = \frac{1+\mu}{1+n} < \frac{p_{t+1}}{p_t} \iff R_a < R$. As for $\frac{\bar{m}_t}{p_{t+1}}$, we have that $\frac{\bar{m}_t}{p_{t+1}} = \frac{\bar{m}_t}{p_t} R_a < \frac{\bar{m}_t}{p_t} R = \frac{\bar{m}_t}{p_t} \frac{1+n}{1+\mu}$, and since $\lim_{t\to\infty} \frac{\bar{m}_t}{p_t} \to 0$, then $\lim_{t\to\infty} \frac{\bar{m}_t}{p_{t+1}} \to 0$. In this stationary equilibrium, agents do not "trust" money.

If $\text{sav}(\cdot)$ is differentiable and $\text{sav}'(\cdot) \neq 0$, the dynamics of $(R_t)_{t=0}^{\infty}$ can be analyzed through the slope,

$$\frac{dR_{t+1}}{dR_t} = \frac{\text{sav}'(R_t)R_t + \text{sav}(R_t)}{\text{sav}'(R_{t+1})} \frac{1+\mu_t}{1+n}. \tag{3.50}$$

There are three cases:

(i) $\text{sav}'(R) > 0$. Gross substituability: the substitution effect dominates the income effect.

(ii) $\text{sav}'(R) = 0$. Income and substitution effects compensate with each other.

(iii) $\text{sav}'(R) < 0$. Complementarity: the income effect dominates the substitution effect.

The introductory example of this section leads to an instance of gross substituability (see Eq. (3.42)). Note that an equilibrium cannot exist in that economy, once we assume agents do not have endowments in the second period, $w_{2,t+1} = 0$, as in this case, savings would be strictly positive, such that the equilibrium condition in Eq. (3.41) would not hold.

These problems do not arise in the monetary economy of this section, in which savings can sustain a monetary equilibrium while being positive and equal to $\frac{\bar{m}_t}{p_t}$. Assume, for example, the Cobb-Douglas utility function, $u(c_{1t}, c_{2,t+1}) = c_{1t}^{l_1} \cdot c_{2,t+1}^{l_2}$, which leads to a real saving function equal to $\text{sav}(R_{t+1}) = \frac{\frac{l_2}{l_1} w_{1t} - \frac{w_{2,t+1}}{R_{t+1}}}{1+\frac{l_2}{l_1}}$. If $w_{2,t+1} = 0$, then, $\text{sav}(R_{t+1}) = \frac{\bar{m}_t}{p_t} = \frac{1}{\nu} w_{1t}$, $\nu \equiv \frac{l_1+l_2}{l_2}$ and, by reorganizing,

$$\bar{m}_t \nu = p_t w_{1t},$$

an equation consistent with the Quantitative Theory of money. In this economy, the sequence of gross returns satisfies, $R_{t+1} = \frac{p_t}{p_{t+1}} = \frac{\bar{m}_t}{\bar{m}_{t+1}} \frac{w_{1,t+1}}{w_{1,t}}$, or

$$R_{t+1} = \frac{(1+n)\cdot(1+g_{t+1})}{1+\mu_{t+1}}, \quad g_{t+1} \equiv \frac{w_{1,t+1}}{w_{1,t}} - 1.$$

Gross inflation, R_t^{-1}, equals the monetary creation factor, corrected for the growth rate of the economy as measured by g_{t+1}, the youngs' endowments growth rate.

As a final example, consider the utility function $u(c_{1t}, c_{2,t+1}) = \left(l c_{1t}^{(\gamma-1)/\gamma} + (1-l) c_{2,t+1}^{(\gamma-1)/\gamma} \right)^{\gamma/(\gamma-1)}$, which collapses to Cobb-Douglas once $\gamma \to 1$. We have:

$$c_{1t} = \frac{R_{t+1} w_{1t} + w_{2,t+1}}{R_{t+1} + K^{\gamma} R_{t+1}^{\gamma}}, \quad c_{2,t+1} = \frac{R_{t+1} w_{1t} + w_{2,t+1}}{1 + K^{-\gamma} R_{t+1}^{1-\gamma}}, \quad \text{sav}\left(R_{t+1}\right) = \frac{K^{\gamma} R_{t+1}^{\gamma} w_{1t} - w_{2,t+1}}{R_{t+1} + K^{\gamma} R_{t+1}^{\gamma}},$$

where $K \equiv \frac{1-l}{l}$. To simplify, set (i) $K = 1$, (ii) $w_{2t} = \mu_t = n = 0$, and (iii) $w_{1t} = w_{1,t+1}$. It can be shown that in this case, $\text{sign}\left(\text{sav}'(R)\right) = \text{sign}\left(\gamma - 1\right)$. Moreover, the dynamics of the gross interest rate, R, are given by:

$$R_{t+1} = f\left(R_t\right) \equiv \left(R_t^{-\gamma} + R^{-1} - 1\right)^{1/(1-\gamma)}. \tag{3.51}$$

The stationary equilibria are solutions to $R = f(R)$, and it is easily seen that one of them is $R = 1$, which corresponds to the monetary steady state. If $\gamma > 1$, the slope in Eq. (3.50) has always the same sign, and the mapping f in Eq. (3.51) has two fixed points, $R_a = 0$ and $R = 1$, with R_a being stable and R being unstable, as illustrated by Figure 3.5 in the case where $\gamma = 2$.

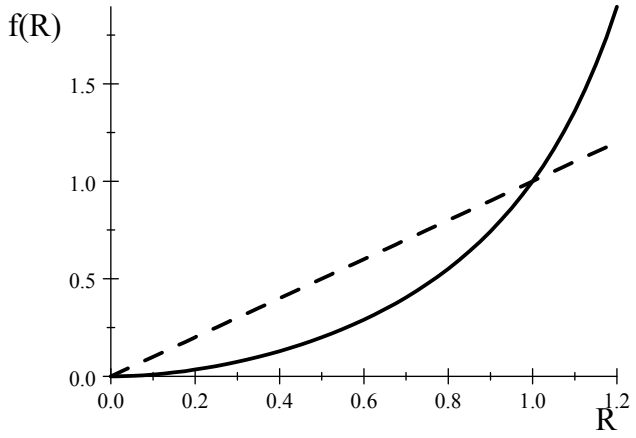


FIGURE 3.5. Stationary equilibria in Eq. (3.51), with $(R) = (R^{-2} + R^{-1} - 1)^{-1}$.

If $\gamma < 1$, R_a is not well-defined, and $R = 1$ is not necessarily unstable. We may have sequences of gross interest rates, R_t, converging towards R, or even the emergence of cycles. These properties can be understood by examining the slope of the map in Eq. (3.50), for $R = 1$, $\left.\frac{dR_{t+1}}{dR_t}\right|_{R_{t+1}=R_t=1} = \frac{\gamma}{\gamma-1}$.

Figure 3.6 depicts an hypothetical shape of the map $R_t \mapsto R_{t+1}$, which is that we might expect to arise in the presence of gross substituability and, at least provided $\frac{\text{sav}'(R)R}{\text{sav}(R)} < -1$ for all R, in the case of complementary. In both cases, the slope, $\frac{dR_{t+1}}{dR_t} > 0$. Moreover, the slope at the monetary state, $R = \frac{1+n}{1+\mu}$, is $\left.\frac{dR_{t+1}}{dR_t}\right|_{R_{t+1}=R_t=R} = 1 + \frac{\text{sav}(R)}{\text{sav}'(R)R} > 0$. The slope at the monetary state, the point M in Figure 3.6, is greater than one, provided $\text{sav}'(\cdot) > 0$. In this case, the monetary state M is unstable, while the autarkic state, the point A in Figure 3.6, is stable.

Note that any path beginning from the right of the monetary state leads to explosive dynamics for R. These dynamics cannot be part of any equilibrium because they would imply a decreasing sequence of prices, p, thereby tilting the agents' budget constraints in such a way to rule out the existence of a solution to the agents' programs. Therefore, the economy needs to starts anywhere between the point A and the point M, although then, we do not have any other piece of information: there exists, in fact, a continuum of points $R_1 \in [R_a, R)$ that are equally likely candidates to the beginning of the equilibrium sequence. Contrary to the representative agent models in the previous sections, the model of this section leads to an indeterminacy of the equilibrium, parametrized by the initial price p_0.

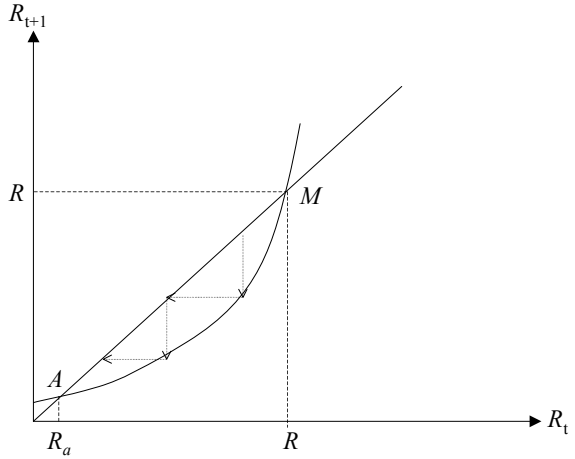


FIGURE 3.6. Gross substitutability

Is the autarkic equilibrium the only possible stable steady state? The answer is in the negative. Consider the case where the map $R_t \mapsto R_{t+1}$ bends backwards and is such that $\frac{dR_{t+1}}{dR_t}\Big|_{R_{t+1}=R_t=R} < -1$, such that the monetary steady state M is stable. A condition for the map $R_t \mapsto R_{t+1}$ to bend backward is that $\frac{\text{sav}'(R)R}{\text{sav}(R)} > -1$, and the condition for $\frac{dR_{t+1}}{dR_t}\Big|_{R_{t+1}=R_t=R} < -1$ to hold is that $\frac{\text{sav}'(R)R}{\text{sav}(R)} > -\frac{1}{2}$. In this case, the point M is reached from any sufficiently neighborhood of M. Figure 3.7 shows a cycle of order two, where $R_{**} = \frac{R^2}{R_*}$.[8]

Note, finally, that to analyze the behavior of the gross interest rate, we need to make reference to backward-looking dynamics, as there exists an indeterminacy of forward-looking dynamics. Finally, there might exist more complex situations where cycles of order 3 exist, giving rise to what is known as a "chaotic" system. Note that these complex dynamics, including those in Figure 3.7, rely on the assumption that $\text{sav}'(R) < 0$, which might be somehow unappealing.

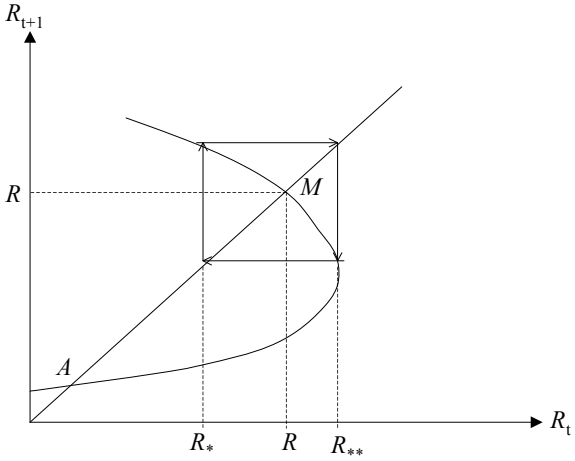


FIGURE 3.7.

[8]For the proof, note that by Eq. (3.49), we have, we have that for a cycle of order 2, (i) $\frac{1+\mu}{1+n}R_* s(R_*) = s(R_{**})$, and (ii) $\frac{1+\mu}{1+n}R_{**} s(R_{**}) = s(R_*)$. Multiplying the two equations side by side leaves the result that $R_{**} = R^2/R_*$.

3.5.4 Money in a model with real shocks

Lucas (1972) develops the first explanation regarding neutrality of money in contexts with overlapping generations and uncertainty. This section presents a simplified version of this model relying on Stokey et Lucas (1989, p. 504). Each agent lives two periods. When young, he works and produces a consumption good, and when old, he consumes. His disutility of work is $-v(n_t)$, where n_t is his labor supply, and v is assumed to satisfy $v', v'' > 0$. Utility drawn from second period consumption is $u(c_{t+1})$, and has the standard properties. He chooses how to work and consumes as follows:

$$\max_{\{n,c\}} \left[-v(n_t) + \beta E_t\left(u\left(c_{t+1}\right)\right)\right] \quad \text{subject to} \quad \begin{cases} m = p_t y_t, \quad y_t = \epsilon_t n_t \\ p_{t+1} c_{t+1} = m \end{cases}$$

where E_t denotes the expectation taken given the information at time t, m is money holdings; y_t is production, obtained through his labor supply n_t, $(\epsilon_t)_{t=0,1,...}$ is a sequence of positive shocks affecting his productivity and, finally, p_t is the price of the consumption good at t. By replacing the first constraint into the second leaves $c_{t+1} = \epsilon_t n_t R_{t+1}$, where $R_{t+1} \equiv \frac{p_t}{p_{t+1}}$, and the first-order conditions are

$$v'(n_t) = \beta E_t\left[u'(\epsilon_t n_t R_{t+1})\epsilon_t R_{t+1}\right].$$

By the equilibrium in the good markets, $\epsilon_{t+1} n_{t+1} = c_{t+1} = \epsilon_t n_t R_{t+1}$, such that the first-order conditions can be re-written as

$$v'(n_t) n_t = \beta E_t\left[u'(\epsilon_t n_t R_{t+1})\epsilon_{t+1} n_{t+1}\right]. \tag{3.52}$$

A rational expectation equilibrium is a labor supply $n_t = g(\epsilon_t)$, where g satisfies Eq. (3.52), i.e.,

$$v'\left(g(\epsilon)\, g(\epsilon)\right) = \beta \int_{\mathcal{E}} u'\left(\tilde{\epsilon} g(\tilde{\epsilon})\right) \tilde{\epsilon} g(\tilde{\epsilon}) dP(\tilde{\epsilon}|\, \epsilon),$$

where $\mathcal{E}$ denotes the support of $\tilde{\epsilon}$. This equation simplifies as soon as productivity shocks are IID, $P(\tilde{\epsilon}|\, \epsilon) = P(\tilde{\epsilon})$, in which case, $v'\left(g(\epsilon)\, g(\epsilon)\right)$ is independent of $\tilde{\epsilon}$, and n is a constant $\bar{n}$.[9] This is a result about the neutrality of money, at least provided such a constant $\bar{n}$ exists. Precisely, we have that $v'(\bar{n}) = \beta \int_{\mathcal{E}} u'\left(\tilde{\epsilon}\bar{n}\right) \tilde{\epsilon} dP(\tilde{\epsilon})$. For example, consider $v(x) = \frac{1}{2}x^2$ and $u(x) = \ln x$, in which case $\bar{n} = \sqrt{\beta}$, $y(\epsilon) = \epsilon\sqrt{\beta}$ and $p(\epsilon) = \frac{m}{\epsilon\sqrt{\beta}}$.

3.5.5 Sunspots

[In progress]

3.6 Optimality

3.6.1 Production economies

Consider the usual law of capital accumulation: $K_{t+1} = S_t = Y(K_t, N_t) - C_t$, for K_0 given. Dividing both sides of this equation by N_t leaves:

$$k_{t+1} = \frac{1}{1+n}\left(y(k_t) - c_t\right), \quad \text{for } k_0 \text{ given.} \tag{3.53}$$

[9] The proof that $g(\epsilon) = \bar{n}$ relies on the following argument. Suppose the contrary, i.e. there exists a point ϵ_0 and a neighborhood of ϵ_0 such that either (i) $g(\epsilon_0 + A) > g(\epsilon_0)$ or (ii) $g(\epsilon_0 + A) < g(\epsilon_0)$, for some strictly positive constant A. We deal with the proof of (i) as the proof of (ii) is nearly identical. Since $v'\left(g(\epsilon)\, g(\epsilon)\right)$ is constant, and $v'' > 0$, we have that $v'\left(g(\epsilon_0 + A)\right) g(\epsilon_0 + A) = v'\left(g(\epsilon_0)\right) g(\epsilon_0) \leq v'\left(g(\epsilon_0 + A)\right) g(\epsilon_0)$. Therefore, $v'\left(g(\epsilon_0 + A)\right)\left(g(\epsilon_0 + A) - g(\epsilon_0)\right) < 0$. Next, note that $v' > 0$, such that $g(\epsilon_0 + A) < g(\epsilon_0)$, contradicting that $g(\epsilon_0 + A) > g(\epsilon_0)$.

The stationary state of the economy is achieved when $k_{t+1} = k_t \equiv k$ and $c_{t+1} = c_t \equiv c$, such that:

$$c = y(k) - (1+n)k.$$

In steady-state, per-capita consumption attains its maximum at:

$$k_* : y'(k_*) = 1 + n. \tag{3.54}$$

The steady state per-capita capital satisfying Eq. (3.54) is said to satisfy the *Golden Rule*. A social planner would be able to increase per-capita consumption at the stationary state, provided $y'(k) < 1 + n$. Indeed, because $y(k)$ is given, we can lower k and have $dc = -(1+n)\,dk > 0$, immediately, and $dc = (y'(k) - (1+n))\,dk > 0$, in the next periods. In fact, this outcome would apply along the entire capital accumulation path of the economy, not only in steady state, as we now illustrate. First, a definition. We say that a path $(k, c)_{t=0}^{\infty}$ is consumption-inefficient if there exists another path $(\tilde{k}, \tilde{c})_{t=0}^{\infty}$ satisfying Eq. (3.53), and such that $\tilde{c}_t \geq c_t$ for all t, with at least a strict inequality for one t. The following is a slightly less general version of Theorem 1 in Tirole (1988, p. 161):

THEOREM 3.3 (Cass-Malinvaud theory). *A path $(k, c)_{t=0}^{\infty}$ is: (i) consumption efficient if $\frac{y'(k_t)}{1+n} \geq 1$ for all t, and (ii) consumption inefficient if $\frac{y'(k_t)}{1+n} < 1$ for all t.*

PROOF. As for Part (i), suppose k_t is consumption efficient, ad let $\tilde{k}_t = k_t + \epsilon_t$ be an alternative consumption efficient path. Since k_0 is given, $\epsilon_t = 0$. Moreover, by Eq. (3.53),

$$(1+n) \cdot (\tilde{k}_{t+1} - k_{t+1}) = y(\tilde{k}_t) - y(k_t) - (\tilde{c}_t - c_t),$$

and because $\tilde{k}$ is consumption-efficient, $\tilde{c}_t \geq c_t$, with at least one strictly equality for some t. Therefore, by concavity of y, and the definition of $\tilde{k}_t$,

$$0 \leq y(\tilde{k}_t) - y(k_t) - (1+n)(\tilde{k}_{t+1} - k_{t+1}) < y(k_t) + y'(k_t)\epsilon_t - y(k_t) - (1+n)\epsilon_{t+1},$$

or $\epsilon_{t+1} < \frac{y'(k_t)}{1+n}\epsilon_t$. Evaluating this inequality at $t = 0$ yields $\epsilon_1 < \frac{y'(k_0)}{1+n}\epsilon_0$, and since $\epsilon_0 = 0$, one has that $\epsilon_1 < 0$. Since $\frac{y'(k_t)}{1+n} \geq 1$ for all t, then $\epsilon_t \to -\infty$ as $t \to \infty$, which contradicts k_t has bounded trajectories. The proof of Part (ii) is nearly identical, except that, obviously, in this case, $\liminf \epsilon_t >> -\infty$. Note, in general, there are infinitely many sequences that allow for efficiency improvements. ‖

The reasoning in this section holds independently of whether the economy has a finite number of agents living forever, or overlapping generations. For example, in the case of overlapping generations, Eq. (3.53) is the capital accumulation path for Diamond's model, once we set $c_t \equiv \frac{C_t}{N_t} = c_{1t} + \frac{c_{2,t+1}}{1+n}$. An important issue is to establish whether actual economies are dynamically efficient? In a seminal contribution, Abel, Mankiw, Summers and Zeckhauser (1989) provide a framework to address this issue and conclude about dynamic efficiency.

[In progress]

3.6.2 Over-accumulation of capital

[In progress]

3.6.3 Bubbles

Diamond's model

[In progress]

3.6.4 Money

We wish to find first-best optima, that is, equilibria that a social planner may choose, by acting directly on agents' consumption, without needing to force the agents to make use of money.[10] Let us analyze, first, the stationary state, $R = \frac{1+n}{1+\mu}$. We show that this state corresponds to the stationary state where consumptions and endowments are constants, and that the agents' utility is maximized when $\mu = 0$. Indeed, since the social planner allocates resources without having regard to money, the only constraint is: $w_n \equiv w_1 + \frac{w_2}{1+n} = c_1 + \frac{c_2}{1+n}$, such that the utility of the "stationary agent" is:

$$u(c_1, c_2) = u\left(w_n - \frac{c_2}{1+n}, c_2\right).$$

The first-order condition is $\frac{u_{c2}}{u_{c1}} = \frac{1}{1+n}$. Instead, the first-order condition in the market equilibrium is $\frac{u_{c2}}{u_{c1}} = \frac{1}{R}$. Therefore, the Golden Rule is attained in the market equilibrium, if and only if $\mu = 0$. The social planner policy converges towards the Golden Rule. Indeed, the social planner solves:

$$\max \sum_{t=0}^{\infty} \vartheta^t u^{(t)}(c_{1t}, c_{2,t+1}), \quad \text{subject to } w_{nt} \equiv w_{1t} + \frac{w_{2t}}{1+n} = c_{1t} + \frac{c_{2t}}{1+n},$$

or $\max \sum_{t=0}^{\infty} \vartheta^t u^{(t)}\left(w_{nt} - \frac{c_{2t}}{1+n}, c_{2,t+1}\right)$, where ϑ is the weight the planner gives to the generation as of time t, and the notation $u^{(t)}$ is meant to emphasize that that endowments may change from one generation to another. The first-order conditions, $\frac{u_{c2}^{(t-1)}}{u_{c1}^{(t)}} = \frac{\vartheta}{1+n}$, lead to the "modified" Golden Rule in stead-state state (modified by the weight ϑ).

[10] In a second-best equilibrium, a social planner would let the market "play" first, by allowing the agents to use money and, then, would parametrize such virtual equilibria by μ_t. The indirect utility functions that arise as a result would then be expressed in terms of these growth rates μ_t. The social planner would then maximize an an aggregator of these utilities with respect to μ_t.

3.7 Appendix 1: Finite difference equations, with economic applications

Let $z_0 \in \mathbb{R}^d$, and consider the following linear system of finite difference equations:

$$z_{t+1} = A \cdot z_t, \quad t = 0, 1, \cdots, \tag{3A.1}$$

for some matrix A. The solution to Eq. (3A.1) is

$$z_t = v_1 \kappa_1 \lambda_1^t + \cdots + v_d \kappa_d \lambda_d^t, \tag{3A.2}$$

where λ_i and v_i are eigenvalues and eigenvectors of A, and κ_i are constants, which will be determined below.

A standard proof of this result relies on the so-called diagonalization of Eq. (3A.1). Consider the system of characteristic equations for A, $(A - \lambda_i I)\, v_i = \mathbf{0}_{d\times 1}$, where λ_i is scalar and v_i a $d \times 1$ column vector, for $i = 1, \cdots, n$, or, in matrix form, $AP = P\Lambda$, where $P = (v_1, \cdots, v_d)$ and Λ is a diagonal matrix with λ_i on its diagonal. We assume that $P^\top = P^{-1}$. By post-multiplying by P^{-1} leaves the *spectral decomposition* of A:

$$A = P\Lambda P^{-1}. \tag{3A.3}$$

By replacing Eq. (3A.3) into Eq. (3A.1), and rearranging terms,

$$y_{t+1} = \Lambda \cdot y_t, \text{ where } y_t \equiv P^{-1} z_t.$$

The solution for y is $y_{it} = \kappa_i \lambda_i^t$, and the solution for z is: $z_t = P y_t = (v_1, \cdots, v_d) y_t = \sum_{i=1}^d v_i y_{it} = \sum_{i=1}^d v_i \kappa_i \lambda_i^t$, which is Eq. (3A.2).

To determine the vector of constants $\kappa = (\kappa_1, \cdots, \kappa_d)^\top$, evaluate the solution at $t = 0$,

$$z_0 = (v_1, \cdots, v_d)\kappa = P\kappa,$$

whence

$$\hat{\kappa} \equiv \kappa(P) = P^{-1} z_0, \tag{3A.4}$$

where the columns of P are vectors belonging to the space of the eigenvectors. Naturally, there are many such vectors. However, the previous formula shows how the constants $\kappa(P)$ need to "adjust" so as to guarantee the stability of the solution with respect to changes in P.

3A.1 Example. Let $d = 2$, and suppose that $\lambda_1 \in (0, 1)$, $\lambda_2 > 1$. The resulting system is unstable for any initial condition, except perhaps for a set with measure zero. This set of measure zero gives rise to the so-called saddlepoint path. We can determine the coordinates of this set. We wish to find the set of initial conditions such that $\kappa_2 = 0$, which, in turn, rules out an explosive system behavior due to the unstable root, $\lambda_2 > 1$. The solution at $t = 0$ is

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = z_0 = P\kappa = (v_1, v_2) \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} v_{11}\kappa_1 + v_{12}\kappa_2 \\ v_{21}\kappa_1 + v_{22}\kappa_2 \end{pmatrix},$$

where we have set $z = (x, y)^\top$. By replacing the second equation into the first, and solving for κ_2, yields:

$$\kappa_2 = \frac{v_{11} y_0 - v_{21} x_0}{v_{11} v_{22} - v_{12} v_{21}},$$

which is zero when

$$y_0 = \frac{v_{21}}{v_{11}} x_0.$$

For this system, the saddlepoint is a line with a slope equal to the ratio of the two components of eigenvector for λ_1—the stable root. Figure 3A.1 depicts the phase diagram for this system, with the "divergent" line satisfying the equation $y_0 = \frac{v_{22}}{v_{12}} x_0$.

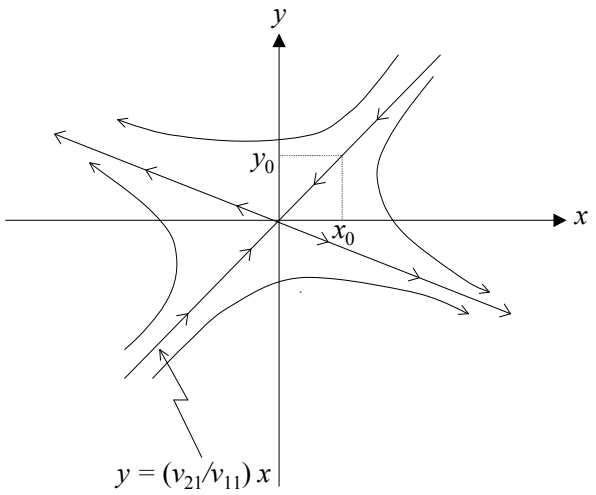


FIGURE 3A.1.

A saddlepoint path brings the following economic content. If x is a *pre-determined* variable, y must "jump" to the saddlepoint $y_0 = \frac{v_{21}}{v_{11}} x_0$, so as to ensure the system does not explode. Note that a conceptual difficulty arises should the system include two pre-determined variables, as in this case, there are no stable solutions in general. However, this possibility is unusual in economics. Consider the next example.

3A.2 Example. Example 3A.1 covers exactly the linearized version of the neoclassic growth model. Indeed, Section 3.3.3 shows that in a small neighborhood of the stationary state (k, c), the deviations $(\bar{k}_t, \bar{c}_t)_{t=0}^{\infty}$ of capital and consumption from (k, c), satisfy Eq. (3.15), reported for convenience here:

$$\begin{pmatrix} \bar{k}_{t+1} \\ \bar{c}_{t+1} \end{pmatrix} = A \begin{pmatrix} \bar{k}_t \\ \bar{c}_t \end{pmatrix}, \quad A \equiv \begin{pmatrix} y'(k) & -1 \\ -\frac{u'(c)}{u''(c)} y''(k) & 1 + \beta \frac{u'(c)}{u''(c)} y''(k) \end{pmatrix}.$$

The two eigenvalues of A are $\lambda_{1/2} = \frac{\mathrm{tr}(A) \mp \sqrt{\mathrm{tr}(A)^2 - 4\det(A)}}{2}$, where (i) $\det(A) = y'(k) = \beta^{-1} > 1$, and (ii) $\mathrm{tr}\,(A) = \beta^{-1} + 1 + \beta \frac{u'(c)}{u''(c)} y''(k) > 1 + \det(A)$. Note that

$$a \equiv \mathrm{tr}(A)^2 - 4\det(A) = \left(\beta^{-1} + 1 + \beta \frac{u'(c)}{u''(c)} y''(k) \right)^2 - 4\beta^{-1} > \left(\beta^{-1} + 1 \right)^2 - 4\beta^{-1} = \left(1 - \beta^{-1} \right)^2 > 0,$$

where the first inequality follows by the standard assumption made on utility, u, and production, y. Therefore, $\lambda_2 = \frac{1}{2}\left(\mathrm{tr}(A) + \sqrt{a}\right) > \frac{1}{2}\left(1 + \det(A) + \sqrt{a}\right) > 1 + \frac{1}{2}\sqrt{a} > 1$. Regarding λ_1, note that $\det(A) > 0$, such that $2\lambda_1 = \mathrm{tr}\,(A) - \sqrt{\mathrm{tr}(A)^2 - 4\det(A)} > 0$; moreover, $\lambda_1 < 1 \iff \mathrm{tr}\,(A) - \sqrt{\mathrm{tr}(A)^2 - 4\det(A)} < 2$, or $(\mathrm{tr}\,(A) - 2)^2 < \mathrm{tr}\,(A)^2 - 4\det(A)$, which can easily be verified. Hence, $\lambda_1 \in (0, 1)$.

Next, consider generalizing the previous examples to higher dimensional spaces. A *convergent*, or *stable subspace* is the counterpart to the saddlepoint in the $d = 2$ case. It is the locus of points such that z_t in Eq. (3A.2) does not explode. (In the case of nonlinear systems, this convergent subspace is termed *convergent*, or *stable manifold*.) Let $\Pi \equiv P^{-1}$, and rewrite Eq. (3A.4), i.e., the system determining the solution for κ, as follows

$$\hat{\kappa} = \Pi z_0.$$

We assume the elements of z and A are ordered in such a way that $\exists s : |\lambda_i| < 1$, for $i = 1, \cdots, s$ and $|\lambda_i| > 1$ for $i = s+1, \cdots, d$. Then, we partition Π as follows:

$$\hat{\kappa} = \begin{pmatrix} \underset{s\times d}{\Pi_{\bar{s}}} \\ \underset{(d-s)\times d}{\Pi_u} \end{pmatrix} z_0.$$

Proceeding similarly as in Example 3A.1, we search for conditions that ensure the system remains "trapped" in the convergent space: we require that $\hat{\kappa}_{s+1} = \cdots = \hat{\kappa}_d = 0$, or,

$$\begin{pmatrix} \hat{\kappa}_{s+1} \\ \vdots \\ \hat{\kappa}_d \end{pmatrix} = \underset{(d-s)\times d}{\Pi_u} z_0 = \mathbf{0}_{(d-s)\times 1}.$$

Let $d \equiv k+k^*$, where k^* is the number of pre-determined variables and k is the number of the remaining, or free, variables. Partition Π_u and z_0 in such a way to disentangle free from pre-determined variables, as follows:

$$\mathbf{0}_{(d-s)\times 1} = \underset{(d-s)\times d}{\Pi_u} z_0 = \begin{pmatrix} \underset{(d-s)\times k}{\Pi_u^{(1)}} & \underset{(d-s)\times k^*}{\Pi_u^{(2)}} \end{pmatrix} \begin{pmatrix} \underset{k\times 1}{z_0^{\text{free}}} \\ \underset{k^*\times 1}{z_0^{\text{pre}}} \end{pmatrix} = \underset{(d-s)\times k}{\Pi_u^{(1)}} \underset{k\times 1}{z_0^{\text{free}}} + \underset{(d-s)\times k^*}{\Pi_u^{(2)}} \underset{k^*\times 1}{z_0^{\text{pre}}},$$

or,

$$\underset{(d-s)\times k}{\Pi_u^{(1)}} \underset{k\times 1}{z_0^{\text{free}}} = - \underset{(d-s)\times k^*}{\Pi_u^{(2)}} \underset{k^*\times 1}{z_0^{\text{pre}}}. \tag{3A.5}$$

The k^*-dimensional vector z_0^{pre} is obviously known, being the vector of pre-determined variables, and $\Pi_u^{(1)}, \Pi_u^{(2)}$ depend on the primitives the economy embedded into the matrix A. Then, Eq. (3A.5) has $d-s$ equations with k unknowns, i.e., the components of z_0^{free}. We assume that $\Pi_u^{(1)}$ has full rank.

We refer s to as the dimension of the convergent subspace, $\mathcal{S}$, say. To explain this terminology, consider the solution for z_t,

$$z_t = v_1\hat{\kappa}_1\lambda_1^t + \cdots + v_s\hat{\kappa}_s\lambda_s^t + v_{s+1}\hat{\kappa}_{s+1}\lambda_{s+1}^t + \cdots + v_d\hat{\kappa}_d\lambda_d^t.$$

For z_t to remain stuck in $\mathcal{S}$, it must be that

$$\hat{\kappa}_{s+1} = \cdots = \hat{\kappa}_d = 0,$$

in which case,

$$\underset{d\times 1}{z_t} = \underset{d\times 1}{v_1}\,\hat{\kappa}_1\lambda_1^t + \cdots + \underset{d\times 1}{v_s}\,\hat{\kappa}_s\lambda_s^t = (v_1\hat{\kappa}_1, \cdots, v_s\hat{\kappa}_s)\cdot\left(\lambda_1^t, \cdots, \lambda_s^t\right)^\top,$$

i.e.,

$$\underset{d\times 1}{z_t} = \underset{d\times s}{\hat{V}} \cdot \underset{s\times 1}{\hat{\lambda}_t},$$

where $\hat{V} \equiv (v_1\hat{\kappa}_1, \cdots, v_s\hat{\kappa}_s)$ and $\hat{\lambda}_t \equiv \left(\lambda_1^t, \cdots, \lambda_s^t\right)^\top$. Finally, for each t, introduce the vector subspace:

$$\langle\hat{V}\rangle_t \equiv \{z_t \in \mathbb{R}^d : \underset{d\times 1}{z_t} = \underset{d\times s}{\hat{V}} \cdot \underset{s\times 1}{\hat{\lambda}_t}, \quad \hat{\lambda}_t \in \mathbb{R}^s\}.$$

Clearly, for each t, $\dim\langle\hat{V}\rangle_t = \text{rank}(\hat{V}) = s$—whence, the terminology.

There are three cases to consider:

(i) $d - s = k$, or $s = k^*$. The dimension of the divergent subspace is equal to the number of free variables; equivalently, the dimension of the convergent subspace is equal to the number of pre-determined variables. The system dynamics are determined in this case, because the pre-determined variables identify one (and only one) point in the convergent space, at which the free variables can jump to ensure the system remains in the convergent space, $z_0^{\text{free}} = -\Pi_u^{(1)-1}\Pi_u^{(2)} z_0^{\text{pre}}$, as illustrated in Example 3A.1 in the case where $d = 2$, $k = 1$, and x_0 is the pre-determined variable: x_0 identifies one (and only one) point in the saddlepoint path, (x_0, y_0), such that the system does not explode starting from that point.

(ii) $d - s > k$, or $s < k^*$. There are no solutions in general lying in the convergent space—a case mentioned just before Example 3A.2.

(iii) $d - s < k$, or $s > k^*$. There are infinitely many solutions that lie on the convergent space, a phenomenon typically referred to as *indeterminacy*. It is in this case that *sunspots* equilibria have the potential to arise. To illustrate, consider again Example 3A.1, in which $d = 2$, $s = 1$. In this example, indeterminacy may only arise if no pre-determined variables exist. Alternatively, in this two-dimensional example, indeterminacy would arise if $d = 2$, $s = 2$ and $k^* = 1$.

3.8 Appendix 2: Neoclassic growth in continuous-time

3.8.1 Convergence from discrete-time

3.8.1.1 Laws

Consider chopping time in the law of population growth as follows:

$$N_{hk} - N_{h(k-1)} = \bar{n} N_{h(k-1)} h, \quad k = 1, \cdots, \ell,$$

where $\bar{n}$ is an *instantaneous rate*, and $\ell = \frac{t}{h}$ is the number of subperiods in the given time period t. The solution is $N_{h\ell} = (1 + \bar{n}h)^{\ell} N_0$, or $N_t = (1 + \bar{n}h)^{t/h} N_0$. By taking limits leaves:

$$N_t = \lim_{h \downarrow 0} (1 + \bar{n}h)^{t/h} N_0 = e^{\bar{n}t} N_0.$$

We proceed similarly with the law of capital accumulation. We have:

$$K_{h(k+1)} = \left(1 - \bar{\delta}h\right) K_{hk} + I_{h(k+1)} h, \quad k = 0, \cdots, \ell - 1,$$

where $\bar{\delta}$ is an instantaneous rate, and $\ell = \frac{t}{h}$. By iterating,

$$K_t = \left(1 - \bar{\delta}h\right)^{t/h} K_0 + \sum_{j=1}^{t/h} \left(1 - \bar{\delta}h\right)^{t/h - j} I_{hj} h.$$

Taking the limits for $h \downarrow 0$ yields:

$$K_t = e^{-\bar{\delta}t} K_0 + e^{-\bar{\delta}t} \int_0^t e^{\bar{\delta}u} I_u du, \tag{3A.6}$$

or in differential form:

$$\dot{K}_t = -\bar{\delta} K_t + I_t. \tag{3A.7}$$

By replacing the IS equation,

$$Y_t = F(K_t, N_t) = C_t + I_t, \tag{3A.8}$$

into Eq. (3A.7), we obtain the law of capital accumulation:

$$\dot{K}_t = F(K_t, N_t) - C_t - \bar{\delta} K_t. \tag{3A.9}$$

3.8.1.2 Discretization issues

Consider the following exact discretization of Eq. (3A.6):

$$K_{t+1} = e^{-\bar{\delta}} K_t + e^{-\bar{\delta}(t+1)} \int_t^{t+1} e^{\bar{\delta}u} I_u du. \tag{3A.10}$$

By identifying with the capital accumulation law in discrete time, $K_{t+1} = (1 - \delta) K_t + I_t$, leaves

$$\bar{\delta} = \ln \frac{1}{1 - \delta}.$$

That is, while $\delta \in [0, 1)$, $\bar{\delta}$ can take on values on the entire real line. In the continuous time model, then, $\lim_{\delta \to 1-} \ln \frac{1}{(1-\delta)} = \infty$. In continuous time, we cannot imagine a "maximal rate of capital depreciation," as this would be infinite!

Let us replace δ into the exact discretization in Eq. (3A.10):

$$K_{t+1} = (1 - \delta) K_t \; + \; \underbrace{e^{-\bar{\delta}(t+1)} \int_t^{t+1} e^{\bar{\delta}u} I_u du}_{= I_{t+1}}.$$

That is, investments over the period from t to $t + 1$ are given by the continuous flow of investments during this period, discounted at the appropriate depreciated rates.

3.8.1.3 Per-capita dynamics

Define $k \equiv \frac{K_t}{N_t}$ and divide both sides of Eq. (3A.9) by N_t,

$$\frac{\dot{K}_t}{N_t} = \frac{F\left(K_t, N_t\right)}{N_t} - c_t - \bar{\delta} k_t = y\left(k_t\right) - c_t - \bar{\delta} k_t.$$

We have: $\dot{k} = d\left(\frac{K_t}{N_t}\right) = \frac{\dot{K}_t}{N_t} - \bar{n} k_t$, which replaced into the previous equation leaves

$$\dot{k} = y\left(k_t\right) - c_t - \left(\bar{\delta} + \bar{n}\right) k_t.$$

It is the capital accumulation contraint used to solve the program in the next section.

3.8.2 The model

Consider the following social planner problem:

$$\left\{ \begin{array}{l} \max_c \int_0^{\infty} e^{-\rho t} u\left(c_t\right) dt \\ \text{s.t. } \dot{k}_t = y(k_t) - c_t - \left(\bar{\delta} + \bar{n}\right) k_t \end{array} \right. \qquad \text{[3A.P1]}$$

where all variables are per-capita. We assume there is no capital depreciation. (Note that for the discrete time model, we assumed, instead, a total capital depreciation.) The Hamiltonian is,

$$H_t = u(c_t) + \lambda_t \left[y(k_t) - c_t - \left(\bar{\delta} + \bar{n}\right) k_t \right],$$

where λ is a co-state variable. As explained below (see Appendix 3), the first-order conditions for this problem are:

$$\left\{ \begin{array}{lcl} \dfrac{\partial H_t}{\partial c_t} & = & 0 \\ \dfrac{\partial H_t}{\partial \lambda_t} & = & \dot{k}_t \\ -\dfrac{\partial H_t}{\partial k_t} & = & \dot{\lambda}_t - \rho \lambda_t \end{array} \right.$$

The first and the third conditions lead to

$$\left\{ \begin{array}{lcl} \lambda_t & = & u'\left(c_t\right) \\ \dot{\lambda}_t & = & \left[\rho + \bar{\delta} + \bar{n} - y'\left(k_t\right)\right] \lambda_t \end{array} \right. \qquad (3A.11)$$

By differentiating the first equation in (3A.11) with respect to time yields

$$\dot{\lambda}_t = \left(\frac{u''(c_t)}{u'(c_t)} \dot{c}_t \right) \lambda_t,$$

and using the third equation in (3A.11),

$$\dot{c}_t = \frac{u'\left(c_t\right)}{u''\left(c_t\right)} \left(\rho + \bar{\delta} + \bar{n} - y'\left(k_t\right)\right). \qquad (3A.12)$$

The equilibrium is the solution of the system consisting of the constraint of the program [3A.P1], and Eq. (3A.12). Similarly as in Section 3.3.3, we analyze the dynamics of the system in a small neighborhood of the stationary state, defined as the solution (c, k) of the constraint of the program [3A.P1], and Eq. (3A.12), when $\dot{c}(t) = \dot{k}(t) = 0$,

$$\left\{ \begin{array}{lcl} c & = & y(k) - \left(\bar{\delta} + \bar{n}\right) k \\ \rho + \bar{\delta} + \bar{n} & = & y'(k) \end{array} \right.$$

A first-order approximation of both sides of the constraint of the program [3A.P1], and Eq. (3A.12), near (c,k), yields:

$$\left\{\begin{array}{rcl}\dot{c}_t &=& -\dfrac{u'(c)}{u''(c)}y''(k)\,(k_t-k)\\ \dot{k}_t &=& \rho\cdot(k_t-k)-(c_t-c)\end{array}\right.$$

where we used the equality $\rho+\bar{\delta}+\bar{n}=y'(k)$. By setting $x_t\equiv c_t-c$ and $y_t\equiv k_t-k$, the previous system can be rewritten as:

$$\dot{z}_t=A\cdot z_t,\quad A\equiv\left(\begin{array}{cc}0 & -\dfrac{u'(c)}{u''(c)}y''(k)\\ -1 & \rho\end{array}\right),$$

where $z\equiv(x,y)^\top$. **Warning! There must be some mistake somewhere.**

We diagonalize this system by setting $A=P\Lambda P^{-1}$, where P and Λ are as in Appendix 1. We have:

$$\dot{\nu}_t=\Lambda\cdot\nu_t,$$

where $\nu\equiv P^{-1}z$. The eigenvalues are solutions of the following quadratic equation:

$$0=\lambda^2-\rho\lambda-\frac{u'(c)}{u''(c)}y''(k).$$

We see that $\lambda_1<0<\lambda_2$, and $\lambda_1\equiv\frac{\rho}{2}-\frac{1}{2}\sqrt{\rho^2+4\frac{u'(c)}{u''(c)}y''(k)}$. The solution for ν_t is:

$$\nu_{it}=\kappa_i e^{\lambda_i t},\ i=1,2,$$

whence

$$z_t=P\cdot\nu_t=v_1\kappa_1e^{\lambda_1t}+v_2\kappa_2e^{\lambda_2t},$$

where the v_is are 2×1 vectors. We have,

$$\left\{\begin{array}{rcl}x_t &=& v_{11}\kappa_1e^{\lambda_1t}+v_{12}\kappa_2e^{\lambda_2t}\\ y_t &=& v_{21}\kappa_1e^{\lambda_1t}+v_{22}\kappa_2e^{\lambda_2t}\end{array}\right.$$

Let us evaluate this solution in $t=0$,

$$\left(\begin{array}{c}x_0\\ y_0\end{array}\right)=P\kappa=\left(\begin{array}{cc}v_1 & v_2\end{array}\right)\left(\begin{array}{c}\kappa_1\\ \kappa_2\end{array}\right).$$

By repeating the reasoning of the previous appendix,

$$\kappa_2=0\Longleftrightarrow\frac{y_0}{x_0}=\frac{v_{21}}{v_{11}}.$$

As in the discrete time model, the saddlepoint path is located along a line that has as a slope the ratio of the components of the eigenvector associated with the negative root. We can explicitely compute such ratio. By definition, $A\cdot v_1=\lambda_1v_1\Leftrightarrow$

$$\left\{\begin{array}{rcl}-\dfrac{u'(c)}{u''(c)}y''(k) &=& \lambda_1v_{11}\\ -v_{11}+\rho v_{21} &=& \lambda_1v_{21}\end{array}\right.$$

i.e., $\frac{v_{21}}{v_{11}}=-\frac{\lambda_1}{\frac{u'(c)}{u''(c)}y''(k)}$ and simultaneously, $\frac{v_{21}}{v_{11}}=\frac{1}{\rho-\lambda_1}$.

3.9 Appendix 3: Notes on optimization of continuous time systems

Consider the following optimization problem. For k_0 given,

$$\begin{cases} J(k_t,t,T) \equiv \max\limits_{(v_\tau)_{\tau\in[0,T]}} E_t\left(\int_t^T e^{-\rho(\tau-t)}u(k_\tau,v_\tau)\,d\tau + e^{-\rho(T-t)}B(k_T)\right) \\ \text{s.t. } dk_\tau = T(k_\tau,v_\tau)\,d\tau + \sigma(k_\tau,v_\tau)\,dW_\tau \end{cases} \tag{3A.13}$$

where u and B are an instantaneous and a terminal payoff, ρ is a subjective instantaneous rate, W_τ is a standard Brownian Motion and, finally, $T(k,v)$ and $\sigma(k,v)$ are given drift and diffusions functions.

We interpret k as a state variable and v as a control. In general, the control depends on the realization of k although we assume that it cannot depend on future observations of k—it can only depend on past values of k. We confine attention to controls known as *feedbacks*, i.e., such that v only depends on the current value of k, $v_\tau(\omega) \equiv \hat{v}(\tau, k_\tau(\omega))$, for each sample path ω of the state variable $k_\tau(\omega)$. The function $\hat{v}$ makes the control and, then, the state variable k, Markovian.

We refer the value function $J(k_t,t,T)$ to as the *undiscounted* indirect utility and, based on famous derivations of Merton (1969, p. 24) in related contexts, we can define

$$I(k_t,t,T) \equiv e^{-\rho t}J(k_t,t,T), \tag{3A.14}$$

and call $I(\cdot)$ *discounted* indirect utility. We shall come back to this discounted utility below (see Eq. (3A.16) below).

We apply what is known as the *stochastic programming principle*. Heuristically, we maximize up to an intermediate point in time, $t+\Delta t$, say, assuming the maximization for the remaining time period $[t+\Delta t, T]$ holds. We have:

$$\begin{aligned} J(k_t,t,T) &= \max_{(v_\tau)_{\tau\in[0,T]}} E_t\left(\int_t^T e^{-\rho(\tau-t)}u(k_\tau,v_\tau)\,d\tau + e^{-\rho(T-t)}B(k_T)\right) \\ &= \max_{(v_\tau)_{\tau\in[0,T]}} \left[E_t\left(\int_t^{t+\Delta t} e^{-\rho(\tau-t)}u(k_\tau,v_\tau)\,d\tau\right)\right. \\ &\qquad \left. + e^{-\rho\Delta t}E_t\left(E_{t+\Delta t}\left(\int_{t+\Delta t}^T e^{-\rho(\tau-t-\Delta t)}u(k_\tau,v_\tau)\,d\tau + e^{-\rho(T-t-\Delta t)}B(k_T)\right)\right)\right] \\ &= \max_{(v_\tau)_{\tau\in[0,\Delta T]}} E_t\left(\int_t^{t+\Delta t} e^{-\rho(\tau-t)}u(k_\tau,v_\tau)\,d\tau\right) + e^{-\rho\Delta t}E_t\left(J(k_{t+\Delta t},t+\Delta t,T)\right), \end{aligned}$$

where the last two lines follow by the law of iterated expectations. By rearranging terms,

$$\begin{aligned} &\left(\frac{1-e^{-\rho\Delta t}}{\Delta t}\right)J(k_t,t,T) \\ &= \max_{(v_\tau)_{\tau\in[0,\Delta T]}} E_t\left(\frac{1}{\Delta t}\int_t^{t+\Delta t} e^{-\rho(\tau-t)}u(k_\tau,v_\tau)\,d\tau\right) + e^{-\rho\Delta t}E_t\left(\frac{1}{\Delta t}\left(J(k_{t+\Delta t},t+\Delta t,T) - J(k_t,t,T)\right)\right). \end{aligned}$$

For small Δt, and under regularity conditions, we have that by Itô's lemma,

$$0 = \max_v \left(u(k,v) + \left(\frac{\partial}{\partial\tau} + L\right)J(k,\tau,T) - \rho J(k,\tau,T)\right), \tag{3A.15}$$

subject to the boundary condition $J(k,T,T) = B(k)$, where $LJ \equiv \frac{\partial J}{\partial k}T + \frac{1}{2}\frac{\partial^2 J}{\partial k^2}\sigma^2$.

Eq. (3A.15) is at the basis of the following derivations. However, these derivations could equivalently be based on the discounted utility $I(\cdot)$ in (3A.14). Note that

$$\left(\frac{\partial}{\partial\tau} + L\right)I(k,\tau,T) = e^{-\rho t}\left(\frac{\partial}{\partial\tau} + L - \rho\right)J(k,\tau,T),$$

and replacing into Eq. (3A.15) leaves

$$0=\max_{v}\left(e^{-\rho t}u\left(k,v\right)+\left(\frac{\partial}{\partial\tau}+L\right)I\left(k,\tau,T\right)\right). \tag{3A.16}$$

However, we shall keep on relying on Eq. (3A.15) rather than (3A.16).

The first order condition for Eq. (3A.15) is

$$0=u_{v}\left(k,\hat{v}\right)+\frac{\partial J\left(k,\tau,T\right)}{\partial k}T_{v}\left(k,\hat{v}\right)+\frac{1}{2}\frac{\partial^{2}J\left(k,\tau,T\right)}{\partial k^{2}}\frac{\partial\sigma^{2}\left(k,\hat{v}\right)}{\partial v},$$

which implicitly defines $\hat{v}\equiv\hat{v}\left(\tau,k\right)$, such that Eq. (3A.15) can be written as:

$$\begin{aligned}\rho J\left(k,\tau,T\right)-\frac{\partial J\left(k,\tau,T\right)}{\partial\tau} &= u\left(k,\hat{v}\left(\tau,k\right)\right)+\lambda\left(k,\tau,T\right)T\left(k,\hat{v}\left(\tau,k\right)\right)+\frac{1}{2}\mu\left(k,\tau,T\right)\sigma^{2}\left(k,\hat{v}\left(\tau,k\right)\right)\\ &\equiv\hat{H}\left(\tau,k,\lambda\left(k,\tau,T\right),\mu\left(k,\tau,T\right)\right),\end{aligned} \tag{3A.17}$$

where

$$\lambda\left(k,\tau,T\right)\equiv\frac{\partial J\left(k,\tau,T\right)}{\partial k},$$

is interpreted as the marginal indirect utility in (3A.13) with respect to k, and

$$\mu\left(k,\tau,T\right)\equiv\frac{\partial\lambda\left(k,\tau,T\right)}{\partial k}=\frac{\partial^{2}J\left(k,\tau,T\right)}{\partial k^{2}}.$$

The function $\hat{H}$ is usually referred to as the optimized *Hamiltonian*, and Eq. (3A.17) is the *Bellman Equation* for diffusion processes.

By Itô's lemma:

$$d\lambda=\left(\frac{\partial\lambda}{\partial\tau}+\frac{\partial\lambda}{\partial k}T+\frac{1}{2}\frac{\partial^{2}\lambda}{\partial k^{2}}\sigma^{2}\right)d\tau+\frac{\partial\lambda}{\partial k}\sigma dW. \tag{3A.18}$$

Moreover, by differentiating both sides of the Bellman Equation (3A.17) with respect to k,

$$\begin{aligned}\rho\frac{\partial J}{\partial k}-\frac{\partial^{2}J}{\partial k\partial\tau} &= \rho\lambda-\frac{\partial\lambda}{\partial\tau}\\ &= \frac{\partial\hat{H}}{\partial k}+\frac{\partial\hat{H}}{\partial\lambda}\frac{\partial\lambda}{\partial k}+\frac{\partial\hat{H}}{\partial\mu}\frac{\partial\mu}{\partial k}\\ &= \frac{\partial\hat{H}}{\partial k}+T\frac{\partial\lambda}{\partial k}+\frac{1}{2}\sigma^{2}\frac{\partial\mu}{\partial k},\end{aligned} \tag{3A.19}$$

where the first equality follows by the definition of λ, and the third equality holds by the definition of the optimized Hamiltonian function, $\hat{H}\left(\tau,k,\lambda,\mu\right)$. Plugging Eq. (3A.19) into Eq. (3A.18) leaves:

$$d\lambda=\left(\rho\lambda-\frac{\partial\hat{H}}{\partial k}\right)d\tau+\frac{\partial\lambda}{\partial k}\sigma dW,$$

which shows that:

$$\rho\lambda-\frac{\partial\hat{H}}{\partial k}=\frac{E\left(d\lambda\right)}{d\tau}=\frac{\partial\lambda}{\partial\tau}+\frac{\partial\lambda}{\partial k}T+\frac{1}{2}\frac{\partial^{2}\lambda}{\partial k^{2}}\sigma^{2}. \tag{3A.20}$$

To summarize, we solve the problem in three steps:

1. Maximize the Hamiltonian function,

$$H\left(\tau,k,\lambda,\mu\right)\equiv u(k,v)+\lambda\cdot T\left(k,v\right)+\frac{1}{2}\cdot\mu\cdot\sigma\left(k,v\right)^{2},$$

with respect to the control v. The result is the optimized Hamiltonian function, $\hat{H}\left(\tau,k,\lambda,\mu\right)$.

2. Impose the condition in Eq. (3A.20).

3. Solve the partial differential equation (3A.15) for J.

Succinctly,

$$\begin{cases} \dfrac{\partial \hat{H}}{\partial v} = 0 \\ \dfrac{\partial \hat{H}}{\partial \lambda} = T \\ \rho\lambda - \dfrac{\partial \hat{H}}{\partial k} = \dfrac{\partial \lambda}{\partial \tau} + \dfrac{\partial \lambda}{\partial k} T + \dfrac{1}{2}\dfrac{\partial^2 \lambda}{\partial k^2}\sigma^2 \end{cases} \tag{3A.21}$$

Note that in the infinite horizon case (i.e., the problem in (3A.13) when $T = \infty$), Eq. (3A.15) collapses to

$$0 = \max_{v} \left(u\left(k, v\right) + LJ\left(k\right) - \rho J\left(k\right) \right),$$

subject to the transversality condition,

$$\lim_{T \to \infty} e^{-\rho(T-t)} E_t \left[J\left(k_T\right) \right] = 0.$$

The deterministic model in Appendix 2 of this chapter is a special case of the current setup. Indeed, the first of Eqs. (3A.11) reveals that λ is not a function of k such that Eqs. (3A.11) are a special case of Eqs. (3A.21), namely for $\sigma \equiv 0$.

References

Abel, A.B., N.G. Mankiw, L.H. Summers and R.J. Zeckhauser (1989): "Assessing Dynamic Efficiency: Theory and Evidence." *Review of Economic Studies* 56, 1-20.

Cochrane, J. H., F. A. Longstaff, and P. Santa-Clara (2008): "Two Trees." *Review of Financial Studies* 21, 347-385.

Farmer, R. (1998): *The Macroeconomics of Self-Fulfilling Prophecies*. Boston: MIT Press.

Hayashi, F. (1982): "Tobin's Marginal q and Average q: A Neoclassical Interpretation." *Econometrica* 50, 213-224.

Kamihigashi, T. (1996): "Real Business Cycles and Sunspot Fluctuations are Observationally Equivalent." *Journal of Monetary Economics* 37, 105-117.

King, R. G. and S. T. Rebelo (1999): "Resuscitating Real Business Cycles." In: J. B. Taylor and M. Woodford (Editors): *Handbook of Macroeconomics*, Elsevier.

Lucas, R. E. (1972): "Expectations and the Neutrality of Money." *Journal of Economic Theory* 4, 103-124.

Lucas, R. E. (1978): "Asset Prices in an Exchange Economy." *Econometrica* 46, 1429-1445.

Lucas, R. E. (1994): "Money and Macroeconomics." In: *General Equilibrium 40th Anniversary Conference*, CORE DP no. 9482, 184-187.

Martin, I. (2011): "The Lucas Orchard." Working Paper Stanford University.

Menzly, L., Santos, T., and P. Veronesi (2004): "Understanding Predictability." *Journal of Political Economy* 112, 1-47.

Merton, R.C. (1969): "Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case." *Review of Economics and Statistics* 51, 247-257.

Pavlova, A. and R. Rigobon (2008): "The Role of Portfolio Constraints in the International Propagation of Shocks." *Review of Economic Studies* 75, 1215-1256.

Prescott, E. (1991): "Real Business Cycle Theory: What Have We Learned?" *Revista de Analisis Economico* 6, 3-19.

Stokey, N. L. and R. E. Lucas, (with E.C. Prescott) (1989): *Recursive Methods in Economic Dynamics*. Harvard University Press.

Tirole, J. (1988): "Efficacité intertemporelle, transferts intergénérationnels et formation du prix des actifs: une introduction." *Melanges économiques. Essais en l'honneur de Edmond Malinvaud*. Paris: Editions Economica & Editions EHESS, 157-185.

Tobin, J. (1969): "A General Equilibrium Approach to Monetary Policy." *Journal of Money, Credit and Banking* 1, 15-29.

Watson, M. (1993): "Measures of Fit for Calibrated Models." *Journal of Political Economy* 101, 1011-1041.

4
Information, security design and financial contracting

4.1 Introduction

Much of the discussion in the previous chapters relies on explanations where symmetrically informed agents invest in existing assets. For example, Chapter 3 contains generalizations of the static CAPM introduced in Chapter 1, in which asset markets are cleared in general equilibrium economies. Moreover, in the next chapter and in Part II (see, e.g., Chapter 9), we shall explain asset prices when markets are free from arbitrage and any frictions while relying on dynamic stochastic general equilibrium models.

However, market imperfections are important, both in theory and in practice. Information problems amount to one of the most notable imperfections, and they were incorporated in many areas of financial economics, extending the predictions arising from frictionless dynamic stochastic general equilibrium models. This chapter provides an introduction to these problems. Part II deals with many additional instances and articulated models of financial markets with frictions, including non-informational frictions.

[In progress]

Section 4.2 discusses three models that illustrate the importance of information problems in financial economics, and the way these problems challenge the main architecture underlying theoretical constructs such as the CAPM. It contains a succinct introduction to problems of adverse selection, moral hazard and signaling in financial markets, as well as the main ideas underlying the contracts that mitigate these problems. The remainder of the chapter describes the main elements that make these contracts endogeneous (i.e., optimal) between parties. Section 4.3 deals with idealized debt contracts, arising in absence of asymmetric information.

[In progress]

4.2 Theoretical challenges to frictionless markets: information problems

Why do certain assets exist? What makes corporations issue and, then, investors agree to trade equity or debt? And why would investors agree to trade even more complex securities? This chapter reviews theories of debt and the firm's capital structure that take into account informational asymmetries between investors and firms' managers. This section illustrates the

motivation underlying these theories while relying on three simple models. The next section provides the reader with historical background regarding the economics of informations. The three models are developed in Sections 4.2.2 through 4.2.4.

4.2.1 The economics of information

Economic theorists provided the historical background underlying information problems. Some agents may have more information than others. For example, insurance companies cannot verify the efforts made by insured individuals to avoid risks, such that markets may fail to arise in the first place, a situation of "moral hazard" explained by Arrow (1963) in a famous contribution; or, to remain in the theme of these lectures, an investor (the "principal") who contemplates investing into mortgage-backed securities cannot verify whether a bank (the "agent") is securitizing loans originated with good due diligence. This conflict of interest is an instance of what is referred to as an agency problem. Parallel to moral hazard is "adverse selection": to illustrate, a trader would only propose to sell at £5 if the assets he proposes are worth less than £5. Finally, how do companies "signal" their value or, how can they transmit credible information to the marketplace? Finally, how to incentivize borrowers or workers to adopt behaviors in line with the interest of their "principals"?

Questions of this kind have been discussed since at least Adam Smith (1776) and, then, for centuries, by many other economists, as explained in detail by Laffont and Martimort (2002). However, it was only during the 1970s that new models appeared to formalize these problems. Economists were stumbling into conceptual difficulties while attempting to incorporate such information asymmetries into elegant general equilibrium models. Their focus then became the analysis of simple market relationships amongst a few agents, covering adverse selection (Akerlof, 1970; Rothschild and Stiglitz, 1976), moral hazard (Ross, 1973, Mirrlees, 1974, 1975; Holmström, 1979; Shavell, 1979), and signaling (Spence, 1973, 1974; and Riley, 1975).

The earliest applications of information theory to finance were in the area of signaling (Leland and Pyle, 1977; Ross, 1977; Bhattacharya, 1979), followed by those in the area of adverse selection, aiming to deal with debt relationships (Townsend, 1979; Gale and Hellwig, 1985), credit markets (Stiglitz and Weiss, 1981) and asset markets (Glosten and Milgrom, 1985; Kyle, 1985), and, then, those aiming to rationalize debt as a solution to moral hazard problems (Innes, 1990).

We review these information-based theories of financial markets in more advanced sections (from Section 4.4) as well as in Part II (e.g., in Chapter 10), but this section illustrates some of their flavor. Thus, so far, securities have always been given exogenously. But in practice, securities may or may not exist or, may or may not trade, according to market forces that we have not analyzed so far. For example, how can we even consider purchasing any security at a price proposed to us? It may well be the case that our counterpart has information that the security will not likely perform and that the price we are offered is too high. It is an *adverse selection* problem: the only securities being proposed for sale are those with fundamental values below their prices.

How can two parties trade in these conditions? They actually could. Our reluctance to buy may weaken if we knew that a liquidity shock (not information) is the motivation underlying our counterparty's decision to sell. Note, however, that in this case, a transaction price (if any) should reflect the buyer's concern of trading with an informed counterparty; therefore, a liquidity motivated trade can "move" prices. These price impacts are actually important. Now, we have that imperfectly liquid markets are somehow inefficient (prices may be affected

by reasons unrelated to the assets' fundamentals) and yet a transaction may well take place anyway; that is, inefficiencies may actually make market function.

How inefficient markets can be? Depending on the benefits of acquiring information relative to their costs, there may well be investors deciding to remain uninformed, as they would simply learn some information held by informed traders through the equilibrium price. It is the celebrated Grossman-Stiglitz "equilibrium degree of disequilibrium" (Grossman and Stiglitz, 1980). This chapter does not deal with these free-riding problems and, in general, with how information acquisition is determined in equilibrium; these issues are studied in Part II (see Chapter 10). Rather, it makes the preliminary point of illustrating how, and somehow paradoxically, the presence of liquidity motivated trades can help trading arise in a market plagued with asymmetric information.

A second information problem regards the very securities creation. Certain derivative assets draw value from events with probability of occurrence that may be affected by the security producers. For example, financial intermediaries can create and sell securities that are backed by mortgage loans. The resulting mortgages become "securitized" and possibly more liquid and easy to trade than the original loans. Chapter 14 provides some details of this securitization process. Securitization is an useful business model. On the one hand, outside investors may want to get some exposure to the original loans, but originating loans requires specialization. On the other hand, the originators of loans only desire a partial exposure to these loans. Securitization would allow to share these risks.

The delicate point of securitization is that financial intermediaries need to be properly incentivized while originating the loans: anticipating that portions of these loans will be distributed to outside investors, financial intermediaries may well extend loans with poor credit standards (pretending these standards are high), which are arguably less expensive to originate than loans with high credit standards. It is a *moral hazard* problem, to be tackled while investors provide originators with adequate inventives. The incentive to originate good loans is to make sure originators may become involved in losses if these losses materialized. It is a solution consistent with standard remedies to moral hazard suggested by the economics of information. To illustrate, in insurance, a car driver is not incentivized to drive properly when he is offered a full insurance contract. In our context, the originator needs to retain a sufficiently high proportion of securitized assets (a "skin in the game") to be incentivized in generating high quality mortgage loans.

A third interesting instance of information problem regards the case in which a firm actually has a very good project, but it is not able to credibly explain the market the project is good. Thus, the firm wants to issue bonds to finance this project but investors walk away from these bonds because the firm has not credibly explained its bonds have a good quality. In the previous moral hazard example, a loan originator needs retain a sufficiently high proportion of the securitized asset. In the context of this new example, the firm needs to signal the market that it is a good type, and the signal has to be costly. For example, the firm may issue an option embedded into its debt, which gives the right but not the obligation to redeem its debt at some pre-specified point; naturally, the option is expensive for the firm. The rationale behind this "callable bond" is that the firm wants the investors to believe that the probability of default is really small and that this will be known at the time of the option exercise, when the cost of capital will be lower exactly for this reason. It is a *signaling problem*. Note, now, and interestingly, that a market may arise (a bond market) due to the presence of a derivative (the callable option embedded into it).

The point of these models is that many attributes of securities studied so far are endogenous. For example, the riskiness of a security in the moral hazard example may well depend on forces that the standard CAPM does not take into account. Alternatively, a market may even fail to arise in all the examples of this section. It is interesting: the CAPM was developed before the birth of the economics of information. Frictions and market failures would become the focus of economists only during the 1970s, as discussed. We now present three models that illustrate these ideas in each of the previous three areas regarding information frictions: adverse selection, moral hazard, and signaling.

4.2.2 Adverse selection

4.2.2.1 The lemons problem

Akerlof (1970) points out that asymmetric information raises a series of problems that may well lead to dysfunctional markets. Consider a market for a continuum of cars with quality q uniformly distributed in $(0, 1)$. First, assume that for any specific car drawn from this distribution, the quality q is common knowledge. Potential buyers have a reservation value for the cars equal to Kq, where $K > 1$ is a preference parameter; that is, buyers would be willing to buy at any price $s \leq Kq$. Instead, sellers would sell at any price $s > q$. Therefore, a transaction will take place for any price $\hat{s} \in (q, Kq)$.

Next, suppose product quality uncertainty: the cars' quality q is *not* common knowledge: only sellers know the true quality. In this case, an equilibrium may fail to emerge. Indeed, suppose s_o is an equilibrium price, which can only be acceptable by sellers when the true quality satisfies $q \leq s_o$. That is, the only sellers who sell are those proposing cars that have a quality worse than s_o: it is the adverse selection problem. The buyers anticipate this adverse selection and update their beliefs on the cars' quality to $E(q|\, s_o)$, where the (uniform) density of the cars' quality conditional on s_o is $\frac{1}{s_o}\mathbb{I}_{q\in(0,s_o)}$; thus, $E(q|\, s_o) = \frac{1}{2}s_o$. Buyers will then be willing to pay no more than $\frac{1}{2}Ks_o$, such that a transaction will occur (if any) only if $s_o < \frac{1}{2}Ks_o$: there are no equilibria unless the buyers have a sufficiently strong preference for cars, i.e., $K > 2$. We now explain how the lemons problem can be mitigated when there is a probability that the motives for selling do not related to quality problems.

4.2.2.2 Trading with noise

The lemons problem suggests that it is not trivial to explain how parties agree to trade financial securities: how do we know that those who want to sell a security do not already know they are selling us lemons? Investor protection laws certainly mitigate these concerns; however, regulation may only serve as a deterrent to episodes of misselling and, in many instances, it only applies after a trade is executed. Furthermore, trading on the basis of asymmetric information may be far from being a crime: some institutional investors may have more information than others simply because they invest in (and pay for) research. Once again, how do we justify trading between parties that have different information quality? It is a very important and beautiful chapter in the history of thought, which these lectures deal with in detail in Chapter 10. We now illustrate how trade may arise in spite of asymmetric information, relying on a famous mechanism: the presence of trades possibly motivated by liquidity needs.

Suppose, then, that there is a fraction of cars equal to $1 - p$ that is being sold because the owners need cash, and assume these cars' owners are ready to sell for any price. With complementary probability p, the car is still a lemon. Therefore, conditional on only observing

a price equal to s_o, the expected quality is

$$E\left(q|\, s_o\right) = pE_l\left(q|\, s_o\right) + (1-p)\, E_{nl}\left(q|\, s_o\right) = p\frac{1}{2}s_0 + (1-p)\,\frac{1}{2}. \tag{4.1}$$

where $E_l\left(q|\, s_o\right)$ and $E_{nl}\left(q|\, s_o\right)$ denote expectations conditional on s_0 and the car being a lemon and not a lemon, respectively. Note that the second quality follows because $E_{nl}\left(q|\, s_o\right) = E\left(q\right) = \frac{1}{2}$: by assumption, the car owners accept to sell at any price when hit by a liquidity shock, such that s_o carries no information content on q.

Trading occurs whenever the equilibrium price s_o satisfies

$$s_0 \leq KE\left(q|\, s_o\right), \tag{4.2}$$

or for any equilibrium price $s_o \in (0, \hat{s}_o]$, where $\hat{s}_o = \frac{(1-p)K}{2-pK}$. The range of trading prices obviously increases with K. For example, when $K = 2$, $\hat{s}_o = 1$ and trading always takes place, as in the lemons markets with $p = 1$. But, and in contrast with the previous lemons market, trading can take place even when consumers have weaker preferences. The solid line of Figure 4.1 depicts the conditional expectation in Eq. (4.1), which is the R.H.S. of (4.2) when $K = 1$ and $p = \frac{1}{2}$. Note that when $K = 1$, there are equilibrium prices that were not in the previous lemons market with $p = 1$, arising for all $s_o \in (0, \hat{s}_o]$, where $\hat{s}_o = \frac{1-p}{2-p}$.

In these examples, noise is what makes markets plagued with asymmetric information work. The simple CAPM does not explain these phenomena, although the microstructure of markets and information problems are important aspects of market behavior, which will be analyzed in greater detail in Chapter 10 of Part II.

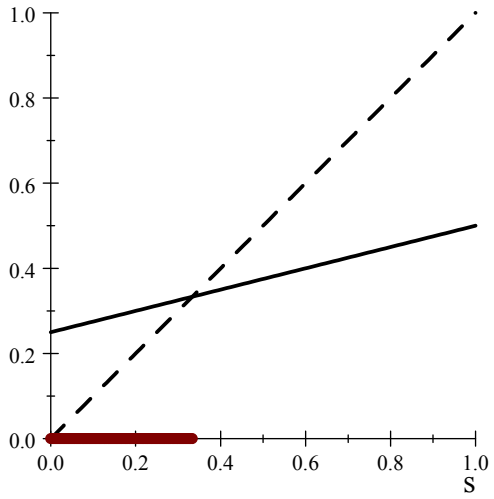


FIGURE 4.1. Equilibrium prices in a model of lemons with noise. The solid line depicts the conditional expectation $E\left(q|\, s\right)$ in Eq. (4.1) when $p = \frac{1}{2}$. Trading takes place for all values of $s : s \leq KE\left(q|\, s\right)$, i.e., assuming $K = 1$, for all $s \in (0, \hat{s}]$, where $\hat{s} = \frac{1}{3}$.

4.2.3 Moral hazard

We study securities markets arising within a "originate-and-distribute" business model: a number of banks originate loans, which they can securitize to a number of sophisticated investors.

Problems of this kind are extensively studied in frameworks of optimal contracting in the presence of asymmetric information.[1] This section discusses a streamlined model that illustrates the general skin-in-the-game ideas, without entering into details of optimal financial contracting, discussed in the next sections of this chapter.

We assume that investors are sophisticated, in that they are able to obtain expected returns equal to ρ, the "alternative investments" return. Moreover, banks and investors act competitively and are specialists in their business: only investors have a reservation expected return equal to ρ; and only banks can manufacture securities out of the original loans. Chapter 14 provides a succinct description of the market for mortgage-backed-securities, which are a very famous instance of these mechanisms illustrated in this section.

We consider a single, representative bank, which extends borrowers a loan of size equal to N that is subject to a loss. The loan rate is implied by the amount of resources made available to borrowers, kN: it is $1-k$, and is determined endogenously (see, e.g., (4.11) and Figure 4.3 below). Loss is denoted with $\tilde{l}N$, where $\tilde{l} \in [0,1]$ is a random variable described in detail below. Therefore, if the bank distributes the entire loan to investors, the expected return on this loan required by the investors satisfies $\frac{1-E(\tilde{l})}{s} \geq 1+\rho$, where $s \equiv \frac{S}{N}$ and S is the price of the loan paid by the investors. The inequality reflects the assumption that the investors would walk away from this deal should the price be not enough to guarantee an expected return at least as large as ρ. By the assumption of investors' competition,

$$s = \frac{1-E(\tilde{l})}{1+\rho}. \tag{4.3}$$

The bank is adverse to the risk of the the loan it has originated. To illustrate, the bank may be subject to Value-at-Risk (VaR) constraints,[2] which make it averse to the loan variance. The bank can mitigate this risk while securitizing the loan: it can decide to sell a fraction χ of the loan it originated for a value equal to χs. We study the incentives the bank has to securitize these loans in a market with symmetric information and, then, how the bank's incentives can be aligned to those of investors in markets plagued by moral hazard.

4.2.3.1 Full insurance, full securitization

Securitization bears resemblance with insurance: the bank that originates the loan may wish to re-sell the loan, thereby insulating itself from the risk of default. Similar to standard conclusions in the insurance literature, and under conditions identified below, the bank finds it optimal to securitize the entire loan. This full insurance case can be understood through a very simple model with two states of nature, such that $\tilde{l} = l < 1$ with probability p and $\tilde{l} = 0$ with complementary probability. Therefore, in the bad state, the payoff to the bank is w_b and, in the bad, the payoff is w_g, where

$$\begin{cases} \dfrac{w_b}{N} = (1-\chi)(1-l) + \chi s - k \\ \dfrac{w_g}{N} = (1-\chi) + \chi s - k \end{cases} \tag{4.4}$$

[1] For example, DeMarzo (2005) considers security designs arising through signaling: a bank retaining a portion of these loans would signal a good loan quality to investors. In Hartman-Glaser, Piskorski and Tchistyi (2012), to mention another contribution, securitization is framed within a dynamic contracting model with moral hazard.

[2] A VaR constraint imposes the bank to limit exposure to risk. Chapter 14 provides an overview regarding VaR in the credit space and existing regulation.

and kN is the original amount of money lent to the borrower. The bank payoffs are equal to the value of the securitized assets sold to the investors net of the resources originally lent to the borrowers, $\chi S - kN$, plus asset performance regarding the retained proportion of the original loan, $(1-\chi)(N-lN)$.

Eqs. (4.4) clearly show that securitization allows the bank to smooth loan risks. Randomness is entirely eliminated with full securitization, $\chi = 100\%$. If $\rho = 0$, it is indeed in the bank's interest to proceed with full securitization. Let us model bank's aversion to risk while assuming the bank maximizes the following objective

$$\max_{\chi\in[0,1]} pu(w_b) + (1-p)\,u(w_g),$$

where w_b and w_g are as in Eqs. (4.4) and u is an increasing and concave function. The first order conditions for this program lead to

$$\frac{u'(w_b)}{u'(w_g)} = \frac{1-p}{p}\frac{\rho + lp}{l(1-p) - \rho(1-l)},$$

where we have used Eq. (4.3) to substitute for the equilibrium price, and used $E(\tilde{l}) = lp$.

We assume that ρ is small enough to make the ratio of the marginal utilities across states strictly positive and bounded. In the special case in which $\rho = 0$, we have that $u'(w_b) = u'(w_g)$, that is, $w_b = w_g$, which can only be achieved through $\chi = 1$—full securitization. This conclusion is reminiscent of standard predictions in the insurance literature, for which a consumer always wishes to achieve full insurance in markets where insurers face perfect competition. In our case, the restriction that $\rho = 0$ corresponds to the insurance market situation in which risks are priced through actuarial methods. We now proceed with a very tractable model that allows us to examine details of this securitization market in the presence of asymmetric information. We first provide a framework in the symmetric information case.

4.2.3.2 Symmetric information again

Let us generalize the two constrains in (4.4), assuming that losses are a continuous random variable, such that the final payoff to the bank is

$$\frac{\tilde{w}}{N} = (1-\chi)(1-\tilde{l}) + \chi s - k, \tag{4.5}$$

where $\tilde{l}$ is uniformly distributed on $(0, X)$ and $X < 1$; thus $E_X(\tilde{l}) = \frac{1}{2}X$ and $var_X(\tilde{l}) = \frac{1}{12}X^2$, where we emphasize the dependence on X for reasons developed below. We assume that the bank has a mean-variance objective, thereby choosing a level of securitization equal to

$$\chi^* = \arg\max_{\chi\in[0,1]} E_X\left(\frac{\tilde{w}}{N}\right) - \frac{\tau}{2} var_X\left(\frac{\tilde{w}}{N}\right), \quad \tau > 0.$$

The trade-off of this optimization problem is the following. On the one hand, by increasing the extent of securitization, χ, the expected payoff increases due to the higher securitization proceeds, χs in Eq. (4.5). On the other hand, a higher χ implies a lower recovery value, $(1-\chi)(1-\tilde{l})$. On average, the net benefit of an increase in securitization is obviously negative, because the price investors pay, s, is lower than the recovery value

$$E_X(s - (1-\tilde{l})) = -\frac{\rho}{1+\rho}(1 - E_X(\tilde{l})).$$

But to this expected negative marginal gain corresponds a positive marginal benefit due to a variance reduction, $\tau(1-\chi)\,var_X(\tilde{l})$. By equating the expected marginal gain to this marginal benefit, and rearranging leaves the optimal retention rate in the securitization deal

$$1-\chi^* = \frac{1-E_X(\tilde{l})}{\tau var_X(\tilde{l})}\frac{\rho}{1+\rho}. \tag{4.6}$$

We assume parameter values are such that the solution is interior. Note the full insurance result: if $\rho = 0$, the optimal securitization rate is $\chi^* = 100\%$. In general, the retention rate $1-\chi^*$ is increasing with ρ: the higher the expected return required by the sophisticated investors, the lower the securitization revenues (i.e., χs) and, then, the incentives to securitize. We now proceed to analyze how this conclusion is modified when the bank and the investors are asymmetrically informed regarding the quality of loans.

4.2.3.3 Asymmetric information: incentive-compatible constraint

Credit standards have been assumed to be given so far. However, the borrowers' quality may be a decision variable: while anticipating that a significant proportion of the loans will be securitized (and, therefore, sold to outside investors) at a relatively high price, the bank may lose the incentives to adopt proper due diligence and, in general, to research for good borrowers. Moral hazard comes into play because the assets quality to end-users (i.e., the likelihood of default of the original loans) now depends on the efforts the bank makes while originating the loans.

If the investors understood this bank's behavior, they would not be willing to pay the price they pay while assuming the bank sells products with good credit standards. To illustrate, and relying on previous notation, suppose that if the bank does not make any efforts while selecting its borrowers, the upper limit to the loss distribution is $X = b$. Instead, if the bank invests in resources to improve the borrowers' credit standards, $X = b-\varepsilon$, where $\varepsilon \in (0, b)$. Now, these investments cost C, and the investors anticipate that the bank, unless properly incentivized, would not invest in these resources while selling the securitized assets for a high price (they would then mis-report to have exterted all efforts). Absent these incentives, the investors are only willing to pay a price per notional equal to s_b, not the higher price $s_{b-\varepsilon}$, where, and consistent with Eq. (4.3),

$$s_b = \frac{1-E_b(\tilde{l})}{1+\rho}. \tag{4.7}$$

How can investors incentivize the bank to originate loans with high standards? It is a principal-agent problem, where investors act as the "principal" and the bank acts as the "agent." The investors want to make sure the bank invests in resources while originating the loans. When does this happen? It must be that the increase in the payoffs accruing to the bank (due to a better loans quality) after securitization exceeds the costs of ensuring a better loans quality, viz

$$\begin{aligned}&(1-\chi)\left(1-E_{b-\varepsilon}(\tilde{l})\right)+\chi s_{b-\varepsilon}-\frac{\tau}{2}(1-\chi)^2\,var_{b-\varepsilon}(\tilde{l})-C\\ &\geq (1-\chi)\left(1-E_b(\tilde{l})\right)+\chi s_{b-\varepsilon}-\frac{\tau}{2}(1-\chi)^2\,var_b(\tilde{l}).\end{aligned} \tag{4.8}$$

The L.H.S. of (4.8) is the payoff per notional accruing to the bank when it exerts an effort. The R.H.S. is the payoff accruing to the bank when it does not.[3] Note that, in both cases, the

[3]Both payoffs are calculated before the amount k extended to the borrowers, which cancels out from the two sides of the inequality.

investors announce to be willing to pay the price that reflects the asset fundamentals when the bank puts efforts, $s_{b-\varepsilon}$. That is, the investors cannot observe whether the bank will exert (or have already exerted) any effort; however, they are willing to sign a contract in which they stand ready to pay the high price $s_{b-\varepsilon}$, but under the *incentive-compatibility* constraint that the inequality (4.8) holds true.

In other words, the contract specifies, ex ante, a securitization rate χ^{**} satisfying (4.8) and a securitization price $s_{b-\varepsilon}$, and the bank will then originate a loan, retaining a proportion equal to $1-\chi^{**}$ while securitizing this loan. In case the bank had already originated a loan with good credit standards (it may, under parameter conditions specified below), it will then decide whether it is in its interest to securitize it at χ^{**}, as explained in more detail in the next subsection. Note that if the bank had already originated the loan without exterting any effort, and this becomes public information, there are obviously no viable ways to proceed with contracting in the way we are explaining in this section: the trivial solution would be that in the symmetric information case (see Eq. (4.6)), with low securitization prices, $X = b$. That is, securitization does not work "when it arrives too late."

In addition to being incentive compatible, the contract must be such that the bank cannot obtain better opportunities while making business elsewhere. These *participation constraints* will be discussed in the next subsection below (see Eqs. (4.10)-(4.11)).

The contract is incentive-compatible if the bank is able to retain a proportion $1-\chi > 0$ of the securitized loans. This retention of risks makes the bank responsible, so to speak: it signals the investors that it has interests aligned regarding risks arising from poor credit standards. Let us illustrate. Assume the costs are $C = c\varepsilon$, for some constant c, and rewrite (4.8) as an equality

$$(1-\chi) + (1-\chi)^2 \frac{\tau}{12}(2b-\varepsilon) = 2c. \tag{4.9}$$

The L.H.S. of this equation reflects the payoff increase the bank experiences while adopting good credit standards: (i) lower expected losses and (ii) lower volatility. The R.H.S. reflects costs regarding the adoption of a rigorous lending scheme. The contract is incentive-compatible in correspondence of securitization levels χ^{**} such that Eq. (4.9) holds true. For $b = 0.8$, $\varepsilon = 0.2$ and $c = 0.25$, we have $\chi^{**} = 0.59546$.[4] Figure 4.2 (solid line) depicts the optimal securitization rate as a function of the investors' sophistication, i.e., χ^* in Eq. (4.6) as a function of ρ, assuming that $X = b$ (solid line). The horizontal dashed line is the optimal retention rate χ^{**}.

[4]To illustrate, Chomsisengphet and Pennington and Cross (2006) report secutitization rates in 2003 to amount for approximately 76% (Conventional loans), 35% (Jumbo) and 59% (Subprime).

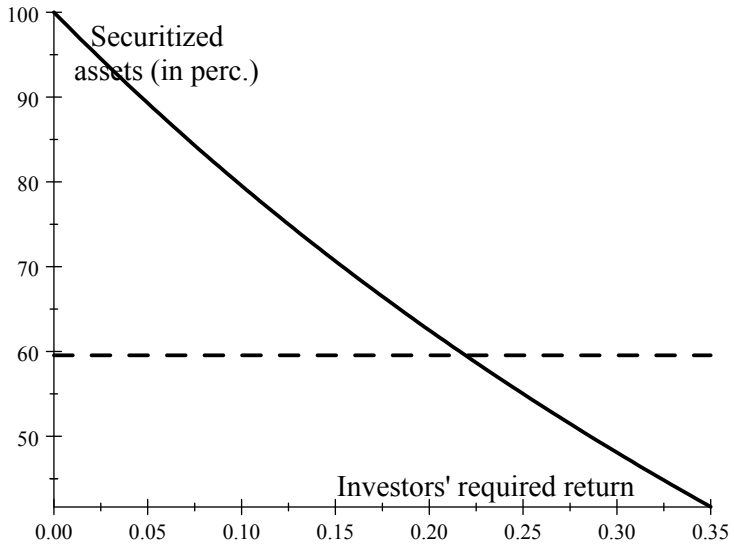


FIGURE 4.2. Proportion of securitized assets as a function of investors' sophistication. The solid line is this proportion in a market with symmetric information and poor credit standards (i.e., χ^* in Eq. (4.6) with $X = b$). The dashed line is the proportion in a contract between banks and investors that is incentive compatible regarding the credit standards adopted by the bank (i.e., χ^{**} solution to Eq. (4.9)). Parameter values are $b = 0.8$, $\varepsilon = 0.2$ and $c = 0.25$.

Note that in a market with symmetric information, a bank would choose a retention rate $1 - \chi^* \approx 40\%$ (approximately the same as in the market with moral hazard and contracting) only when the investors' expected return $\rho \approx 22\%$: the price of the securitized assets has to be sufficiently low to discourage the bank to securitize too much. But in a market plagued with moral hazard, and within more realistic assumptions on investors' sophistication (i.e., $\rho < 22\%$), the retention rate suggested by the contract is higher (i.e., $\chi^* > \chi^{**}$), and increases with ρ. The bank would like to securitize more, but the investors know that while securitizing more, the bank would not be properly incentivized to adopt the credit standards compatible with the high price $s_{b-\varepsilon}$, and would therefore not take the promised amount of efforts.[5]

The whole point of this model is to show that high retention rates are needed to sustain a securitization deal that bears high quality assets. Without such retention rates, a market may well fail to arise, at least in correspondence of the high prices suggested by the contract (in our examples, $s_{b-\varepsilon}$). Having observed default experiences higher than those implied by prices is symptomatic of misselling by the originators or poor understanding of market conditions (i.e., the banks' incentive problems) by institutional investors.[6] For example, during the 2007 subprime crisis, default frequencies for securitized loans were considerably higher for those loans

[5]For example, consider the extreme case $\chi = 100\%$. In this case, the equality in (4.9) is not satisfied in the presence of due diligence costs.

[6]In the model, the credit quality X (the upper limit of the distribution of $\tilde{l}$) can be inferred through Eq. (4.7) as $X = 2(1 - s_X(1 + \rho))$. The frequency of default decreases with the securitized price. Therefore, if the securitized asset is sold at $s_{b-\varepsilon}$ (when, in fact, the risk is $X = b$), one should observe a default frequency higher than predicted by the original sale price $s_{b-\varepsilon}$. Asset mispricing may create a feedback loop and lead to the emergence of systemic risk, under conditions hypothesized in Chapter 14 (see Section 14.6).

with higher ease of securitization (see Keys, Mukherjee, Seru, and Vig, 2010). We shall return to these observations in Chapter 14, when we shall describe additional critical aspects regarding the unfolding of the 2007 credit crisis (see Section 14.6).

4.2.3.4 Participation constraints

We now discuss participation constraints. The bank must be incentivized to sign the contract in the first place. First, the bank has the alternative of not securitizing the loans; therefore, the payoff accruing to the bank while securitizing at χ^{**} must exceed the payoff it would obtain while not securitizing at all:

$$(1-\chi^{**})\left(1-E_{b-\varepsilon}(\tilde{l})\right)+\chi^{**}s_{b-\varepsilon}-\frac{\tau}{2}\left(1-\chi^{**}\right)^{2}var_{b-\varepsilon}(\tilde{l})-C\geq\left(1-E_{b}(\tilde{l})\right)-\frac{\tau}{2}var_{b}(\tilde{l}). \quad (4.10)$$

Second, the bank has the alternative of not originating the loan and invest the original amount of cash kN in the risk-free asset and the market portfolio; therefore, this alternative must yield a payoff that is less than that originating from the securitization deal:

$$(1-\chi^{**})\left(1-E_{b-\varepsilon}(\tilde{l})\right)+\chi^{**}s_{b-\varepsilon}-\frac{\tau}{2}\left(1-\chi^{**}\right)^{2}var_{b-\varepsilon}(\tilde{l})-C-k\geq\frac{1}{2\tau}\lambda^{2}+Rk-k, \quad (4.11)$$

where λ is the Sharpe ratio on the market portfolio (see Chapter 1, Eq. (1.50)).[7] This constraint determines the loan rate $1-k$.

The inequality (4.10) is satisfied for the parameter values of this exercise (and for the values of ρ in Figure 4.2). Furthermore, with these parameter values, and without the option of securitizing, the bank would be better off while exerting an effort rather than not (that is, (4.8) holds for $\chi=0$) and, still, the banks prefers signing the contract for χ^{**} rather than exerting an effort and retaining the loan (albeit for values of $\rho\leq 18\%$). When properly designed, securitization is a means that allows the bank to mitigate risks consistent with its risk-appetite, not a mere re-allocation of mispriced risks.

Finally, Figure 4.3 plots the loan rate the bank requires when (4.11) holds as an equality, the risk-free rate $r=0.01$ and the market Sharpe ratio is $\lambda=0.4$ (solid line); the dashed line is the value of an hypothetical loan rate the bank would require when it generates and retains 100% of a high quality loan and pays a cost for it (and does not compare this investment to the market portfolio). For comparison, the expected loss is $E_{b-\varepsilon}(\tilde{l})=30\%$ and $E_{b}(\tilde{l})=40\%$. The required loan rate is increasing in ρ because a smaller securitized price, $s_{b-\varepsilon}$ requires a lower

[7] By investing $w\equiv k$ into the risk-free asset and the market portfolio leaves next period wealth equal to $w'=\pi\left(b_{M}-r\right)+Rw+\pi B\tilde{u}$, consistent with notation set up in Chapter 1 (see Sections 1.2 and 1.3). Given its mean-variance objective function, the bank would select a proportion of the market portfolio equal to π^{*} in Eq. (1.40) (Chapter 1), such that its value function, $E\left(w'\right)-\frac{\tau}{2}var\left(w'\right)$, would equal the first two terms of the R.H.S. of (4.11).

initial investment k to maintain the balance in (4.11).

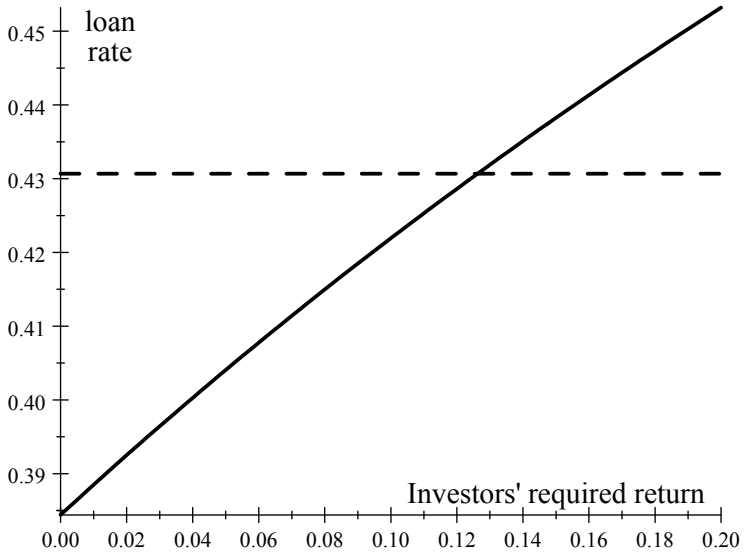


FIGURE 4.3. The solid line depicts the loan rate $1-k$ required by the participation constraint (4.11) (as an equality), as a function of investors' sophistication. The dashed line is the loan rate required when the bank makes an effort and does not securitize the loan. Parameter values are as in Figure 4.2, and $r=0.01$ and $\lambda=0.4$.

4.2.4 Signaling

[In progress]

4.3 The classics: capital structure and Modigliani-Miller propositions

Chapter 2 explains that Modigliani and Miller (1958) first rely on the notion of absence of arbitrage and derive "asset price restrictions." The scope of their analysis is to assess how leverage affects the value of the firm. We now review their important contributions, while relying on their original arguments as well as those relatively more modern arguments, based on Arrow-Debreu security pricing.

4.3.1 The irrelevance of debt

Modigliani and Miller (1958) consider an economy where firms can be sorted by the expected returns of their shares, according to the sector, or class, they belong to. Let π be the constant, expected profit paid off by the each firm within sector k, and E_U be the price of an unlevered firm's share. Under standard conditions, we have that $E_U=\sum_{t=1}^{\infty}(1+\rho_k)^{-t}\,\pi$, where ρ_k is the risk-adjusted discount rate prevailing in sector k, such that the return on equity (ROE) for the unlevered firm is,

$$\rho_k=\frac{\pi}{E_U},$$

a constant for all the unlevered firms belonging to sector k. Naturally, the value of the firm is equal to the value of equity, $V_U=E_U$, say. Next, consider a levered firm operating within the

k-th sector. This firm issues debit with nominal value equal to D such that is value denoted as V_L, equals the sum of equity and debt, $V_L = E_L + D$. In the absence of any market frictions, we have the following irrelevance result:

THEOREM 4.1 (Modigliani & Miller). *In the absence of arbitrage and frictions, the market value of any firm is independent of its capital structure and is given by discoutning its expected profits at the discount rate appropriate to its class: $V_j = \frac{\pi}{\rho_k}$, for any firm $j \in \{U, L\}$ in sector k.*

In other words, the return on investment (ROI), defined as $\rho_k = \frac{\pi}{V_j}$, is the same for two firms that earn the same expected profit π, regardless of the capital structure. Naturally, the ROE and ROI are the same for the unlevered firm.

Modigliani and Miller rely derive these results relying on the following arguments. Consider two firms: a first, unlevered and a second, levered. They both earn the same expected profit, π. Suppose to purchase the shares of the unlevered firm and borrow the same amount of money issued by the levered firm. In the absence of arbitrage or any frictions, the value of this portfolio should equal the value of the levered firm, which is possible as soon as the value of the levered and the unlevered firm are the same.

Formally, let us be given an arbitrary constant $\alpha \in (0, 1)$, and consider the following trade: (i) we buy $N_U = \alpha \frac{E_L + D}{E_U} = \alpha \frac{V_L}{V_U}$ of the unlevered firm; (ii) we sell $N_L = \alpha$ shares of the levered firm. These two trades make the balance of the position worth $-N_U E_U + N_L E_L = -\alpha D$, and so (iii) we borrow αD at the interest rate r, to make this initial position worthless. This portfolio yields: (i) $+N_U \pi$, due to the purchase of the shares of the unlevered firm, (ii) $-N_L (\pi - rD)$, due to the sale of the shares of the levered firm, which of course has to pay interests on its debt, and (iii) $-r\alpha D$, arising to honour the debt we are making to build up the worthless portfolio. Summing up, the profits are $N_U \pi - N_L (\pi - rD) - r\alpha D = \alpha \left(\frac{V_L}{V_U} - 1 \right) \pi$. If $V_L > V_U$, we have an arbitrage opportunity as we may make money out of a worthless portfolio, and if $V_L < V_U$, we have an arbitrage as well, as we could reverse the positions of the worthless portfolio. Therefore, it must be that $V_L = V_U = E_U = \frac{\pi}{\rho_k}$.

4.3.2 Irrelevance through Arrow-Debreu security pricing

[In progress]

4.3.3 The cost of capital and financial frictions

[In progress]

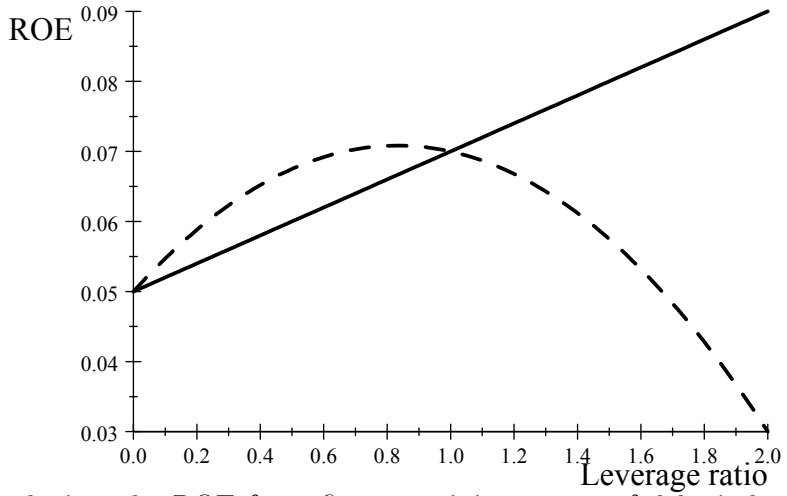
WACC: The weigthed cost of capital.

We have: $\pi = \text{ROI} \cdot V$. Therefore,

$$\text{ROE} = \frac{\pi - rD}{E} = \frac{\text{ROI} \cdot (E + D) - rD}{E} = \text{ROI} + (\text{ROI} - r) \frac{D}{E}.$$

If the interest rate on debt is not affect by the firm capital structure, ROE is increasing in the leverage ratio, $\frac{D}{E}$, once ROI $> r$. Underlying this result is the assumption uof Theorem 2.8 that no-arbitrage trades can be implemented with a cost of borrowing money equal to that of the firm. If borrowers and lenders had asymmetric information, this assumption would break down. For example, debt markets might be concerned about the size of the leverage ratio. For

example, assume that $r = f(\ell)$, where $\ell = \frac{D}{E}$, and in particular that $f(\ell) = 0.03\ell$. Then, we have that: $\text{ROE} = \text{ROI} + (\text{ROI} - 0.03\ell)\,\ell$. The picture below depicts the behavior of ROE as a function of ℓ, assuming that $\text{ROI} = 5\%$ and that the risk-free rate in case of no such frictions is $r = 3\%$.



The solid line depicts the ROE for a firm sustaining a cost of debt independent of the leverage ratio, with $\text{ROI} = 5\%$ and $r = 2\%$. The dashed line is the ROE for a firm that has a cost of debt increasing in the leverage ratio ℓ, $r(\ell) = 0.03\ell$.

Consider the firm with cost of capital depending on the current leverage rato, ℓ. For a low level of ℓ, the ROE increases with ℓ, so as to magnify the difference $\text{ROI} - 0.03\ell$ through the multiplying effect $(\text{ROI} - 0.03\ell)\,\ell$. However, for higher leverage ratios, the difference $\text{ROI} - 0.03\ell$ becomes thinner and thinner, and an increase in ℓ then leads to marginally lower ROE. In this example, there is an interior value for the leverage ratio that maximizes the ROE, which is, approximately, $\ell = 0.83$.

Mention we shall now deal with agency problems, bankrutpcy costs etc. in detail.

[In progress]

4.4 Symmetric information again: full insurance

Why do firm issue debt? One possibility is risk sharing, that is, the possibility that market mechanisms allocate risks according to agents' attitude towards such risks (see Chapter 2, Section 2.62). However, we know that this explanation is problematic.

Consider a risk-averse entrepreneur who has access to a technology to produce a random output $\tilde{x} \in [0, \bar{x}]$ (for some $\bar{x} > 0$), drawn from a probability distribution with density g. The entrepreneur is penniless; he needs an amount of capital equal to I to implement the production plan. Outside investors are willing to provide the entrepreneur with such capital and require an exogenous rate of return equal to ρ. The entrepreneur and outside investors both require limited liability. The contract between these two parties specifies a re-payment schedule $D(\tilde{x})$ due to the investors, which maximizes the entrepreneur's utility, $\max_{D(\tilde{x})\in[0,\tilde{x}]} E\left(u(\tilde{x} - D(\tilde{x})\right)$, subject to the lender break-even condition $E\left(D(\tilde{x})\right) = (1+\rho)\,I$:

$$\max_{D(x)\in[0,x]} \int_0^{\bar{x}} u\left(x - D\left(x\right)\right) g\left(x\right) dx, \quad \text{s.t.} \quad (1+\rho)\,I = \int_0^{\bar{x}} D\left(x\right) g\left(x\right) dx. \tag{4.12}$$

The first order conditions for this problem are

$$\left(u'\left(x-D\left(x\right)\right)-\lambda\right)g\left(x\right)=\eta_0\left(x\right)-\eta_1\left(x\right),$$

where λ is the Lagrange multiplier for the lender break-even condition in (4.12) and $\eta_0\left(x\right)$ and $\eta_1\left(x\right)$ are the limited liability constraints ($D\left(x\right)\geq 0$ and $x\geq D\left(x\right)$, respectively).

When investors' limited liability is not binding, equity is constant, i.e., $\omega\left(x\right)\equiv x-D\left(x\right)=\bar{\omega}$; that is, investors simply redistribute income until the entrepreneur's marginal utility of equity is the same across states, similarly to the "implicit contract theory" of real wages and employment of the 1970s.[8] When, instead, investors' limited liability is binding at zero, $\omega\left(x\right)=x$. All in all, debt and equity play roles that are somehow opposite to those we would have expected. Risk-sharing motives do not justify the existence of standard debt contracts. We now proceed with theories of debt and capital structure relying on asymmetric information.

4.5 Debt and moral hazard

Innes (1990) provides an explanation of debt as a solution to a moral hazard problem: the role of a well-designed financial structure is to mitigate the underprovision of entrepreneurial efforts that may likely occur once a project is funded.

Consider a risk-neutral and penniless entrepreneur; he has access to a technology to produce a random payoff $\tilde{x}\in\left[0,\bar{x}\right]$ ($\bar{x}>0$) through entrepreneurial effort and investment funds equal to I. Investors are willing to finance this project but they can only observe the entrepreneur ex post output $\tilde{x}$, not the efforts he puts in place; furthermore, they require an expected return equal to ρ. We assume that $\tilde{x}$ is continuous with a probability density $g\left(x;e\right)$, where $e\geq 0$ denotes effort. Effort affects production favourably, in that the Monotone Likelihood Ratio Property (MLRP) holds:

$$\frac{\partial}{\partial x}\left(\frac{g_e\left(x;e\right)}{g\left(x;e\right)}\right)>0, \tag{4.13}$$

where the subscript denotes a partial derivative; furthermore, we assume that $\int xg\left(x;0\right)dx=0$. Finally, entrepreneurial efforts cost $c\left(e\right)$, for some increasing and convex function c. We assume that g and c satisfy all regularity conditions that lead to an interior solution to the contracting problem below.[9]

The interpretation of the MLRP property is that higher output realizations are more likely to have been generated by production processes with higher efforts.

[Provide a succinct discussion regarding MLRP]

Entrepreneurs need to be incentivized to exert costly efforts once they receive fundings. Investors propose a contract by which the entrepreneur re-pays $D\left(x\right)$ when the realization of output is $\tilde{x}=x$. Optimal contracting amounts to a re-payment function $D\left(\cdot\right)$ that alignes the incentives of the principal (the investors) to those of the agent (the entrepreneur). Precisely, the contract aims to maximize the entrepreneur's utility or NPV (expected equity minus efforts),

$$\max_{\left\{D(x)\in[0,x],e\right\}}\int_0^{\bar{x}}\left(x-D\left(x\right)\right)g\left(x;e\right)dx-c\left(e\right), \tag{4.14}$$

[8] See, e.g., Chapter 9 in Blanchard and Fischer (1989).

[9] In particular, the value of (4.14) is non-negative as the entrepreneur can always choose to exert no efforts and get a zero NPV (net of efforts).

subject to the incentive compatibility constraint

$$c'(e) = \int_0^{\bar{x}} (x - D(x)) g_e(x;e)\, dx \tag{4.15}$$

and the lender break-even condition

$$(1+\rho) I = \int_0^{\bar{x}} D(x) g(x;e)\, dx. \tag{4.16}$$

Note that the re-payment function $D(\cdot)$ in (4.14) is subject to limited liability on both sides. The incentive compatibility constraint in (4.15) reflects entrepreneurial tradeoffs: marginal costs and benefits of effort compensate with each other. The Lagrangian for this problem is (assuming that limited liability is not binding),

$$\mathcal{L} = \int_0^{\bar{x}} \omega(x) \left(1 - \lambda + \mu \frac{g_e(x;e)}{g(x;e)}\right) g(x;e)\, dx + \lambda \left(\int_0^{\bar{x}} x g(x;e)\, dx - (1+\rho) I\right) - c(e) - \mu c'(e), \tag{4.17}$$

where $\omega(x) \equiv x - D(x)$ denotes the entrepreneur's equity, and μ and λ are Lagrange multipliers for (4.15) and (4.16). Therefore, the optimal pre-payment schedule $D(\cdot)$ is "bang-bang" and equity satisfies

$$\omega(x) = \begin{cases} 0, & \text{if } \lambda > 1 + \mu \frac{g_e(x;e)}{g(x;e)} \\ x, & \text{if } \lambda < 1 + \mu \frac{g_e(x;e)}{g(x;e)} \end{cases}$$

such that, by the MLRP in (4.13), there exists a x^* such that

$$\omega(x) = \begin{cases} 0, & \text{for } x < x^* \\ x, & \text{for } x > x^* \end{cases} \tag{4.18}$$

The left panel of Figure 4.4 depicts the equity schedule $\omega(x)$ (dashed line) and the right depicts the optimal re-payment schedule $D(x)$ (dashed line). The re-payment schedule is a debt contract, in that it specifies a "variable payment" in the bad states of nature, i.e., $D(x) = x$ for $x < x^*$. However, this contract is unusual: it is non-monotonic and exhibits a downward jump: no payments are due to the lender in the good states of nature! To obviate this feature, we now impose a monotonic condition, stipulating that the optimal re-payment schedule is non-decreasing.[10]

Consider the class of continuous but piecewise linear re-payment schedules and, then, equity schedules. It can be shown (see Appendix 1) that there exists a $D^* < x^*$ such that debt takes the standard form of the right panel of Figure 4.4 and, accordingly, equity satisfies

$$\omega(x) = \max\{x - D^*, 0\}. \tag{4.19}$$

It is instructive to consider the "first best" effort, which is the solution to (4.14) subject to (4.16) but not to the incentive compatibility constraint (4.15),

$$\hat{e} = \arg\max_e \int_0^{\bar{x}} x g(x;e)\, dx - (1+\rho) I - c(e).$$

[10] Innes justifies this assumption by explaining that a non-monotonicity would give the entrepreneur the incentive to secretly borrow from another party, report a high cash flow and reduce his repayment and, finally, repay the side loan.

It is possibile to show that in the moral hazard model, effort is less than $\hat{e}$. The intution is that because debt is monotonic in some regions (even without requiring it is always so), some of the efforts made by the entrepreneur are to the benefits of investors.

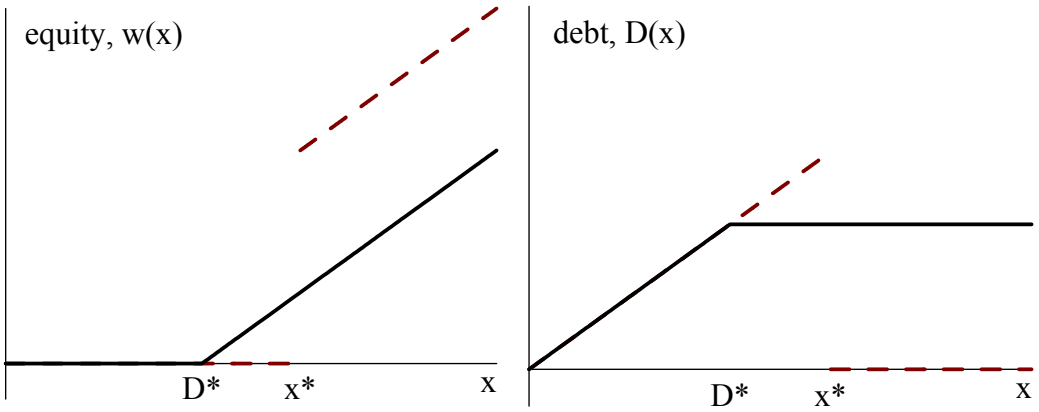


FIGURE 4.4. The left panel of this figure depicts the equity schedule $\omega(x)$ in (4.18) (dashed line) and in (4.19) (solid line) that arise while dispensing with and imposing the monotonicty requirement for the re-payment schedule $D(x)$. The right panel depicts the corresponding re-payment schedules.

4.6 Debt and adverse selection with costly state verification

Gale and Hellwig (1985); Townsend (1979).

4.7 Capital structure and incomplete contracting

Aghion and Bolton (1992); Hart and Moore (1994).

4.8 Appendix 1: Details regarding debt and moral hazard

Proof of Eq. (4.19). Note, first, that monotonicity of $D(\cdot)$ requires that $\omega'(x) \leq 1$. We are considering continuous and piecewise linear equity schedules. We conjecture, and later verify, that there exists a $D^* < x^*$ that maximizes the entrepreneur's profits, such that

$$\omega(x) = \begin{cases} 0, & \text{for } x \in [0, D^*) \\ b_0 + b_1 x, & \text{for } x \geq D^* \end{cases}$$

for two constants b_0 and b_1. Because $\omega(x)$ is continuous, we need that $b_0 = -b_1 D^*$, such that $\omega(x) = b_1(x - D^*)$ for $x \geq D^*$. Therefore, the first term on the R.H.S. of the Lagrangian $\mathcal{L}$ in the main text (see (4.17)) is

$$\mathcal{L}_1(D) \equiv \int_0^{\bar{x}} \omega(x) \underbrace{\left(1 - \lambda + \mu \frac{g_e(x;e)}{g(x;e)}\right) g(x;e)}_{\equiv \zeta(x)} dx = b_1 \underbrace{\int_D^{\bar{x}} (x - D)\zeta(x)\, dx}_{\equiv \mathcal{L}_{11}(D)}.$$

Maximizing $\mathcal{L}_1(D)$ with respect to D leads to the following first order conditions $D^* : 0 = \int_{D^*}^{\bar{x}} \zeta(x)\, dx$. By the MLRP, $\zeta(\cdot)$ changes its sign at x^*, with $\zeta(x) < 0$ for $x < x^*$ and $\zeta(x) > 0$ for $x > x^*$. Therefore, the previous first order condition can only be satisfied if $D^* < x^*$. The second order condition for a maximum is indeed $\mathcal{L}_1''(D^*) = \zeta(D^*) < 0$.

Furthermore, because $\omega'(x) \leq 1$, we need that $b_1 \leq 1$. In particular, we claim that $\mathcal{L}_{11}(D^*) > 0$, in which case $b_1 = 1$ and Eq. (4.19) follows. To show that $\mathcal{L}_{11}(D^*) > 0$, note that

$$\mathcal{L}_{11}(D^*) = \int_{D^*}^{\bar{x}} x\zeta(x)\, dx - D^* \int_{D^*}^{\bar{x}} \zeta(x)\, dx = \int_{D^*}^{\bar{x}} x\zeta(x)\, dx,$$

where the second equality follows by the first order conditions. Therefore,

$$\begin{aligned} \mathcal{L}_{11}(D^*) &= \int_{D^*}^{x^*} x\zeta(x)\, dx + \int_{x^*}^{\bar{x}} x\zeta(x)\, dx \\ &> x^* \int_{D^*}^{x^*} \zeta(x)\, dx + \int_{x^*}^{\bar{x}} x\zeta(x)\, dx \\ &> x^* \int_{D^*}^{x^*} \zeta(x)\, dx + x^* \int_{x^*}^{\bar{x}} \zeta(x)\, dx \\ &= x^* \int_{D^*}^{\bar{x}} \zeta(x)\, dx \\ &= 0, \end{aligned}$$

where the first and the second inequalities follow as $\zeta(x) < 0$ for $x \in [D^*, x^*]$ and $\zeta(x) > 0$ for $x \in [x^*, \bar{x}]$, and the last equality follows by the first order conditions.

References

Aghion, Ph. and P. Bolton (1992): "An Incomplete Contracts Approach to Financial Contracting." *Review of Economic Studies* 77, 473-494.

Akerlof, G.A. (1970): "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84, 488-500.

Arrow, K.J. (1963): "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review* 53, 941-973.

Bhattacharya, S. (1979): "Imperfect Information, Dividend Policy, and the Bird in the Hand Fallacy." *Bell Journal of Economics* 10, 259-270.

Chomsisengphet, S. and A. Pennington-Cross (2006): "The Evolution of the Subprime Mortgage Market." *Federal Reserve Bank of St. Louis Review* 88, 31-56.

DeMarzo, P. (2005): "The Pooling and Tranching of Securities: A Model of Informed Intermediation." *Review of Financial Studies* 18, 1-35.

Gale, D. and M. Hellwig, (1985): "Incentive-Compatible Debt Contracts: The One-Period Problem." *Review of Economic Studies* 52, 647-663.

Glosten, L.R. and P.R. Milgrom (1985): "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders." *Journal of Financial Economics* 14, 71-100.

Grossman, S.J. and J.E. Stiglitz (1980): "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70, 393-408.

Hart, O.D. and J. Moore (1994): "A Theory of Debt Based on the Inalienability of Human Capital." *Quarterly Journal of Economics* 109, 841-879.

Hartman-Glaser, B., Piskorski, T. and A. Tchistyi (2012): "Optimal Securitization with Moral Hazard." *Journal of Financial Economics* 104, 186-202.

Holmström, B. (1979): "Moral Hazard and Observability." *Bell Journal of Economics* 10, 74-91.

Innes, R.D. (1990): "Limited Liability and Incentive Contracting with Ex-ante Action Choices." *Journal of Economic Theory* 52, 45-67.

Keys, B., T. Mukherjee, A. Seru, and V. Vig (2010): "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans." *Quarterly Journal of Economics* 125, 307-362.

Kyle, A.S. (1985): "Continuous Auctions and Insider Trading." *Econometrica* 53, 1335-1335.

Laffont, J.-J. and D. Martimort (2002). *The Theory of Incentives: The Principal-Agent Model.* Princeton: Princeton University Press.

Leland, H. and D.H. Pyle (1977): "Informational Asymmetries, Financial Structure, and Financial Intermediation." *Journal of Finance* 32, 371-87.

Mirrlees, J. (1974): "Notes on Welfare Economics, Information, and Uncertainty." In Balch, M., Mc Fadden, D. and S. and Wu (Editors): *Essays on Economic Behavior under Uncertainty*. Amsterdam: North Holland.

Mirrlees, J.A., (1975): "The Theory of Moral Hazard and Unobservable Behavior: Part I." Nuffield College, Oxford, unpublihsed manuscript.

Modigliani, F. and M. Miller (1958): "The Cost of Capital, Corporation Finance and the Theory of Investment." *American Economic Review* 48, 261-297.

Riley, J. "Competitive Signalling." *Journal of Economic Theory* 10, 174-186.

Ross, S.A. (1973): "The Economic Theory of Agency: The Principal's Problem." *American Economic Review* 63, 134-139.

Ross, S.A. (1977): "The Determination of Financial Structure: The Incentive-Signalling Approach." *Bell Journal of Economics* 8, 23-40.

Rothschild, M. and J. Stiglitz (1976): "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *Quarterly Journal of Economics* 90, 29-649.

Shavell, S. (1979): "On Moral Hazard and Insurance." *Quarterly Journal of Economics* 93, 541-562.

Smith, A. (1776): *The Wealth of Nations*. New York: The Modern Library.

Spence, M. (1973): "Job Market Signaling." *Quarterly Journal of Economics* 87, 355-374.

Spence, M. (1974): *Market Signaling: Information Transfer in Hiring and Related Screening Processes*. Cambridge: Harvard University Press.

Stiglitz, J.E. and A. Weiss (1981): "Credit Rationing in Markets with Imperfect Information." *American Economic Review* 71, 393-410.

Townsend, R.M. (1979): "Optimal Contracts and Competitive Markets with Costly State Verification." *Journal of Economic Theory* 21, 265-293.

5
Continuous time models

5.1 Introduction

This chapter is an introduction to asset pricing models cast in continuous time. As such, it does not not introduce any new economic concept against what we have already learned in previous chapters. Nevertheless, the power of continuous time methods is that they allow to deal with issues arising in economies and markets more complex than those in the previous chapters. Moreover, on an applied perspective, continuous time methods are extremely useful to evaluate derivative instruments that draw value from complex events, such as those relating to baskets of credit events, capital market volatility, or history-dependent developments in fixed income markets, to name a few (see Part III). Continuous time models pose challenges to econometricians—we only observe a discrete realization of an idealized continuous time data generating process. The next chapter surveys tools, based on simulations, which allow us to mitigate these challenges.

This chapter aims to two scopes. The first is to explain in detail how the principle of absence of arbitrage works in continuous time: how do asset prices need to "drift" to ensure that there is no arbitrage? How many possible "drifts" would we expect to see in arbitrage-free markets?

The second objective is to develop technical details about the properties of asset prices in continuous time. For example, we shall see that asset prices, once restricted by absence of arbitrage, satisfy partial differential equations under regularity conditions. Yet asset prices are discounted expectations of their future payoffs, taken under the risk-neutral probability. How are these properties tied together? We shall explain how these properties are tied to each other, by introducing the celebrated Feynman-Kac theorem, which provides a probabilistic representation of the solution to a partial differential equation. Moreover, what is the relation between the risk-neutral probability and the "physical" probability? How do we need to tilt the physical probability to determine the risk-neutral? How many risk-neutral probabilities exist, in incomplete markets or in markets with frictions? Are there natural "pricing probabilities" arising in complex contexts such as those in which interest rates are random? And how these pricing probabilities relate to the notion of *numéraires*? Girsanov theorem is the starting point that we need to deal with these fundamental questions.

The models we consider in this chapter diffusion models (with some extensions that accommodate for jumps), which are the workhorse in finance. Diffusion models are, so to speak, those where the variations of a variable of interest are driven by a deterministic component (the "drift") and a stochastic one (the "diffusion"). Heuristically, the diffusion component is normally distributed over an infinitesimal amount of time, being proportional to the variations of what is known as a "Brownian motion." We typically assume that the fundamentals of the economy follow diffusion processes, and that asset prices are rational, in that they are a function of these fundamentals. Absence of arbitrage restricts the set of all possible pricing functions. The fundamental tool with which we link asset prices to fundamentals is Itô's lemma, a device we need to build new processes (in our case, the asset prices) from old ones (the fundamentals of the economy). The complication in finance is that these new processes, albeit a function of the fundamentals, are not given in advance; instead, they are the focus of research.

The chapter is organized as follows. The next section is an introduction to methods. It deals with details leading to the birth of continuous time finance—the Black & Scholes formula of evaluation of European options; it also illustrates how continuous time models obtain as limiting cases of the discrete time models in the previous chapter, and also describes basic properties of long-lived asset prices, such as (i) the fundamental relations that link expected returns, volatilities (the "betas") and risk-premiums (the "lambdas"); and (ii) a representation of the price-dividend ratio in terms of certain possibly varying discount rates—the risk-adjusted discount rates. These derivations turn out to be useful while discussing the properties of equity markets in Part II (see Chapters 8 and 9).

Section 5.3 ... Finally, the Appendix provides technical details omitted from the main text, including a self-contained appendix containing notions of stochastic calculus.

[In progress]

5.2 An introduction to no-arbitrage and equilibrium

5.2.1 Time

By definition, long-lived assets do not have an expiration date and their prices do not explicitly depend on calendar time,[1] at least insofar as the state variables driving them do not explicitly depend on time. Naturally, it does not mean that these prices cannot be time-varying, or stochastic. On the contrary, they can well be driven by the random state variables, as we shall explain in a few sections.

Derivatives do, instead, have an expiration date, and their value reflects the time left to maturity, when they will be worth the terminal payoff. The presence of an expiration date implies that derivative prices solve partial differential equations, which also involve the price sensitivity to time. In the absence of any expiration date, asset prices are still solutions to partial differential equations, but their sensitivity to time is zero, at least provided that the state variables driving are Markov. In the simple case of a single state variable, long-lived asset prices are, then, solutions to ordinary differential equations, as in the Lucas model with a single state variable of Section 5.3.

The next section develops the simplest possible setting where we can illustrate the main tools of continuous time finance, which is the Black and Scholes (1973) market leading to the evaluation formula for a European option. In the absence of arbitrage, the derivative price is

[1] See, e.g., Eq. (5.28) below.

solution to a certain partial differential equation, the solution of which, we can represent as a conditional expectation taken under the risk-neutral probability. In Section 5.3.3, we provide the link between this risk-neutral probability and the original probability.

5.2.2 The origins: Black & Scholes

5.2.2.1 Self-financed strategies

A self-financed portfolio leads to a situation where the change in value of the portfolio between two instants t and $t+dt$ is determined as a mark-to-market P&L: the change in the asset prices times the quantities of the same assets held at time t: there is no injection or withdrawal of funds between any two instants. For example, let θ_1 and S be the number of shares and the price of some risky asset, and θ_2 and b be the number of some riskless assets and its price. Then, the value of a self-financed portfolio, $V = S\theta_1 + \theta_2 b$, satisfies:

$$dV = \theta_1 dS + \theta_2 db = \pi \left(\frac{dS}{S} - \frac{db}{b} \right) + \frac{db}{b} V,$$

where $\pi \equiv S\theta_1$ and the second equality follows by simple calculations. If the portfolio strategy involves risky assets distributing a dividend process, and consumption, the value of the self-financed portfolio satisfies (see Appendix 2 for details):

$$dV_t = \left(\frac{dS_t}{S_t} + \frac{D_t}{S_t} dt - r dt \right) \pi_t + r V_t dt - c_t dt, \tag{5.1}$$

where D_t is the dividend process.

5.2.2.2 Black & Scholes partial differential equation

Why are partial differential equations so important in finance? Suppose that the price of a stock follows a geometric Brownian motion:

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t, \quad \mu, \sigma > 0,$$

and that there exists a riskless accounting technology, or money market account (MMA, henceforth) making spare money evolve as:

$$\frac{dM_t}{M_t} = r dt,$$

where $r > 0$. Finally, suppose that there exists another asset, a "call option," which gives rise to a payoff equal to $(S_T - K)^+$ at some future date T, where K is the "strike," or exercise price of the option. Let $(c_t)_{t \in [0,T]}$ be the option price process. We wish to figure out what this price looks like while formulating as few assumptions as possible. We ignore dividend issues, assume there are no transaction costs, and rule out any other frictions. We assume *rational expectations*, that is, there exists a function $f : c_t = f(t, S_t)$, and assume that this function is as differentiable as needed for an application of Itô's lemma, such that:

$$dc = (Lf)\, dt + f_S \sigma S dW,$$

where Lf is the "infinitesimal generator," defined as $Lf = \frac{\partial f}{\partial t} + \frac{1}{2}\sigma^2 S^2 f_{SS} + \mu S f_S$, with subscripts denoting partial derivatives. Next, we create the following portfolio: θ units of the risky asset

and invest as a residual of the wealth, θ_M unit in the MMA, M_t. As explained, for any portfolio strategy to be "self-financed," the value V of the resulting portfolio, $V_t = \theta_t S_t + \theta_{Mt} M_t$, must be such that:

$$\begin{aligned} dV_t &= \theta_t dS_t + \theta_{Mt} dM_t \\ &= (\theta_t \mu S_t + r\theta_{Mt} M_t)\, dt + \theta_t \sigma S_t dW_t. \end{aligned}$$

Now, set $V_0 = c_0$. Moreover, let us actually conjecture that we could choose θ_t and θ_{Mt} such that $V_t = c_t$ for all t, i.e. that the self-financed strategy can replicate the option price. We obviously need to check this conjecture below. For now, note that if $V_t = c_t$ for all t, the drifts and diffusion coefficients of both V_t and c_t have to be the same, by a result stated in Appendix 1, known as the *unique decomposition property*. Diffusion terms are the same when $\hat{\theta} = f_S$. Replacing this $\hat{\theta}$ into the dynamics of V produces:

$$dV = (\mu S f_S + r\theta_M M)\, dt + f_S \sigma S dW. \tag{5.2}$$

Next, take $\hat{\theta}_M : \text{drift}(V) = \text{drift}\,(c)$, i.e.,

$$\hat{\theta}_M = \frac{Lf - \mu f_S S}{rM}. \tag{5.3}$$

Because we are conjecturing that $V_t = c_t$ for all t:

$$f = c = \hat{V} \equiv \hat{\theta} S + \hat{\theta}_M \hat{M} = f_S S + \frac{Lf - \mu f_S S}{r}.$$

By the definition of L, and rearranging terms,

$$\frac{\partial f}{\partial t} + \frac{1}{2}\sigma^2 S^2 f_{SS} + rSf_S - rf = 0, \quad \forall (t, S) \in [0, T) \times \mathbb{R}_{++}, \tag{5.4}$$

subject to the "boundary condition" $f(T, S) = (S - K)^+$ for all S. Eq. (5.4) is a *partial differential equation*; that is, we are searching for an unknown function f, which has to be such that once it and its partial derivatives are plugged into the left hand side of the first line, we obtain zero. Moreover, the same functions must pick up the boundary condition. The solution to this is the celebrated Black and Scholes (1973) formula.[2]

It remains to verify the conjecture that $V_t = c_t$ holds for all t. We deal with this issue in the next subsection while relying on a slightly more general perspective than Black & Scholes.

5.2.2.3 Non-hedgeable claims

Suppose that there exists a traded asset and that its price satisfies,

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t,$$

with obvious notation. We want to replicate a generic "risk," a process $\varphi_t \equiv \varphi\,(S_t, t)$. By Itô's lemma,

$$d\varphi = \varphi_S dS + \left(\frac{\partial \varphi}{\partial t} + \frac{1}{2}\sigma^2 S^2 \varphi_{SS}\right) dt.$$

[2]The case where $r = 0$ is dealt with similarly. If $r = 0$, we need to have $Lf - \mu f_S S$, by Eq. (5.2) or, analogously, by Eq. (5.3), which leads to Eq. (5.4), with $r = 0$. In this case, the portfolio process is set equal to $\hat{M}_0 = V_0 - \hat{\theta}_0 S_0$ at inception of the strategy, and then, kept constant, such that at each point in time, $V_t = \hat{\theta}_t S_t + \hat{M}_0$, with $dV_t = \hat{\theta}_t dS_t$.

Consider as usual a self-financed strategy V_t, the value of which satisfies:

$$dV_t = \theta_t dS_t + \theta_{M,t} dM_t.$$

We have,

$$dV - d\varphi = (\theta - \varphi_S)\, dS + \left(r\theta_M M - \left(\frac{\partial \varphi}{\partial t} + \frac{1}{2}\sigma^2 S^2 \varphi_{SS} \right) \right) dt. \tag{5.5}$$

Therefore, and consistently with the Black-Scholes derivation in the previous subsection, a *necessary* condition for V_t to replicate φ_t is that

$$\hat{\theta}_t = \varphi_S(S_t, t), \quad \text{for all } t \in [0, T]. \tag{5.6}$$

Next, suppose φ is the price of a *traded* asset, one that delivers a payoff equal to $\varphi(S_T, T) = \psi(S_T)$ at time T, for some function ψ, and that at the same time, satisfies the partial differential equation (5.7) below. We need to have that $V_0 = \varphi_0^{\$}$ where $\varphi_0^{\$}$ denotes the market price of the asset. For suppose not and, e.g., $V_0 < \varphi_0^{\$}$. We could sale short the asset for $\varphi_0^{\$}$, and implement a self-financing strategy with $\hat{\theta}_t$, such that $V_0 = \varphi(S_0, 0)$, where φ satisfies,

$$L\varphi(S,t) = r\varphi(S,t), \quad \text{with boundary condition } \varphi(S_T, T) = \psi(S_T). \tag{5.7}$$

We claim that $V_t = \varphi(S_t, t)$ for all $t \in [0, T]$. Indeed, by Eq. (5.5),

$$\begin{aligned} dV - d\varphi &= \left(r\theta_M M - \left(\frac{\partial \varphi}{\partial t} + \frac{1}{2}\sigma^2 S^2 \varphi_{SS} \right) \right) dt \\ &= \left(r(V - \varphi_S S) - \left(\frac{\partial \varphi}{\partial t} + \frac{1}{2}\sigma^2 S^2 \varphi_{SS} \right) \right) dt \\ &= r(V - \varphi)\, dt. \end{aligned} \tag{5.8}$$

where the first equality follows by Eq. (5.6); the second by the self-financing condition, $\theta_M M = V - \theta S$ and again by (5.6); and the third by (5.7). Because $V_0 = \varphi(S_0, 0)$, then (5.8) implies $V_t = \varphi(S_t, t)$ for all t, as claimed, and then $V_T = \varphi(S_T, T) = \psi(S_T)$ too, which allows to honour the short-sale of the asset.[3] Note that these arguments do not require that a market for this asset exists over the life of the asset.

The crucial assumption underlying the property that the strategy value replicates the risk $\varphi(S_t, t)$, $V_t = \varphi(S_t, t)$ for all t, is that $\varphi(S_t, t)$ *is the price of a traded asset*, i.e., Eq. (5.7) holds true. We can show the converse. That is, suppose the strategy value replicates the risk through Eq. (5.6). In this case, the L.H.S. of (5.5) is zero, and by Eq. (5.6),

$$\frac{\partial \varphi}{\partial t} + \frac{1}{2}\sigma^2 S^2 \varphi_{SS} = r\theta_M M = r(V - \theta S) = r(\varphi - \varphi_S S). \tag{5.9}$$

Then, the risk, φ, satisfies Eq. (5.7).

Next, suppose that φ is *not* the price of a traded asset—for example, $\varphi(\cdot) = S^2$. Then, the counterpart to (5.6) would be $\hat{\theta}_t = \varphi'(S_t)$, the counterpart to (5.8) would be,

$$dV - d\varphi = \left(r\theta_M M - \frac{1}{2}\sigma^2 S^2 \varphi'' \right) dt, \tag{5.10}$$

[3]Suppose the opposite, i.e. that $\varphi_0^{\$} < V_0 = \varphi(0, S_0)$. We could then buy the asset for $\varphi_0^{\$}$ and, hence, claim for $\psi(S_T)$ at T. We could at the same time, "short-sell" the portfolio for $V_0 > \varphi_0^{\$}$ (guaranteeing a positive profit at time-0), which we could update through through the self-financed strategy $\hat{\theta}_t$, up to time T, when it will deliver $-V_T = -\psi(S_T)$, which we could honour through the long static position in the asset initiated at time-0.

and the latter is identically zero when the following counterpart to (5.9) holds,

$$\frac{1}{2}\sigma^2 S^2 \varphi'' = r\theta_M M = r\left(V - \theta S\right) = r\left(\varphi - \varphi' S\right).$$

However, this ordinary differential equation cannot hold, unless $r + \sigma^2 = 0$. That is, assume that $V_t = \varphi\left(S_t\right)$ for all t; then, the left-hand side of Eq. (5.10) cannot equal zero, contradicting that $V_t = \varphi\left(S_t\right)$ holds for all t.

All in all, we cannot replicate $\varphi\left(S_t\right) = S_t^2$ through a self-financed strategy. Naturally, we could replicate the payoff at maturity, S_T^2, provided a market exists for a claim to S_T^2 (see Eq. (5.22) in the following section), not necessarily a market over the entire life of this derivative. The price of this asset is obviously not $\varphi\left(S_\tau\right) = S_\tau^2$ for each τ though. To determine the *tracking error* of the portfolio $\varphi\left(S_t\right) = S_t^2$, note that

$$\begin{aligned} dV - d\varphi &= \left(r\theta_M M - \frac{1}{2}\sigma^2 S^2 \varphi''\right) dt \\ &= \left(r(V - \varphi' S) - \frac{1}{2}\sigma^2 S^2 \varphi''\right) dt \\ &= \left(rV - \left(2r + \sigma^2\right) S^2\right) dt. \end{aligned} \tag{5.11}$$

Eq. (5.11) confirms that $dV = d\varphi$ only when $L\varphi = r\varphi$, which would imply that $r + \sigma^2 = 0$: an asset price equal to S_t^2 would be sustained only under this implausible parameter restriction.

An alternative to a self-financed strategy generating a tracking error, is a strategy, which albeit not being self-financed, guarantees that the risk process is replicable. For example, one could set $\hat{\theta}_t = \varphi'\left(S_t\right)$, and

$$\hat{\theta}_{Mt} M_t = \varphi\left(S_t\right) - \hat{\theta}_t S_t. \tag{5.12}$$

While Eq. (5.12) allows by construction to replicate φ, it also obviously generates an *hedging cost* ε_t that satisfies: $d\varepsilon_t = d\varphi - dV$. In the context of our example,

$$\begin{aligned} dV - d\varphi &= \left(r\theta_M M - \frac{1}{2}\sigma^2 S^2 \varphi''\right) dt \\ &= \left(r(\varphi - \varphi' S) - \frac{1}{2}\sigma^2 S^2 \varphi''\right) dt \\ &= -\left(r + \sigma^2\right) S^2 dt, \end{aligned} \tag{5.13}$$

where V is a self-financed strategy implemented with $\hat{\theta}_t$.

In other words, the second line of (5.11) uses the *self-financing condition* $\hat{\theta}_{Mt} M_t = V_t - \hat{\theta}_t S_t$, while the second line of (5.13) relies on the *replicability condition* (5.12). Assuming $\varepsilon_t = 0$, the tracking error arising from (5.13) is strictly positive process, and at time T it equals,

$$\varepsilon_T = \left(r + \sigma^2\right) \int_t^T e^{-(r+\sigma^2)(T-s)} V_s ds.$$

In Section 5.6, we shall return to the issues regarding replicability of claims in the context of incomplete markets.

5.2.2.4 Surprising cancellations

The option price predicted by the Black and Scholes model is independent of the drift of the underlying asset. After reading this chapter, the reader will interpret this result as follows. Asset prices (rescaled by the money market account) are martingales under the so-called risk-neutral probability, say Q. Therefore, their value is equal to the discounted expectation of its payoff under Q, that is, the probability under which the stock price drifts proportionally to r. In other words, μ doesn't matter.

Let us analyze the details of this property in the context of the previous replicating arguments, by relying on a very simple example. Assume that,

$$\left\{ \begin{array}{lcl} \dfrac{dS_t}{S_t} & = & \mu_t dt + \sigma dW_{1,t} \\ d\mu_t & = & \mu_{D,t} dt + \sigma_{D,t} dW_{2,t} \end{array} \right. \tag{5.14}$$

where σ is a constant, $\mu_{D,t}$ and $\sigma_{D,t}$ are some function of calendar time and S_t and μ_t, and $W_{j,t}$, $j = 1, 2$, are two standard Brownian motions.

Consider the function $\varphi_t = \varphi(S_t, \mu_t, t)$, and assume it is solution to the following partial differential equation, generalizing (5.7),

$$L^{S,\mu}\varphi = r\varphi + \varphi_S (\mu - r) S + \varphi_\mu(\mu_D - \hat{\mu}_D), \quad \text{with boundary condition } \varphi(T, S_T) = \psi(S_T), \tag{5.15}$$

where $L^{S,\mu}\varphi(S, \mu, t)$ is the infinitesimal generator of the diffusion process (5.14), and $\hat{\mu}_D$ is some function, interpreted as the drift of μ_t under the risk-neutral probability. Note that once $\varphi_\mu = 0$ for each (t, S, μ), Eq. (5.15) collapses to the Black-Scholes equation (5.7). We shall return to this point soon.

Next, consider a self-financing strategy invested into the asset and the money market account, just as in the previous section. We have, generalizing Eq. (5.5), that

$$\begin{aligned} dV - d\varphi = {} & (\theta - \varphi_S) dS - \varphi_\mu d\mu \\ & + \left(r\theta_M M - \left(\frac{\partial \varphi}{\partial t} + \frac{1}{2}\sigma^2 S^2 \varphi_{SS} + \frac{1}{2}\sigma_D^2 \varphi_{\mu\mu} \right) \right) dt. \end{aligned}$$

Let $\theta = \varphi_S$, use the definition of a self-financing strategy (i.e., $\theta_M M = V - \theta S = V - \varphi_S S$), and (5.15), such that the previous equation collapses to

$$dV - d\varphi = r(V - \varphi) dt - \varphi_\mu \left(d\mu - \varphi_\mu \hat{\mu}_D dt \right).$$

That is, taking $\varphi_\mu = 0$ implies that the strategy is self-financing and replicates the option value. By no-arbitrage, the price of the option is then the Black-Scholes price even in this market with random drift. Chapter 11 (see Section 11.6) explains that it is in general impossible to replicate the option in case the asset return volatility is random: the Black-Scholes price is generally no longer the no arbitrage price in a market with random volatility.

5.2.3 Asset prices as Feynman-Kac representations

5.2.3.1 A digression on more general partial differential equations

The Black-Scholes equation (5.4) is a typical (in fact the first) example of partial differential equations in finance. It leads to an equation of the so-called parabolic type, as we shall explain soon. More generally, let us be given,

$$a_0 + a_1 F_t + a_2 F_S + a_3 F_{SS} + a_4 F_{tt} + a_5 F_{St} = 0, \tag{5.16}$$

subject to some boundary condition. This partial differential equation is called: (i) elliptic, if $a_5^2 - 4a_3a_4 < 0$; (ii) parabolic, if $a_5^2 - 4a_3a_4 = 0$; (iii) hyperbolic, if $a_5^2 - 4a_3a_4 > 0$. The typical partial differential equations arising in finance are of the parabolic type. For example, the Black-Scholes function $F = e^{rt}f$ is parabolic. The following section explains how to provide a probabilistic representation to these parabolic partial differential equations.

5.2.3.2 Feynman-Kac solutions to partial differential equations

The typical situation that we encounter in finance is that the asset price is a function F that solves a parabolic partial differential equation, i.e. a special case of Eq. (5.16):

$$-r(x,t)F(x,t) + F_t(x,t) + \mu(x,t)F_x(x,t) + \frac{1}{2}\sigma^2(x,t)F_{xx}(x,t) = 0, \quad \forall (t,x) \in [0,T) \times \mathbb{R}, \tag{5.17}$$

with the boundary condition, $F(x,T) = g(x,T)$ for all x, and the function g is the final payoff. Somehow surprisingly, define, now, a stochastic differential equation, with drift and diffusion μ and σ in Eq. (11.27),

$$dZ_t = \mu(Z_t,t)\,dt + \sigma(Z_t,t)\,dW_t, \quad Z_0 = x. \tag{5.18}$$

where W_t is a Brownian motion. Under regularity conditions on μ, σ, r, the solution F to Eq. (5.17) is

$$F(x,t) = E\left(e^{-\int_t^T r(Z_s,s)ds} g(Z_T,T) \middle| z_t = x \right), \tag{5.19}$$

where Z is solution to Eq. (5.18), and the expectation is taken with respect to the distribution of Z in Eq. (5.18). Note that the existence of the Feynman-Kac representation does not ensure per se the existence of a solution to a given partial differential equation.

Eq. (5.19) can be used to represent the solution to the Black & Scholes partial differential equation (5.4), with "auxiliary" stochastic differential equation (5.18) collapsing to,

$$\frac{dS_t}{S_t} = rdt + \sigma d\tilde{W}_t,$$

where $\tilde{W}_t$ is a Brownian motion, which is defined under the risk-neutral probability, due to the drift of $\frac{dS}{S}$ being equal to the risk-free rate, r. That is, by Eq. (5.19), the price of an option in a Black & Scholes market is the risk-neutral expectation of the final payoff, discounted at the risk-free rate.

The Feynman-Kac representation of the solution to partial differential equations is quite useful. First, computing expectations is generally both easier and more intuitive than finding a solution to partial differential equations through guess and trial. Second, except for specific cases, the solution to asset prices is unknown, and a natural way to cope with this problem is to go for Monte-Carlo methods—approximation of the expectation in Eq. (5.19) through simulations and use of the law of large numbers. Finally, the Feynman-Kac representation theorem is useful for some theoretical reasons we shall see later in this chapter.

5.2.3.3 A few heuristic proofs

We provide heuristic proofs of the Feynman-Kac representation theorem; rigourous details of this proofs would go well beyond the purpose of this lectures.[4] We consider a slightly more

[4]For the more technical details, excellent sources of reference are still Friedman (1975) and Karatzas and Shreve (1991).

general case in which

$$Lf + q - rf = 0, \tag{5.20}$$

subject to some boundary condition. We interpret f as the asset price and q as its instantaneous dividend rate. We assume that q is some function $q(x,t) \equiv q_t$. We have, as usual, that $Lf = \frac{\partial f}{\partial t} + \mu f_x + \frac{1}{2}\sigma^2 f_{xx}$.

So suppose that there exists a solution to Eq. (5.20), and define

$$y_t \equiv e^{-\int_0^t r_u du} f(t, z_t) + \int_0^t e^{-\int_0^u r_s ds} q_u du,$$

where

$$dz_t = \mu_t dt + \sigma_t dW_t.$$

By Itô's lemma,

$$dy = e^{-\int_0^t r_u du}\left((Lf + q - rf)dt + f_z \sigma dW\right) = e^{-\int_0^t r_u du} f_z \sigma dW,$$

where the second equality follows by Eq. (5.20). Therefore, and under regularity conditions, y is a martingale, with $y_0 = E(y_T)$, such that

$$f(0, z_0) = y_0 = E(y_T) = E\left(e^{-\int_0^T r_t dt} f(T, z_T)\right) + E\left(\int_0^T e^{-\int_0^u r_s ds} q_u du\right).$$

That is, the asset price at time-0, $f(0, z_0)$, equals the expected discounted value at T and the sum of the expected discounted dividends up to maturity. We now move to deal with another pillar of continuous-time evaluation: the Girsanov theorem.

5.2.4 The Girsanov theorem

5.2.4.1 Motivation again

Consider the Black & Scholes partial differential equation (5.4). We now know that by the Feynman-Kac theorem, the price f can be represented as the discounted expectation of its terminal payoff, $f(T, S) = (S - K)^+$,

$$f(t, S_t) = e^{-r(T-t)} E_t^Q (S_T - K)^+, \tag{5.21}$$

where E_t^Q denotes the expectation taken conditionally upon the information set at t, and with respect to a new probability Q (say), under which S_t is solution to,

$$\frac{dS_t}{S_t} = rdt + \sigma d\tilde{W}_t,$$

where $\tilde{W}$ is a Brownian motion under Q. We refer to Q as the *risk-neutral probability* for reasons similar to those explained in Chapters 1 through 3: under Q, the expected rate of return of both the derivative and the underlying asset price is the riskless rate, r.

Note that the methodology leading to Eq. (5.21) applies to other derivatives than Black-Scholes. Consider, for example, a "quadratic derivative," i.e. one that pays off the square of the asset price, at time T, S_T^2. The arguments made in the previous sections, relating to the replication of this derivative are still the same—with a portfolio process that still includes

$\hat{\theta} = f_S$ units of the risky assets, to be calculated below. The price of this derivative, then, still satisfies Eq. (5.21), although the boundary condition is $f(T,S) = S^2$, such that the price can be expressed as,

$$f(t,S_t) = e^{-r(T-t)} E_t^Q \left(S_T^2\right) = e^{-r(T-t)} S_t^2 e^{\left(2r+\sigma^2\right)(T-t)}. \tag{5.22}$$

This implies that the "hedging portfolio" is $\hat{\theta}_\tau = 2S_\tau e^{\left(r+\sigma^2\right)(T-\tau)}$. Indeed, it is easy to check that the value of the replicating strategy, $V_\tau = \hat{\theta}_\tau S_\tau + \hat{M}_\tau$, coincides with Eq. (5.22), as $\hat{\theta}_\tau = 2\frac{f(\tau,S_\tau)}{S_\tau}$ and, using Eq. (5.3) and Eq. (5.22), $\hat{M}_\tau = -f(\tau,S_\tau)$. Chapter 11 provides a thorough discussion of the hedging portfolio in a Black-Scholes market.

5.2.4.2 Relation between Q and P: heuristic details

How does the risk-neutral probability Q relate to the physical one, P? Consider the following heuristic details. Note that under P, S_t is solution to

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t,$$

where W_t is a Brownian motion under P. Moreover, define $\eta_T \equiv \frac{Q(d\omega)}{P(d\omega)}$, and note that

$$\begin{aligned} E_t^Q (S_T - K)^+ &= \int (S_T(\omega) - K)^+ Q(d\omega) \\ &= \int \eta_T(\omega) (S_T(\omega) - K)^+ P(d\omega) \\ &\equiv E_t \left(\eta_T (S_T - K)^+\right). \end{aligned}$$

We claim that η_t is solution to

$$\frac{d\eta_t}{\eta_t} = -\lambda dW_t, \quad \eta_0 = 1,$$

where $\lambda = \frac{\mu - r}{\sigma}$. Indeed, define $y_t \equiv \eta_t f_t$. Because $f_T = (S_T - K)^+$, we have that $y_0 = f_0 = e^{-rT} E(y_T)$. That is, $e^{-rt} y_t$ is a martingale under the probability P. But by Itô's lemma,

$$d\left(e^{-rt} y\right) = e^{-rt}\eta \left(\frac{\partial f}{\partial t} + \frac{1}{2}\sigma^2 S^2 f_{SS} + (\mu - \lambda\sigma) S f_S - rf\right) dt + e^{-rt}\eta (\sigma S f_S - \lambda f) dW.$$

Under standard regularity conditions (including pathwise integrability), $e^{-rt} y_t$ is a martingale when

$$0 = \frac{\partial f}{\partial t} + \frac{1}{2}\sigma^2 S^2 f_{SS} + (\mu - \lambda\sigma) S f_S - rf. \tag{5.23}$$

On the other hand, we know that f is solution to the Black-Scholes partial differential equation

$$0 = \frac{\partial f}{\partial t} + \frac{1}{2}\sigma^2 S^2 f_{SS} + rS f_S - rf. \tag{5.24}$$

Comparing Eq. (5.23) and Eq. (5.24) reveals $\lambda = \frac{\mu - r}{\sigma}$. The interpretation of λ is that of a *unit risk-premium* for investing in the stock. We shall return to this important interpretation below.

The point of the previous derivations is that it is *as if* we start from the original probability space under which

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t, \tag{5.25}$$

and, then, define a new Brownian motion $d\tilde{W} = dW + \lambda dt$, such that replacing $\tilde{W}$ into Eq. (5.25) delivers

$$\frac{dS_t}{S_t} = rdt + \sigma d\tilde{W}_t.$$

under some new probability space. Note that the Brownian motion we initially begun is obviously defined under P, although "pricing" is "operating" under a different space. This is one amongst the most crucial concepts in Financial Economics, and is formalized in the next section.

5.2.4.3 The theorem

Consider two probabilities Q and P, linked through $Q(A) = \int_A \eta(\omega) P(d\omega)$ for $A \in F$, and the Radon-Nikodym derivative,

$$\left.\frac{dQ}{dP}\right|_{F_T} = \eta_T = \exp\left(-\frac{1}{2}\int_0^T \|\lambda_t\|^2 dt - \int_0^T \lambda_t dW_t\right),$$

for some process λ,[5] and W_t is a Brownian motion under P. Then,

$$\tilde{W}_t = W_t + \int_0^t \lambda_u du,$$

is a Brownian motion under the probability Q.

The Girsanov theorem is the counterpart to the static and heuristic change of probability dealt with in Chapter 1 (see Section 1.4). Intuitively, in continuous-time, a change in probability leads to a "mean-shift" of the random variable we are changing the distribution of. The fact this change does not entail a change in volatility is well-known in continuous-time finance. The economic interpretation is that, other things being equal, the shift towards the left reflects a more pessimistic assessment of the asset return distribution.[6]

Finally, consider the following few additional details regarding asset evaluation. Suppose we wish to price a claim expiring at T. Heuristically, we have

$$E_t^Q(X_T) = \int X_T dQ = \int X_T \left(\frac{dQ}{dP}\right) dP = E_t\left[\left(\frac{dQ}{dP}\right) X_T\right] = E_t(\eta_T X_T).$$

Similarly, we can "update" the previous formula as time unfolds, as follows

$$E_t^Q(X_T) = \eta_t^{-1} E_t(\eta_T X_T),$$

where

$$\frac{d\eta_t}{\eta_t} = -\lambda_t dW_t, \quad \eta_0 = 1.$$

[5] This process is required to satisfy the so-called Novikov's condition, $E(\exp(\frac{1}{2}\int_0^T \|\lambda_t\|^2 dt)) < \infty$. This condition ensures that Q integrates to one, $E(\eta_T) = 1$ or $\int dQ = 1$.

[6] This property does not necessarily hold with other models, or even in discrete time settings, such as those generated by a binomial distribution, as illustrated by the tree models in Chapter 8 or in Part III (see Chapters 11 and 12).

Section 5.4 explains additional properties regarding this change of probabilities. We now proceed to provide more economic structure regarding the risk-premium λ_t in a slightly more general context, one that also includes evaluation of long-lived assets. We shall see that η_t relates to the stochastic discount factor through what is known as the "pricing kernel."

5.2.5 The APT in continuous time

We explain how asset prices link to a number of state variables by deriving a continuous time version of the APT (see Chapter 2), in which the asset expected excess returns form a linear combination of the asset exposures to factors, with weights equal to the unit risk-premiums the market requires to bear the risk arising from each of these factors. We begin with a heuristic derivation of the pricing kernel in continuous time as the limit of a discrete time model; then, we characterize the market expected returns in terms of the APT while relying on a diffusive model.

5.2.5.1 Prices and pricing kernels, from discrete to continuous time

Let S_t be the price of a long-lived asset as of time t, and $D_{t+h} \cdot h$ the dividend paid by this asset at t over a small trading period h. We know that in the absence of arbitrage, there exists a positive process $m_{t,\cdot}$, known as the *stochastic discount factor*, such that the price of any asset is the expectation of its future payoff, weighted with $m_{t,t+h}$,

$$S_t = E_t \left[m_{t,t+h} \left(S_{t+h} + D_{t+h} \cdot h \right) \right], \tag{5.26}$$

where E_t is the conditional expectation given the information set at time t. For example, in an economy with a representative risk-neutral agent, we have that $m_{t,t+h} = e^{-r_t h}$, where r_t is the risk-free rate per unit of time.

Given the stochastic discount factor, we define, as usual, the *pricing kernel*, or state-price, process, ξ_t, as the process that grows by the stochastic discount factor:

$$\xi_{t+h} = m_{t,t+h} \xi_t, \quad \xi_0 \equiv 1.$$

In terms of the pricing kernel, Eq. (5.26) is $0 = E_t \left(\xi_{t+h} S_{t+h} - S_t \xi_t \right) + E_t \left(\xi_{t+h} D_{t+h} h \right)$, which in the limiting case of h that tends to zero, is

$$0 = E_t \left[d \left(\xi_t S_t \right) \right] + \xi_t D_t dt. \tag{5.27}$$

Eq. (5.27) can be integrated to yield, $\xi_t S_t = E_t(\int_t^T \xi_u D_u du) + E_t \left(\xi_T S_T \right)$. Assuming that $\lim_{T \to \infty} E_t \left(\xi_T S_T \right) = 0$, the asset price satisfies,

$$\xi_t S_t = E_t \left(\int_t^\infty \xi_u D_u du \right). \tag{5.28}$$

Note that, $E_t \left(m_{t,t+h} \right) - 1 = e^{-r_t h} - 1 \approx -r_t h$,[7] whence,

$$\frac{\xi_{t+h} - \xi_t}{\xi_t} \approx -r_t h + \left(m_{t,t+h} - E_t \left(m_{t,t+h} \right) \right). \tag{5.29}$$

In the presence of risk, and risk-averse agents, the innovations to the stochastic discount factor will drive fluctuations of the pricing kernel.

[7] Just set $S_{t+h} + D_{t+h} \equiv 1$ in Eq. (5.26), and interpret S_t as the price of default-free zero coupon bond.

Many of the models in this chapter are cast within a diffusion setting, such that the pricing kernel satisfies the continuous time limit to Eq. (5.29),

$$\frac{d\xi_t}{\xi_t} = -r_t dt - \lambda_t \cdot dW_t, \tag{5.30}$$

where W is a vector Brownian motion, supposed to drive fluctuations of the asset prices, and λ is the vector of unit risk-premiums.

The interpretation of ξ_t in Eq. (5.30) is simple. Without risk, $\xi_\tau = \xi_t e^{-r(\tau-t)}$—the stochastic discount factor is simply the usual discount factor. In the presence of risk, the discount factor varies stochastically, driven by the same sources of variation affecting asset prices, dW_t. Naturally, some components of λ_t are zero if some of these sources of variation do not receive compensation.

5.2.5.2 Expected returns, lambdas and betas

We determine the differential $d(\xi S)$ in Eq. (5.27) by relying on Itô's lemma,

$$d(\xi S) = Sd\xi + \xi dS + d\xi dS = \xi S\left(\frac{d\xi}{\xi} + \frac{dS}{S} + \frac{d\xi}{\xi}\frac{dS}{S}\right),$$

such that Eq. (5.27) can be written as,

$$E_t\left(\frac{dS}{S}\right) + \frac{D}{S}dt = -E_t\left(\frac{d\xi}{\xi}\right) - E_t\left(\frac{d\xi}{\xi}\frac{dS}{S}\right). \tag{5.31}$$

Eq. (5.31) holds for any asset price, including those not distributing dividends and locally riskless, S_0 say, such that $E_t\left(\frac{d\xi}{\xi}\frac{dS_0}{S_0}\right) = 0$, where $\frac{dS_{0t}}{S_{0t}} = r_t dt$ and r_t is the short term rate process. By (5.31), then,

$$E_t\left(\frac{d\xi_t}{\xi_t}\right) = -r_t dt,$$

consistently with the discrete time counterpart of the pricing kernel in Eq. (5.29).

By Eq. (5.31), the expected returns satisfy:

$$E_t\left(\frac{dS}{S}\right) + \frac{D}{S}dt = rdt - E_t\left(\frac{d\xi}{\xi}\frac{dS}{S}\right). \tag{5.32}$$

We know from previous sections that the expectations in Eq. (5.32) can be expressed in terms of partial derivatives, implying that the asset price solves a certain partial differential equation. We will develop this theme in detail in Part II (see Chapters 8 and 9). Now, we wish to further our interpretation of λ in Eq. (5.30) as a vector of unit risk-premiums.

Note that the asset price, S, is obviously driven by the same Brownian motions driving r and ξ, such that $E_t\left(\frac{d\xi}{\xi}\frac{dS}{S}\right) = -\text{Vol}\left(\frac{dS}{S}\right) \cdot \lambda dt$, where $\text{Vol}\left(\frac{dS}{S}\right)$ denotes the instantaneous volatility of asset returns; note that this volatility could be a vector when there are more than a state variables driving the asset price, S. Substituting this result into the R.H.S. of Eq. (5.32) leaves:

$$E_t\left(\frac{dS}{S}\right) + \frac{D}{S}dt = rdt + \underbrace{\text{Vol}\left(\frac{dS}{S}\right)}_{\text{"betas"}} \cdot \underbrace{\lambda}_{\text{"lambdas"}} dt. \tag{5.33}$$

Expected returns equal the short-term rate plus a risk-premium arising due to the randomness of the very same returns. This premium is the product of the instantaneous risks related to the asset price fluctuations (the "betas," $\text{Vol}\left(\frac{dS}{S}\right)$) times the unit risk-premiums that compensate for each individual source of these instantaneous risks (the "lambdas," λ). Eq. (5.33) is an APT relation, the continuous time counterpart to those developed in Chapter 1: the only assumption underlying it is absence of arbitrage, i.e., a positive stochastic discounting factor exists. We now proceed to a decomposition of these expected returns.

5.2.5.3 Risk-adjusted discount rates

How to discount future cash flows in a model with multiple sources of risk? It sounds like there might be an obvious answer to this question: we should use the APT, i.e., Eq. (5.33). It is actually a subtle point. Eq. (5.33) provides predictions of expected returns, but expected returns are not necessarily risk-adjusted discount rates. Naturally, were dividends and asset prices driven by one (and the same) factor, expected returns and risk-adjusted discount rates would be the same. However, the two notions deviate in a multifactor framework.

We illustrate these points while relying on a simplifying assumption, namely that the price-dividend ratio, p say, is independent of the dividends D, and driven by a vector of state variables, y, such that:

$$S\left(D,y\right)=D\cdot p\left(y\right). \tag{5.34}$$

Such a "scale-invariant" property of asset price arises in many economies (see Part II for more detailed discussion). For example, it does if (i) the dividends D are geometric Brownian motions and (ii) the state variables y do not depend on D. In this case, the price in Eq. (5.34) satisfies: $\frac{dS}{S}=\frac{dp}{p}+\frac{dD}{D}$ such that by Eq. (5.32),

$$E_t\left(\frac{dS}{S}\right)+\frac{D}{S}dt=\mathcal{R}dt-E_t\left(\frac{dp}{p}\frac{d\xi}{\xi}\right), \tag{5.35}$$

where $\mathcal{R}$ is defined as follows:

$$\mathcal{R}\equiv r-\frac{E_t\left(\frac{dD}{D}\frac{d\xi}{\xi}\right)}{dt}=r\ +\underbrace{\sigma_0}_{\text{"cash-flow beta"}}\cdot\underbrace{\lambda_{\text{CF}}}_{\text{"cash-flow lambda"}},$$

and $\sigma_0\equiv\text{Vol}\left(\frac{dD}{D}\right)$ (possibly a vector) and λ_{CF} denote the unit-risk premium required to compensate for the randomness in the dividend process. Note that Eq. (5.32) is decomposition: we can always find the appropriate vector λ_{CF} such that Eq. (5.32) holds true.

We refer to $\mathcal{R}$ as the *risk-adjusted discount rates*. They equal the safe interest rate r, plus the premium, $-E_t\left(\frac{dD}{D}\frac{d\xi}{\xi}\right)$, arising to compensate for the stochastic fluctuations of dividends. Eq. (5.35) tells us that if the price-dividend ratio is constant, $E_t\left(\frac{dp}{p}\frac{d\xi}{\xi}\right)=0$, the risk adjusted discount rates are the same as the expected returns, just as in the one-factor Lucas economy discussed in Chapter 3.

If we assume that the price-dividend ratio is constant, and that aggregate dividends D equal consumption, then, σ_0 would be too small to make the expected returns predicted by this model consistent with the data. It is the celebrated equity premium puzzle dealt with in much detail in Part II. But Eq. (5.35) reveals that expected returns could be inflated if the price-dividend

ratio fluctuates, driven by additional state variables requiring compensation:

$$E_t\left(\frac{dp}{p}\frac{d\xi}{\xi}\right) = -\underbrace{\text{Vol}\left(\frac{dp}{p}\right)}_{\text{"price betas"}} \cdot \underbrace{\lambda_p}_{\text{"price lambdas"}} dt. \tag{5.36}$$

This term does indeed represent a wedge between the expected returns and the risk-adjusted discount rates (see Eq. (5.35)); therefore, it carries the potential to mitigate the equity premium puzzle. Note that Eq. (5.36) describes the most natural channel through which expected returns are inflated: the state variables y fluctuate and lead the price-dividend ratio to fluctuate, thereby affecting the realized returns. If these fluctuations require compensation (meaning that their innovations affect the pricing kernel, $\lambda_p \neq 0$), they affect the expected returns. The "price beta" are the asset returns exposure to factors arising through this channel, and the "price lambdas" are the corresponding unit premiums to bear these risks.

Chapter 8 explains that in addition to the equity premium puzzle, time variation in *returns volatility* is a persavive empirical property of asset prices. This property can be rationalized by time variation in risk-adjusted discount rates. Heuristically, note that returns volatility relates to the "price betas" in Eq. (5.36). In a diffusion setting, the price beta relates to the semi-elasticity of the price-dividend ratio with respect to y, $\text{Vol}\left(\frac{dp}{p}\right) = \frac{p'(y)}{p(y)}\text{Vol}(dy)$; that is, the critical ingredient of the "price beta" is the very same price-dividend ratio in Eq. (5.28). Note, also, that the price-dividend ratio can be re-expressed in terms of the risk-adjusted discount rates, as follows:

$$p\left(y_t\right) = \mathbb{E}\left(\int_t^\infty \frac{D_\tau^*}{D_t}\cdot e^{-\int_t^\tau \mathcal{R}(y_u)du}\middle| y_t\right), \quad \frac{D_\tau^*}{D_t} = e^{\left(g_0-\frac{1}{2}\sigma_0^2\right)(\tau-t)+\sigma_0\left(\hat{W}(\tau)-\hat{W}_t\right)}, \tag{5.37}$$

where $\mathbb{E}$ and $\hat{W}_\tau$ denote the expectation and a Brownian motion under the risk-neutral probability, and $g_0 \equiv E_t\left(\frac{dD_t}{D_t}\right)$.

Eq. (5.37) is a present value formula in which a fictitious risk-unadjusted dividend growth, D_τ^*/D_t, is discounted using the risk-adjusted discount rates, $\mathcal{R}\left(y_t\right)$. It is a representation of asset prices with its own interest. It also helps explain results detailed in Chapter 8, by which the sensitivity of the price-dividend ratio to changes in y_t is affected by how $\mathcal{R}(y_t)$ reacts in response to changes in y_t. Intuitively, if the risk-adjusted discount rates $\mathcal{R}(y_t)$ undergo through large and swings in bad times, the volatility predicted by Eq. (5.37) would likely exhibits some of the interesting countercyclical statistics that we see in the data. Chapter 8 explains these properties in detail, as mentioned.

5.2.6 Example: no-arbitrage in Lucas tree

5.2.6.1 Family of prices

Consider the Lucas (1978) model with one tree and one perishable good taken as the numéraire—the continuous time version of the model in Chapter 3. We assume that the dividend is a geometric Brownian motion,

$$\frac{dD_t}{D_t} = gd\tau + \sigma_0 dW_t, \tag{5.38}$$

for two positive constants g and σ_0. We assume no-sunspots, and denote the rational pricing function with $S \equiv S(D)$. By Itô's lemma,

$$dS_t = \underbrace{gD_tS'(D_t) + \frac{1}{2}\sigma_0^2 D_t^2 S''(D_t)}_{\equiv \mu_{St} S_t} dt + \underbrace{\sigma_0 D_t S'(D_t)}_{\equiv \sigma_{St} S_t} dW_t,$$

such that, by Eq. (5.1), wealth satisfies,

$$dV_t = \left(\pi_t \left(\mu_{St} + \tfrac{D_t}{S_t} - r\right) + r_t V_t - c_t\right) dt + \pi_t \sigma_{St} dW_t.$$

Below, we shall show that in the absence of arbitrage, there must be some process λ, the "unit risk-premium", such that,

$$\mu_S + \frac{D}{S} - r = \lambda \sigma_S. \tag{5.39}$$

Let us assume that the short-term rate, r, and the risk-premium, λ, are both constant. Below, we shall show that such an assumption is compatible with a general equilibrium economy. By the definition of μ_S and σ_S, Eq. (5.39) can be written as,

$$0 = \frac{1}{2}\sigma_0^2 D^2 S''(D) + (g - \lambda\sigma_0) DS'(D) - rS(D) + D. \tag{5.40}$$

Eq. (5.40) is a second order differential equation. Its solution, provided it exists, is the rational price of the asset. To solve Eq. (5.40), we initially assume that the solution, S_{F} say, tales the following simple form,

$$S_{\mathrm{F}}(D_t) = K \cdot D_t, \tag{5.41}$$

where K is a constant to be determined. Next, we verify that this is indeed *one* solution to Eq. (5.40). Indeed, if Eq. (5.41) holds, then, by plugging this guess and its derivatives into Eq. (5.40) leaves $K = (r - g + \lambda\sigma_0)^{-1}$, and assuming K is bounded and positive,

$$S_{\mathrm{F}}(D_t) = \frac{1}{r + \lambda\sigma_0 - g} D_t. \tag{5.42}$$

This is a Gordon-type formula. It merely states that prices are risk-adjusted expectations of future expected dividends, where the risk-adjusted discount rate is given by $r + \lambda\sigma_0$. Hence, in a comparative statics sense, stock prices are inversely related to the risk-premium, a quite intuitive conclusion.

Eq. (5.42) can be thought to be the Feynman-Kac representation to Eq. (5.40), viz

$$S_{\mathrm{F}}(D_t) = \mathbb{E}_t\left(\int_t^\infty e^{-r(\tau - t)} D_\tau d\tau\right), \tag{5.43}$$

where $\mathbb{E}_t[\cdot]$ is the conditional expectation taken under the risk neutral probability Q (say), the dividend process follows,

$$\frac{dD_t}{D_t} = (g - \lambda\sigma_0) d\tau + \sigma_0 d\tilde{W}_t,$$

and $\tilde{W}_\tau = W_\tau + \lambda(\tau - t)$ is a another standard Brownian motion defined under Q. Formally, the true probability, P, and the risk-neutral probability, Q, are tied up by the Radon-Nikodym derivative,

$$\eta = \frac{dQ}{dP} = e^{-\lambda(W_\tau - W_t) - \frac{1}{2}\lambda^2(\tau - t)}. \tag{5.44}$$

These pricing results relate to the assumption r and λ are both constant. We didn't specify the exact economic conditions this is true. It is the reason we refer the prices predicted by this model as a "family of prices." The next section provides more structure, through a restriction on preferences that leads to the pricing results summarized so far.

5.2.6.2 Equilibrium with CRRA

How do precisely preferences affect asset prices? In Eq. (5.42), the asset price relates to the interest rate, r, and the risk-premium, λ. But in equilibrium, agents preferences affect r and λ. However, such an impact can have a non-linear pattern. For example, when the risk-aversion is low, a small change of risk-aversion can make the interest rate and the risk-premium change in the same direction. If the risk-aversion is high, the effects may be different, as the interest rate reflects a variety of factors, including precautionary motives.

To illustrate these features within the simple case of CRRA preferences, consider, first, the dynamics of wealth under the risk-neutral probability, Q, such that by Eq. (5.1),

$$dV = (rV - c)\, d\tau + \pi\sigma_S d\tilde{W}. \tag{5.45}$$

We assume that the following transversality condition holds,

$$\lim_{\tau\to\infty} \mathbb{E}_t \left(e^{-r(\tau-t)} V_\tau\right) = 0. \tag{5.46}$$

By integrating Eq. (5.45), and using the previous transversality condition,

$$V_t = \mathbb{E}_t \left(\int_t^\infty e^{-r(\tau-t)} c_\tau d\tau\right). \tag{5.47}$$

Note that Eqs. (5.43) and (5.47) imply that in equilbrium, i.e. $D = c$, we also have that $S = V$.

Next, consider a representative agent with instantaneous utility of consumption $u(c)$ and subjective discount rate ρ, who solves the following intertemporal optimization problem,

$$\max_c E_t \left(\int_t^\infty e^{-\rho(\tau-t)} u\left(c_\tau\right) d\tau\right) \quad \text{s.t. } V_t = E_t \left(\int_t^\infty m_{t,\tau} c_\tau d\tau\right), \tag{P1}$$

where the constraint follows by a change of probability in Eq. (5.47) using Eq. (5.44) and, accordingly,

$$m_{t,\tau} \equiv \frac{\xi_\tau}{\xi_t} = e^{-\left(r+\frac{1}{2}\lambda^2\right)(\tau-t)-\lambda(W_\tau - W_t)}.$$

Consider the Lagrangean

$$\mathcal{L} \equiv E_t \left(\int_t^\infty e^{-\rho(\tau-t)} u(c_\tau) d\tau\right) + \kappa \cdot \left(V_t - E_t \left(\int_t^\infty m_{t,\tau} c_\tau d\tau\right)\right),$$

where κ is a Lagrange multiplier. The first order conditions are $e^{-\rho(\tau-t)} u'\left(c_\tau\right) = \kappa \cdot m_{t,\tau}$, and in equilibrium, $c = D$, such that, using the definition of $m_{t,\tau}$,

$$u'\left(D_\tau\right) = \kappa \cdot e^{-\left(r+\frac{1}{2}\lambda^2-\rho\right)(\tau-t)-\lambda(W_\tau - W_t)}. \tag{5.48}$$

That is, by Itô's lemma,

$$\begin{aligned} du'(D_\tau) &= \left(u''(D_\tau) D_\tau g + \frac{1}{2}\sigma_0^2 D^2 u'''(D_\tau)\right) d\tau + u''(D_\tau) D_\tau \sigma_0 dW_\tau \\ &= u'(D_\tau)\left((\rho - r)\, d\tau - \lambda dW_\tau\right), \end{aligned}$$

where the first equality follows by expanding the L.H.S. and the second equality follows by expanding the R.H.S. of Eq. (9A.21). But drifts and volatilities of the first and the second equality have to be the same, whence

$$r = \rho - \frac{u''(D) D}{u'(D)} g - \frac{1}{2}\sigma_0^2 D^2 \frac{u'''(D)}{u'(D)} \quad \text{and} \quad \lambda = -\frac{u''(D) D}{u'(D)} \sigma_0.$$

If λ is constant (our assumption), we can reverse-engineer the utility functions that are consistent with this assumption. We integrate the second of the previous equations, and find that $u(D) = \frac{D^{1-\gamma}-1}{1-\gamma}$, where $\gamma \equiv \frac{\lambda}{\sigma_0}$ is the CRRA. Therefore, under CRRA preferences,

$$r = \rho + \gamma g - \frac{1}{2}\gamma(\gamma+1)\sigma_0^2.$$

We can, then, replace the expressions for the short-term rate and the risk-premium into Eq. (5.42), leaving then the following special case of Eq. (5.41):

$$S(D_\tau) = \frac{1}{\rho - (1-\gamma)\left(g - \frac{1}{2}\gamma\sigma_0^2\right)} D_\tau. \tag{5.49}$$

Finally, we check that the transversality condition (5.46) holds in equilibrium, $S = V$. Under the same conditions that ensure $S(D)$ is finite, we have that

$$\begin{aligned} \lim_{\tau\to\infty} \mathbb{E}_t\left(e^{-r(\tau-t)} V_\tau\right) &= \lim_{\tau\to\infty} \mathbb{E}_t\left(e^{-r(\tau-t)} S_\tau\right) \\ &= \lim_{\tau\to\infty} E_t\left(m_{t,\tau} S_\tau\right) \\ &= \lim_{\tau\to\infty} E_t\left(e^{-\left(r+\frac{1}{2}\lambda^2\right)(\tau-t) - \lambda(W_\tau - W_t)} S_\tau\right) \\ &= S_t \lim_{\tau\to\infty} E_t\left(e^{\left(g-\frac{1}{2}\sigma_0^2 - r - \frac{1}{2}\lambda^2\right)(\tau-t) + (\sigma_0 - \lambda)(W_\tau - W_t)}\right) \\ &= S_t \lim_{\tau\to\infty} e^{-(r-g+\sigma_0\lambda)(\tau-t)} \\ &= S_t \lim_{\tau\to\infty} e^{-\left(\rho - (1-\gamma)\left(g - \frac{1}{2}\gamma\sigma_0^2\right)\right)(\tau-t)} \\ &= 0. \end{aligned} \tag{5.50}$$

5.2.6.3 Bubbles

We may refer the transversality condition in Eq. (5.46) to as a *no-bubble condition*. To illustrate, note that Eq. (5.40) has an infinite number of solutions that take the following form

$$S(D_\tau) = S_{\mathrm{F}}(D_\tau) + A D_\tau^\delta, \quad A, \delta \text{ constants}. \tag{5.51}$$

Indeed, by plugging Eq. (5.51) into Eq. (5.40) shows that Eq. (5.51) holds if and only if

$$0 = \delta\left(g - \lambda\sigma_0\right) + \frac{1}{2}\delta\left(\delta - 1\right)\sigma_0^2 - r. \tag{5.52}$$

This quadratic equation in δ has two solutions that satisfy $\delta_1 < 0$ and $\delta_2 > 0$. Therefore, the asset price function does actually take the following form:

$$S\left(D_\tau\right) = S_{\mathrm{F}}\left(D_\tau\right) + A_1 D_\tau^{\delta_1} + A_2 D_\tau^{\delta_2},$$

where $\lim_{D\to 0}\left|S\left(D\right)\right| = \infty$ if $A_1 \neq 0$. We set $A_1 = 0$ to rule out that the price explodes as the dividend gets small, leaving

$$S\left(D_\tau\right) = S_{\mathrm{F}}\left(D_\tau\right) + \mathcal{B}\left(D_\tau\right), \tag{5.53}$$

where $\mathcal{B}\left(D\right) \equiv A_2 D^{\delta_2}$.

The component $S_{\mathrm{F}}\left(D\right)$ is the risk-adjusted present value of the expected dividends (see Eq. (5.43)) and, hence, is the asset *fundamental value*. The second component is, then, a bubble, that is, the difference between the asset market value $S\left(D\right)$ and the fundamental value $S_{\mathrm{F}}\left(D\right)$. We seek conditions under which $S\left(D\right)$ in Eq. (5.53) satisfies the transversality condition in Eq. (5.46). We have

$$\lim_{\tau\to\infty}\mathbb{E}_t\left(e^{-r(\tau-t)}S_\tau\right) = \lim_{\tau\to\infty}\mathbb{E}_t\left(e^{-r(\tau-t)}S_{\mathrm{F}}\left(D_\tau\right)\right) + \lim_{\tau\to\infty}\mathbb{E}_t\left(e^{-r(\tau-t)}\mathcal{B}\left(D_\tau\right)\right).$$

By Eq. (5.50), the fundamental value of the asset satisfies the transversality condition, under the condition the denominator in Eq. (5.49) is strictly positive. Regarding the bubble, we have

$$\begin{aligned}\lim_{\tau\to\infty}\mathbb{E}_t\left(e^{-r(\tau-t)}\mathcal{B}\left(D_\tau\right)\right) &= A_2 \cdot \lim_{\tau\to\infty}\mathbb{E}_t\left(e^{-r(\tau-t)}D_\tau^{\delta_2}\right)\\ &= A_2 \cdot D_t^{\delta_2} \cdot \lim_{\tau\to\infty} e^{(\delta_2(g-\lambda\sigma_0)+\frac{1}{2}\delta_2(\delta_2-1)\sigma_0^2-r)(\tau-t)}\\ &= A_2 \cdot D_t^{\delta_2},\end{aligned}$$

where the last line holds as δ_2 satisfies Eq. (5.52). That is, the bubble cannot satisfy the transversality condition and we set $A_2 = 0$: in this economy, the transversality condition in Eq. (5.46) holds if and only if there are no bubbles.

5.2.6.4 Reflecting barriers and absence of arbitrage

Next, suppose that insofar as the dividend D_τ fluctuates above a certain level $\underline{D} > 0$, everything goes as in the previous section but that, as soon as the dividends level hits a "barrier" $\underline{D}$, it is "reflected" back with probability one. In this case, we say that the dividend follows a process with *reflecting barriers*. Alternatively, one may think that the firm has "infinitely deep pockets," in that it can always guarantee that the dividends it distributes are solution to,

$$\frac{dD_\tau}{D_\tau} = gd\tau + \sigma_0 dW_\tau + dU_\tau, \tag{5.54}$$

where U_τ is a continuous, non-negative process that increases only when the dividends D hit $\underline{D}$. We may think of the firm as operating on an "infinitesimal intervention" scale.

Eq. (5.54) thus generalizes Eq. (5.38), in that dividends cannot lower below a threshold. How does the price behave in this context? By Itô's lemma, and Eq. (5.54),

$$\begin{aligned} dS_\tau &= S'\left(D_\tau\right) dD_\tau + \frac{1}{2}\sigma_0^2 D_\tau^2 S''\left(D_\tau\right) d\tau \\ &= \left(S'\left(D_\tau\right) gD_\tau + \frac{1}{2}\sigma_0^2 D_\tau^2 S''\left(D_\tau\right)\right) d\tau + S'\left(D_\tau\right)\sigma_0 D_\tau dW_\tau + S'(D_\tau) dU_\tau. \end{aligned} \tag{5.55}$$

We claim that to ensure absence of arbitrage, the following *smooth pasting* condition must hold,

$$S'(\underline{D}) = 0. \tag{5.56}$$

Indeed, after hitting the barrier, $\underline{D}$, the dividend is reflected back for the part exceeding $\underline{D}$. Since the reflection takes place with probability one, the asset is locally riskless at the barrier $\underline{D}$. Therefore, absence of arbitrage requires the price moves at $\underline{D}$, only by its predictable drift component in Eq. (5.55), which it does when $S'\left(D_\tau\right) = 0$ for $D_\tau = \underline{D}$. Note then that the last component in Eq. (5.55), $S'(\underline{D})dU_\tau = 0$ for all τ and that by standard arguments (i.e. Eq. (5.39)),

$$\mu_{S,\tau} + \frac{D_\tau}{S_\tau} - r = \lambda\sigma_S\left(D_\tau\right) \equiv \lambda\frac{\sigma_0 D_\tau S'\left(D_\tau\right)}{S\left(D_\tau\right)},$$

where $\mu_{S,t} \equiv \frac{d}{d\tau}E_t\left(\frac{S_\tau}{S_t}\right)\Big|_{\tau=t}$. By Eq. (5.56), we have that at $D_t = \underline{D}$,

$$\mu_S + \frac{\underline{D}}{S} = r.$$

In words, suppose that we hold the asset during the infinitesimal period in which "reflection" takes place. During this infinitesimal amount of time, the asset is locally riskless (due to reflection) and, thus, cannot earn more or less than the riskless rate. Indeed, by the previous relation, and by imposing that $\sigma_S(\underline{D}) = 0$, we have that wealth satisfies (see Eq. (5.1)),

$$dV = \left(\pi\left(\mu_S + \tfrac{\underline{D}}{\underline{S}} - r\right) + rV - c\right) d\tau + \pi\sigma_S(\underline{D})dW = \left(rV - c\right) d\tau.$$

This example illustrates how arbitrage restrictions operate through Eq. (5.39).

To solve the model, note that while dividends fluctuate above the barrier, $D_\tau > \underline{D}$, the price is still as in Eq. (5.51),

$$S(D_\tau) = \frac{D_\tau}{r - g + \lambda\sigma_0} + A_1 D_\tau^{\delta_1} + A_2 D_\tau^{\delta_2}.$$

We set $A_2 = 0$, meaning that the effects of a lower barrier $\underline{D}$ on the dividend dynamics are not perceptible when dividends are large. To determine A_1, we pin down the price at the barrier $\underline{D}$. We have: $K \equiv S_{\mathrm{F}}\left(D\right)/\,D$, and

$$0 = S'\left(\underline{D}\right) = K + \delta_1 A_1 \underline{D}^{\delta_1 - 1}, \quad \underline{S} \equiv S\left(\underline{D}\right) = K\underline{D} + A_1\underline{D}^{\delta_1},$$

where the second condition is the *value matching condition*, which we impose to ensure continuity of the price with respect to D and, hence, absence of arbitrage. The solution to the previous system is

$$\underline{S} = \frac{1 - \delta_1}{-\delta_1} K\underline{D} \quad \text{and} \quad A_1 = \frac{K}{-\delta_1}\underline{D}^{1-\delta_1}. \tag{5.57}$$

The price is an increasing and convex function of the fundamentals, D.[8] Convexity reflects the "option-like" feature of the dividends barrier, $\underline{D}$. Note, also, that the model takes $\underline{D}$ as given. Section 5.6 presents a model where distribution of dividends is endogenous, that is, dividends are "controlled" as part of an optimization problem.

5.3 Distorsions and numéraires

Consider a framework where given a strictly positive random variable x_T, we define a Radon-Nikodym derivative as,

$$\left.\frac{dP_2}{dP_1}\right|_{\mathbb{F}_T} = \rho_T(T) = \frac{x_T}{E_t(x_T)},$$

where $\mathbb{F}_T$ denotes the information set at time T, and $E_t(\cdot)$ is the expectation under P_1, conditional on the information set at time t. Define a corresponding *density process* for $\rho_T(\tau)$ as,

$$\rho_T(\tau) = \frac{E_\tau(x_T)}{E_t(x_T)}, \quad \tau \in [t,T].$$

We obviously have that $E_\tau[\rho_T(T)] = \rho_T(\tau)$ for all τ, and in particular, $\rho_T(t) = 1$. Therefore, $\rho_T(\cdot)$ is a martingale starting from 1, with $\rho_T(t) = E_t[\rho_T(T)] = 1$, and P_2 is thus a probability.

We now develop a series of examples to illustrate how naturally changes of probability arise in finance. A leading example is one where P_1 is the physical probability, and the random variable x_T is the marginal utility of consumption at T, $x_T = u'(c_T)$, as in (the continuous time version of) the consumption-based asset pricing models of Chapters 2 and 3. Another fundamental example spans many fixed income security pricing problems described in Chapter 13. In these problems, x_T is the inverse of the money market account at T, $x_T = e^{-\int_t^T r_u du}$, and P_1 is the risk-neutral probability, such that

$$\rho_T(\tau) = \frac{\mathbb{E}_\tau\left(e^{-\int_t^T r_u du}\right)}{P_t(T)} = \frac{e^{-\int_t^\tau r_u du} P_\tau(T)}{P_t(T)},$$

where $P_t(T)$ is the time t price of a zero coupon bond expiring at T.

The next section develops the leading example of change of probabilities, the consumption-based probabilities, as a benchmark. Section 5.4.2 reviews general notions, and examples, of pricing, based on the numéraires being involved into the definition of a given asset pricing problems.

5.3.1 Leading example: consumption-based probabilities

Consider a basic consumption-based model, where an investor can invest at two dates only, at t and at $T > t$, without any intermediate consumption, such that the price of an asset satisfies,

$$S_t = E_t\left(\frac{e^{-\rho(T-t)}u'(c_T)}{u'(c_t)} S_T\right) = e^{-r(T-t)}\mathbb{E}_t(S_T),$$

[8] This model takes the barrier $\underline{D}$ as given. However, one may be interested in "controlling" the dividend D_τ in a way that as soon as the price S_τ hits an exogenous given level $\underline{S}$, the dividend is activated to induce the price to increase back. It is straightforward to see that the solution for the dividend level that triggers this "infinitesimal" intervention is obtained by inverting the first equation in (5.57) for the dividend threshold, viz $\underline{D} = \frac{-\delta_1}{1-\delta_1} K^{-1}\underline{S}$.

with the usual notation. In this model, we define the risk-neutral probability Q through the Radon-Nikodym derivative of Q against P,[9]

$$\left.\frac{dQ}{dP}\right|_{\mathbb{F}_T} = \underbrace{\frac{e^{-\rho(T-t)}E_t\left[u'\left(c_T\right)\right]}{u'\left(c_t\right)}e^{r(T-t)}}_{=1} \cdot \frac{u'\left(c_T\right)}{E_t\left[u'\left(c_T\right)\right]}.$$

Assuming decreasing marginal utility as usual, we have that the risk-neutral probability distorts the physical, by assigning more weight to the bad states of nature, i.e. those where consumption is low, as explained in Chapter 2.

We derive implications for a continuous-time model where a representative agent has CRRA equal to γ. In equilibrium, $c_T = S_T$, such that if consumption is a geometric Brownian motion with parameters (μ, σ), we have that the density process,

$$\rho_T\left(\tau\right) = \frac{E_\tau\left(S_T^{-\gamma}\right)}{E_t\left(S_T^{-\gamma}\right)} = \left(\frac{S_\tau}{S_t}\right)^{-\gamma} e^{\left(\gamma\mu - \frac{1}{2}\sigma^2\gamma(\gamma+1)\right)(\tau-t)},$$

such that,

$$\frac{d\rho_T\left(\tau\right)}{\rho_T\left(\tau\right)} = -\gamma\sigma dW_t,$$

where W_t is a Brownian motion under P. By Girsanov's theorem, then, we have that $\tilde{W}_t = W_t + (\gamma\sigma)\,t$ is a Brownian motion under Q, such that the stock price is solution to,

$$\frac{dS_t}{S_t} = \underbrace{\left(\mu - \gamma\sigma^2\right)}_{=r}dt + \sigma d\tilde{W}_t.$$

To reconcile prices in the risk-neutral world with prices in the true, we need the average path of the stock price to be lower under the risk-neutral probability than under the physical. Since the stock is traded, we also have that $\mu - \gamma\sigma^2 = r$, as indicated—the risk-premium is somehow "hidden" in this example with complete markets. When markets are incomplete, risk-premiums terms would, instead, show up, as in the Gamma process example of Heston (1993,b).

[Develop Gamma processes here.]

5.3.2 Numéraire pricing

5.3.2.1 Definition

The formulation of the fundamental theorems of asset pricing (FTAP) in the previous chapters relies on the risk-neutral probability—the probability under which the asset prices discounted by the value of the money market account are martingales, to prevent arbitrage. That is, the money market account is the numéraire in this no-arbitrage context. We can formulate equivalent versions of this theorem, utilizing different numéraires, and different equivalent probabilities. Consider the following definition, which we shall state assuming no dividends for simplicity.

[9] This derivation relies on the assumption that the short-term rate is constant. Section 7.5.1 in Chapter 7 contains a derivation of the more general case of stochastic interest rates in a Markov setting.

DEFINITION 5.1 (Numéraires and equivalent probabilities). *A* numéraire *is any asset with a price process $\mathcal{N}_t > 0$, say. Given a numéraire, a probability $Q^{\mathcal{N}}$ is an* equivalent probability *(or* measure*) if any asset price process S_t normalized by $\mathcal{N}_\tau$ satisfies,*

$$\frac{S_t}{\mathcal{N}_t} = E_t^{Q^{\mathcal{N}}}\left(\frac{S_T}{\mathcal{N}_T}\right).$$

A critical task in this and previous chapters is the formulation of the FTAP—markets are arbitrage-free if and only if there exists an equivalent martingale probability. We now provide examples of existence of numéraires and associated probabilities $Q^{\mathcal{N}}$.

5.3.2.2 Examples of numéraires and arbitrage pricing laws

Risk-neutral probability

The risk-neutral probability corresponds to the money market account numéraire, as mentioned, with value process given by,

$$\mathcal{N}_\tau^r \equiv e^{\int_0^\tau r_u du}.$$

Suppose, for example, that the short-term rate is constant. In terms of the consumption probabilities of Section 5.1, we have that,

$$\frac{S_t}{\mathcal{N}_t^r} = \mathbb{E}_t\left(\frac{S_T}{\mathcal{N}_T^r}\right), \quad \mathcal{N}_t^r = e^{rt}. \tag{5.58}$$

Alternatively, and still assuming interest rates are constant, we may express Eq. (5.58) as,

$$\frac{S_t}{\mathcal{N}_t^P} = \mathbb{E}_t\left(\frac{S_T}{\mathcal{N}_T^P}\right), \quad \mathcal{N}_t^P = e^{-r(T-t)} = P_t(T),$$

with the obvious condition that $\mathcal{N}_T^P = P_T(T) = 1$. If interest rates are constant, the money market account yields the same as an investment of $\frac{1}{P_t(T)}$ units of a zero coupon bond, such that we can we can re-interpret the zero-coupon bond as the numéraire. Next, we generalize this context to one with stochastic interest rates.

Forward probability

Suppose, instead, that interest rates are random, such that

$$S_t = \mathbb{E}_t\left(e^{-\int_t^T r_u du} S_T\right) = P_t(T)\cdot \mathbb{E}_t\left(\frac{e^{-\int_t^T r_u du}}{P_t(T)} S_T\right) = P_t(T)\cdot \mathbb{E}_t^{Q^{\mathcal{N}_P}}(S_T), \tag{5.59}$$

where $Q^{\mathcal{N}_P}$ is the so-called *forward probability*, defined through the Radon-Nikodym derivative against Q, as follows,

$$\left.\frac{dQ^{\mathcal{N}_P}}{dQ}\right|_{\mathbb{F}_T} = \frac{e^{-\int_t^T r_u du}}{P_t(T)}. \tag{5.60}$$

Accordingly, in this context of random interest rates, we can define the zero coupon bond as a numéraire, such that, by Eq. (5.59),

$$\frac{S_t}{\mathcal{N}_t^P} = \mathbb{E}_t^{Q^{\mathcal{N}_P}}\left(\frac{S_T}{\mathcal{N}_T^P}\right), \quad \mathcal{N}_t^P = P_t(T), \tag{5.61}$$

where the equality follows by the boundary condition for the price of a zero coupon bond, $\mathcal{N}_T^P = 1$.

In Chapter 13, we shall make reference to this probability to price interest rate derivatives without risk of default. The price of these derivatives is,

$$C_t = \mathbb{E}_t \left(e^{-\int_t^T r_u du} \Psi \left((r_s)_{s\in[t,T]} \right) \right), \tag{5.62}$$

where $\Psi(\cdot)$ is the payoff, possibly a function(al) of the entire path of the short-term rate over the life of the derivative. Note the main difficulty in Eq. (5.62). The discounting factor $e^{-\int_t^T r_u du}$ obviously correlates with the payoff, and complicates the evaluation of this derivative. However, we can use the forward probability to express C_t, as follows,

$$C_t = P_t(T) \cdot \mathbb{E}_t^{Q^{\mathcal{N}_P}} \left(\Psi \left(\{r_s\}_{s\in[t,T]} \right) \right). \tag{5.63}$$

As the Radon-Nikodym derivative defined in Eq. (5.60) makes clear, the new probability, $Q^{\mathcal{N}_P}$, distorts Q, in that it assigns a higher weight to events where interest rates paths are lower. The drift of r under $Q^{\mathcal{N}_P}$ is indeed lower, as formally shown in Appendix 3 of Chapter 13.[10] In other words, we have that upon using Eq. (5.63) instead of Eq. (5.62), we get rid off the randomness of the discounting factor $e^{-\int_t^T r_u du}$, by multiplying the expected payoff by $P_t(T)$. At the same time, this expected payoff is calculated in world ruled by a probability $Q^{\mathcal{N}_P}$, where interest rates are on average lower than under Q, which compensates for the fact the very same expected payoff receives an "haircut" of $P_t(T) < 1$.

Note that the calculations underlying Eq. (5.61) can be used to check the internal consistency of the definition of the new numéraire and equivalent probabilities with absence of arbitrage. There are no arbitrage opportunities if there exists a probability $Q^{\mathcal{N}_P}$ with Radon-Nikodym derivative against Q equal to:

$$\left. \frac{dQ^{\mathcal{N}_P}}{dQ} \right|_{\mathbb{F}_T} = \frac{\mathcal{N}_t^r}{\mathcal{N}_T^r} \cdot \frac{\mathcal{N}_T^P}{\mathcal{N}_t^P}. \tag{5.64}$$

Indeed, we have, simply, that

$$\frac{S_t}{\mathcal{N}_t^P} = \mathbb{E}_t^{Q^{\mathcal{N}_P}} \left(\frac{S_T}{\mathcal{N}_T^P} \right) = \mathbb{E}_t \left(\frac{S_T}{\mathcal{N}_T^P} \frac{\mathcal{N}_t^r}{\mathcal{N}_T^r} \cdot \frac{\mathcal{N}_T^P}{\mathcal{N}_t^P} \right),$$

which is the first equality in Eq. (5.59).

Annuity probability

Consider, now, a class of contracts, where the payoff Π_T is given by difference between a constant K and some risk of interest, which is not necessarily traded, re-scaled by a variable measurable at time T,

$$\Pi_T \equiv \mathcal{N}_T^A (R_T - K). \tag{5.65}$$

The interpretation of Π_T in this section is that of the value of an interest rate swap at time T, whereby a counterparty A shall pay a fixed interest rate K to counterpart B, in exchange of the variable LIBOR, over a certain horizon, called the tenor. In Eq. (5.65), R_T is the so-called

[10] Note that this statement certainly holds when the short-term rate is a Markov process. In more complex models such as those with stochastic volatility, the bond price sensitivity to movements in the state variables may switch sign and render this statement model-dependent.

swap rate, and $\mathcal{N}_T^A$ is the value at time T of an annuity of one dollar over the tenor of the swap. Swaps contracts are amongst the most important interest rate derivatives dealt with in practice, and are motivated and explained in detail in Chapter 13.

The annuity factor, $\mathcal{N}_T^A$, is a valid numéraire in this market. Accordingly, let us define the *annuity probability* arising in this context, through the counterpart to the Radon-Nikodym derivative in Eq. (5.64),

$$\left.\frac{dQ^{\mathcal{N}_A}}{dQ}\right|_{\mathbb{F}_T} = \frac{\mathcal{N}_t^r}{\mathcal{N}_T^r} \cdot \frac{\mathcal{N}_T^A}{\mathcal{N}_t^A},$$

such that, we can price all contracts in this market.[11] First, the value at t of a forward starting interest rate swap, i.e. a swap agreed at time t, say Π_t, is given by:

$$\frac{\Pi_t}{\mathcal{N}_t^A} = \mathbb{E}_t^{Q^{\mathcal{N}_A}} \left(\frac{\Pi_T}{\mathcal{N}_T^A} \right),$$

and that of a swaption, i.e. an option to enter into a future swap, is Π_t^+, where:

$$\frac{\Pi_t^+}{\mathcal{N}_t^A} = \mathbb{E}_t^{Q^{\mathcal{N}_A}} \left(\frac{\max\{\Pi_T, 0\}}{\mathcal{N}_T^A} \right) = \mathbb{E}_t^{Q^{\mathcal{N}_A}} \left(R_T - K \right)^+.$$

Note that forward swap rate, defined as the value of K determined at time t, such that Π_t is zero, satisfies,

$$K = \mathbb{E}_t^{Q^{\mathcal{N}_A}} \left(R_T \right).$$

Accordingly, we can define the forward swap rate as the process $R_\tau = \mathbb{E}_\tau^{Q^{\mathcal{N}_A}} \left(R_T \right)$, for $\tau \leq T$, which is obviously a martingale under $Q^{\mathcal{N}_A}$. These properties are extremely useful, when it comes to pricing these important derivatives.

Defaultable probability

Finally, we consider the so-called *survival-contingent probability*, studied in Chapter 14, where the numéraire, $\mathcal{N}_T^{A^D}$, is a defaultable annuity of one dollar, paid off over a certain tenor, which is interpreted as the period over which two counterparties exchange credit risk. The option to enter into this deal is called a credit default swaption, and the payoff is as in Eq. (5.65), with $\Pi_T^+ \equiv \mathcal{N}_T^{A^D} \left(\text{cds}_T - K \right)^+$, where cds_T is the credit default risk premium as of time T, and K is a constant.

5.3.2.3 Martingales and numéraires

The examples in the previous section can be gathered to provide a general insight. Consider a forward starting agreement, originated at time t, with payoff equal to,

$$\mathcal{N}_T \left(X_T - K \right),$$

where $\mathcal{N}_T$ is a numéraire, $Q^{\mathcal{N}}$ is an equivalent probability, and X_T is measurable at time T. Then, the "forward risk," defined as

$$X_\tau = \mathbb{E}_\tau^{Q^{\mathcal{N}}} \left(X_T \right),$$

is a martingale under $Q^{\mathcal{N}}$, and $K = X_t$ is the value that sets the value of the forward starting agreement equal to zero.

[11] Note that under this annuity probability, the events that matter most, are those where interest rates paths are low, similarly as with the forward probability in the previous section.

5.3.2.4 Marking-to-market updates of forward starting agreements

Suppose we are long an amount θ_t of a forward starting agreement at time t and strike K_t such that the value of this agreement at t right after the trade is $\pi_{t^+} \equiv \theta_t \mathcal{N}_t (X_t - K_t)$ and the value at $t+1$ before the trade is $\pi_{t+1} \equiv \theta_t \mathcal{N}_{t+1} (X_{t+1} - K_t)$. Choosing the strike K_t that clears π_{t^+} delivers the forward risk, $K_t = X_t$, such that, by aggregating,

$$\begin{aligned}\sum_{t=1}^{n} \pi_{t+1} &= \sum_{t=1}^{n} \theta_t \mathcal{N}_{t+1} (X_{t+1} - X_t) \\ &= \sum_{t=1}^{n} \theta_t \left[\mathcal{N}_t (X_{t+1} - X_t) + (\mathcal{N}_{t+1} - \mathcal{N}_t) (X_{t+1} - X_t) \right].\end{aligned}$$

By chopping the trading interval in small pieces we have that under standard regularity conditions,

$$\pi_T - \pi_0 = \int_0^T \theta_t \mathcal{N}_t dX_t + \int_0^T \theta_t d\mathcal{N}_t dX_t.$$

5.4 Martingales and arbitrage

5.4.1 The information framework

We still consider a Lucas' type economy, but with a finite horizon $T < \infty$. Consider a probability space $(\Omega, \mathcal{F}, P)$, let W be a standard d-dimensional Brownian motion, and let $\mathbb{F} = (\mathbb{F}_t)_{t \in [0,T]}$ be the P-augmentation of the natural filtration $\mathbb{F}_\tau^W = \sigma (W_s,\ s \leq \tau)$ generated by W, with $\mathbb{F} = \mathbb{F}_T$. We assume that prices, interest rates and dividends are all adapted to $\mathbb{F}$.

We consider a money market account, m trees and, sometimes, additional assets in zero net supply—"inside money." Let $S_{0\tau} = e^{\int_t^\tau r_u du}$ denote the value of a unit money market account, where r_τ is the short-term rate. The trees entitle to receive dividends $D_{i\tau}$, $i = 1, \cdots, m$, the numéraire, and the price of each tree at time-τ, denoted with $S_{i\tau}$, is taken to satisfy

$$\frac{dS_{i\tau}}{S_{i\tau}} = \bar{\mu}_{i\tau} dt + \sigma_{i\tau} dW_\tau, \quad i = 1, \cdots, m, \tag{5.66}$$

where $\bar{\mu}_{i\tau}$ and $\sigma_{i\tau}$ satisfy the same regularity conditions as r_τ,[12] and $\sigma_{i\tau} \in \mathbb{R}^d$. We let $\sigma_\tau \equiv [\sigma_{1\tau}, \cdots, \sigma_{m\tau}]^\top$ and assume that rank $(\sigma_\tau) = m \leq d$ for each τ. All assets are exchanged without frictions.

We assume that the dividend D_i is solution to

$$\frac{dD_{i\tau}}{D_{i\tau}} = g_{i\tau} d\tau + \sigma_{0i,\tau} dW_\tau,$$

where $g_{i\tau}$ and $\sigma_{0i,\tau}$ are $\mathbb{F}_\tau$-adapted, with $\sigma_{0i} \in \mathbb{R}^d$.

The value of an investment strategy net of dividends is $V_\tau \equiv \sum_{i=0}^m \theta_{i\tau} S_{i\tau}$, where $\theta_{i\tau}$ denotes the number of assets i invested at time-τ. The vector of all such θ_i is taken to be a predictable process in $\mathbb{R}^{m+1}$, satisfying standard regularity conditions.[13] Similarly as in Section 5.4.1, we

[12] Namely, $E_t(\int_t^T r_\tau du) < \infty$.

[13] Namely, it is required that this vector θ (say) satisfies $E_t(\int_t^T \|\theta_\tau\|^2 d\tau) < \infty$ to rule out arbitrage through "doubling strategies"—doubling the bets until we win.

say that a strategy is self-financing if its value V is the solution to

$$dV_\tau = \left(\pi_\tau^\top (\mu_\tau - \mathbf{1}_m r_\tau) + r_\tau V_\tau - c_\tau\right) d\tau + \pi_\tau^\top \sigma_\tau dW_\tau, \tag{5.67}$$

where $\mathbf{1}_m$ is a m-dimensional vector of ones, $\pi \equiv [\pi_1, \cdots, \pi_m]^\top$ with generic element $\pi_i \equiv \theta_i S$, and μ is a m-dimensional vector with generic element $\bar{\mu}_i + \frac{D_i}{S_i}$. The solution to Eq. (5.67) is

$$\frac{V_\tau}{S_{0\tau}} = V_t - \int_t^\tau \frac{c_u}{S_{0u}} du + \int_t^\tau \frac{\pi_u^\top (\mu_u - \mathbf{1}_m r_u)}{S_{0u}} du + \int_t^\tau \frac{\pi_u^\top \sigma_u}{S_{0u}} dW_u. \tag{5.68}$$

We require V to be strictly positive.

5.4.2 Viability

Let $\bar{g}_i = \frac{S_i}{S_0} + \bar{z}_i$, where $d\bar{z}_i = \frac{1}{S_0} dz_i$ and $z_{i\tau} = \int_t^\tau D_{iu} du$. We can generalize the definition of risk-neutral probability in Eq. (5.44) by introducing a set $\mathcal{Q}$ of *risk-neutral probabilities* $\mathcal{Q}$ under which $\bar{g}_i$ is a martingale. We can show the equivalent of Theorem 2.8 in Chapter 2, that is, the set $\mathcal{Q}$ is not empty if and only if the market is arbitrage-free. We rely on Girsanov's theorem of Section 5.3.3. Given a $\mathbb{F}_t$-adapted process λ_t, define, for each $\tau \in [t, T]$,

$$W_{0t} = W_t + \int_t^\tau \lambda_u du, \tag{5.69}$$

a standard Brownian motion under a probability Q equivalent to P, with Radon-Nikodym derivative equal to

$$\eta_T^0 \equiv \left.\frac{dQ}{dP}\right|_{\mathbb{F}_T} = \exp\left(-\frac{1}{2}\int_t^T \|\lambda_\tau\|^2 d\tau - \int_t^T \lambda_\tau^\top dW_\tau\right), \tag{5.70}$$

such that η_τ^0 is a martingale under P. Under Q, Eq. (5.66) is

$$d\bar{g}_{i\tau} = d\left(\frac{S_{i\tau}}{S_{0\tau}}\right) + d\bar{z}_{i\tau} = \frac{S_{i\tau}}{S_{0\tau}} \left((\mu_{i\tau} - \sigma_{i\tau}\lambda_\tau - r_\tau) d\tau + \sigma_{i\tau} dW_{0\tau}\right).$$

But if $\bar{g}_i$ is a martingale for each i under $Q \in \mathcal{Q}$, then

$$\mu_\tau - \mathbf{1}_m r_\tau = \sigma_\tau \lambda_\tau. \tag{5.71}$$

Therefore, by Eqs. (5.67), (5.69) and (5.71), we have that, for $\tau \in [t, T]$,

$$\frac{V_\tau}{S_{0\tau}} = V_t - \int_t^\tau \frac{c_u}{S_{0u}} du + \int_t^\tau \frac{\pi_u^\top \sigma_u}{S_{0u}} dW_{0u}. \tag{5.72}$$

Consider the following definition:

DEFINITION 5.2 (Arbitrage opportunity). *A portfolio π is an arbitrage opportunity if for any zero-consumption decision, $c_\tau \equiv 0$, one has that $V_t \leq S_{0T}^{-1} V_T$ and* $\Pr\left(S_{0T}^{-1} V_T - V_t > 0\right) > 0$.

We have:

THEOREM 5.3. *There are no arbitrage opportunities if and only if the set of risk-neutral probabilities $\mathcal{Q}$ is not empty.*

A proof of this fundamental result is in the Appendix. Note that the if part follows easily, by Eq. (5.72). The only if part is more elaborated, and it can be outlined as follows. We want to claim that absence of arbitrage implies that $\mathcal{Q}$ is no-empty or, equivalently (by Girsanov's theorem), that there exists λ satisfying Eq. (5.71). Now, if Eq. (5.71) didn't hold, we could implement an arbitrage, as follows. We could find a nonzero $\underline{\pi} : \underline{\pi}^\top \sigma = 0$ and $\underline{\pi}^\top (\mu - \mathbf{1}_m r) \neq 0$. Then, we could use $\underline{\pi}$ when $\mu - \mathbf{1}_m r > 0$ and $-\underline{\pi}$ when $\mu - \mathbf{1}_m r < 0$, therebying obtaining an appreciation rate of V larger than r in spite of having zeroed uncertainty through $\underline{\pi}^\top \sigma = 0$. Eq. (5.71) rules out this arbitrage, because we have that for each π, $\pi^\top (\mu - \mathbf{1}_m r) = \pi^\top \sigma \lambda$. More precisely, define

$$\langle \sigma^\top \rangle^\perp \equiv \left\{ x \in L^2_{t,T,m} : \sigma^\top x = \mathbf{0}_d \right\} \quad \text{and} \quad \langle \sigma \rangle \equiv \left\{ z \in L^2_{t,T,m} : z = \sigma u,\ u \in L^2_{t,T,d} \right\}.$$

Then, we may formalize the previous reasoning as follows. The excess return vector, $\mu - \mathbf{1}_m r$, must be orthogonal to all vectors in $\langle \sigma^\top \rangle^\perp$,[14] and since $\langle \sigma \rangle$ and $\langle \sigma^\top \rangle^\perp$ are orthogonal, $\mu - \mathbf{1}_m r \in \langle \sigma \rangle$, or $\exists \lambda \in L^2_{t,T,d} : \mu - \mathbf{1}_m r = \sigma \lambda$.

5.4.3 Market completeness

Consider the following definition:

DEFINITION 5.4 (Market completeness). *Markets are* dynamically complete *if for each random variable $Y \in L^2(\Omega, \mathcal{F}, P)$, we can find a portfolio process $\pi : V_T^{x,\pi,0} = Y$ a.s.*

The previous definition is the natural continuous-time counterpart to that given in the discrete-time case (see Chapter 2). Consistently with the conclusions in Chapter 2, we shall explain that, in continuous-time, markets are dynamically complete if and only if (i) $m = d$ and (ii) the price volatility matrix of the available assets (primitives and derivatives) is nonsingular. We shall provide a sketch of the proof for the sufficiency part of this statement (see, e.g., Karatzas (1997 pp. 8-9) for the converse), which relates to the existence of fully spanning dynamic strategies. So given a $Y \in L^2(\Omega, \mathcal{F}, P)$, let $m = d$ and suppose the volatility matrix σ is nonsingular. Let us consider the Q-martingale:

$$M_\tau \equiv E^Q \left(S_{0T}^{-1} \cdot Y \middle| \mathbb{F}_\tau \right). \tag{5.73}$$

By the representation theorem of continuous local martingales as stochastic integrals with respect to Brownian motions,[15] there exists a $\varphi \in L^2_{0,T,d}(\Omega, \mathcal{F}, Q)$ such that

$$M_\tau = M_t + \int_t^\tau \varphi_u^\top dW_{0u}.$$

[14] To see that $\langle \sigma \rangle$ and $\langle \sigma^\top \rangle^\perp$ are orthogonal spaces, note that

$$\begin{aligned} \left\{ x \in L^2_{t,T,m} : x^\top z = 0,\ z \in \langle \sigma \rangle \right\} &= \left\{ x \in L^2_{t,T,m} : x^\top \sigma u = 0,\ u \in L^2_{t,T,d} \right\} \\ &= \left\{ x \in L^2_{t,T,m} : x^\top \sigma = \mathbf{0}_d \right\} \\ &= \left\{ x \in L^2_{t,T,m} : \sigma^\top x = \mathbf{0}_d \right\} \\ &\equiv \langle \sigma^\top \rangle^\perp. \end{aligned}$$

[15] See, e.g., Karatzas and Shreve (1991, Theorem 4.2 p. 170).

We wish to find a portfolio process π such that the discounted wealth process net of consumption, $S_{0\tau}^{-1}V_\tau$, equals M_τ under P, or, equivalently, under Q. Let $V_\tau^{x,\pi}$ denote the wealth at τ achieved by a portfolio process π when the wealth at time-t is x. By Eq. (5.72),

$$\frac{V_\tau^{x,\pi}}{S_{0\tau}} = x + \int_t^\tau \frac{\pi_u^\top \sigma_u}{S_{0u}} dW_{0u},$$

such that, by identifying, the portfolio we are searching for is $\hat{\pi}^\top = S_0\varphi^\top\sigma^{-1}$. Set $x = M_t$. Then, $M_\tau = S_{0\tau}^{-1}V_\tau^{M_t,\hat{\pi}}$ and, in particular, $M_T = S_{0T}^{-1}V_T^{M_t,\hat{\pi}}$. Comparing with Eq. (5.73) leaves $V_T^{M_t,\hat{\pi}} = Y$. Therefore, and based on the fact that there exists a unique λ satisfying $\mu - \mathbf{1}_m r = \sigma\lambda \iff m = d$, we have the following result:

THEOREM 5.5. *$\mathcal{Q}$ is a singleton if and only if markets are complete.*

Similarly as in the discrete time models of Chapter 2, when markets are incomplete, absence of arbitrage does not allow us to identify a unique risk-neutral probability. The next results provide additional characterizations of the many risk-neutral probabilities in $\mathcal{Q}$, in the incomplete markets case. Let $L^2_{0,T,d}(\Omega, \mathcal{F}, P)$ be the space of all $\mathbb{F}_t$-adapted processes x in $\mathbb{R}^d$ satisfying $\int_0^T \|x_u\|^2 du \in (0, \infty)$, and define $\langle\sigma\rangle^\perp \equiv \{x \in L^2_{0,T,d}(\Omega, \mathcal{F}, P) : \sigma_t x_t = \mathbf{0}_m\}$. Let

$$\hat{\lambda} = \sigma^\top \left(\sigma\sigma^\top\right)^{-1} (\mu - \mathbf{1}_m r).$$

Under the usual regularity conditions, $\hat{\lambda}$ can be interpreted as the process of unit risk-premia. In fact, *all* processes belonging to the set $\mathcal{Z} = \{\lambda : \lambda_t = \hat{\lambda}_t + \nu_t,\ \nu \in \langle\sigma\rangle^\perp\}$ are bounded and, then, are interpreted as unit risk-premia processes. More precisely, define the Radon-Nikodym derivative of Q with respect to P on $\mathcal{F}_T$:

$$\hat{\eta}_T \equiv \left.\frac{dQ}{dP}\right|_{\mathbb{F}_T} = \exp\left(-\frac{1}{2}\int_0^T \left\|\hat{\lambda}_t\right\|^2 dt - \int_0^T \hat{\lambda}_t^\top dW_t\right),$$

and the density process of all $Q \approx P$ on $(\Omega, \mathcal{F})$,

$$\eta_t = \hat{\eta}_t \cdot \exp\left(\frac{1}{2}\int_0^t \|\nu_u\|^2 du - \int_0^t \nu_u^\top dW_u\right), \quad t \in [0, T]),$$

a strictly positive P-martingale. We have the following results, which follows for example by He and Pearson (1991, Proposition 1 p. 271) or Shreve (1991, Lemma 3.4 p. 429):

PROPOSITION 5.6. *$Q \in \mathcal{Q}$ if and only if it takes the form*: $Q(A) = E(1_A\eta_T)$, $\forall A \in \mathcal{F}_T$.

To summarize, we have that $\dim(\langle\sigma\rangle^\perp) = d - m$. The previous result nitidly shows that when markets are incomplete, there are many risk-neutral probabilities. Such a result was shown in great generality by Harrison and Pliska (1983).[16]

[16]The so-called Föllmer and Schweizer (1991) measure, or minimal equivalent martingale measure, is defined as: $\hat{P}^*(A) \equiv E(1_A\hat{\eta}_T)$, for each $A \in \mathcal{F}_T$.

5.5 Equilibrium with a representative agent

This and the next section consider equilibrium implications of portfolio choices relying on the existence of representative agent. In a nutshell, we derive optimal portfolio choices of an investor in a dynamic setting, thereby extending mean-variance results in Chapter 1. In this section, we solve for these choices relying on dynamic programming, the approach introduced by Merton (1969); in Section 5.6, we rely on a relatively more recent approach, known as "martingale methods." An equilibrium is defined as the situation in which, amongst others, the asset prices dynamics are such that the investor is happy with his portfolio holdings. These results are known since Merton (1971, 1973) and Breeden (1979).

[In progress]

5.5.1 Constant investment opportunity sets

An agent faces the following problem:

$$\mathcal{J}\left(V_t\right) = \max_{c} E_t\left(\int_t^{\infty} e^{-\rho(\tau-t)} u\left(c_\tau\right) d\tau\right), \quad \text{s.t. Eq. (5.67) holds.}$$

Under regularity conditions, the value function $J\left(\cdot\right)$ satisfies the following Hamilton-Jacobi-Bellman equation:

$$0 = \max_{c,\pi}\left(u\left(c\right) + \mathcal{J}'\left(V\right)\left(\pi^\top(\mu - \mathbf{1}_m r) + rV - c\right) + \frac{1}{2}\mathcal{J}''\left(V\right)\pi^\top\sigma\sigma^\top\pi - \rho\mathcal{J}\left(V\right)\right). \tag{5.74}$$

The first order conditions lead to:

$$u'\left(c\right) = \mathcal{J}'\left(V\right) \quad \text{and} \quad \pi = \left(\frac{-\mathcal{J}'(V)}{\mathcal{J}''(V)}\right)\left(\sigma\sigma^\top\right)^{-1}\left(\mu - \mathbf{1}_m r\right). \tag{5.75}$$

To solve for the value function, replace Eqs. (5.75) into Eq. (5.74); the result is an ordinary differential equation solved by $\mathcal{J}$. In the CRRA case, $u\left(c\right) = c^{1-\gamma}$, the solution is $\mathcal{J}(V) = a\frac{V^{1-\gamma}-b}{1-\gamma}$, for two constants a, b,[17] such that consumption and portfolio policies satisfy

$$c = a^{-\frac{1}{\gamma}}V, \qquad \frac{\pi}{V} = \frac{1}{\gamma}\left(\sigma\sigma^\top\right)^{-1}\left(\mu - \mathbf{1}_m r\right).$$

5.5.2 Stochastic opportunity sets

Next, we consider a case of potentially incomplete markets, in which an exogenous state process y_t drives asset prices. It is solution to

$$dy_t = \varphi\left(y_t\right)dt + \Sigma\left(y_t\right)dW_t, \tag{5.76}$$

where W_t is a d-dimensional Brownian motion and φ and Σ are some vector and matrix valued functions that satisfy the usual regularity conditions. This setup is somehow abstract for now. These state variables can be inflation, exchange rates, industrial production growth, etc.

[17] To determine these two constants, replace optimal consumption, $c = a^{-1/\gamma}V$, into Eq. (5.74), use the conjecture for the value function $\mathcal{J}$, and solve for the undetermined coefficients, obtaining $a = (\frac{\rho - r(1-\gamma)}{\gamma} - \frac{(1-\gamma)Sh}{2\gamma^2})^{-\gamma}$ and $b = (a\rho)^{-1}$, where $Sh \equiv (\mu - \mathbf{1}_m r)^\top(\sigma\sigma^\top)^{-1}(\mu - \mathbf{1}_m r)$. As $\gamma \to 1$, we have that $\mathcal{J}\left(V_t\right) = \frac{1}{\rho}\ln V_t$ and $c_t = \rho V_t$.

Naturally, an important task in financial economics is to explain how some of these variables in these examples (e.g., the rate of economic growth) are affected by asset prices, but the scope of this chapter is to examine market behavior in this simplified setting; Part II contains more elaborated explanations of market behavior (see Chapters 8 and 9). Similarly as for the discrete time model of Lucas (1978) (see Chapter 3), we assume that there are assets that entitle to dividends (the numéraire) of m trees and that their prices are driven by y_t. That is, if S_{it} denotes the time-t price of the i-th asset, S_{it} is some function $S_{it} \equiv S_i(t, y_t)$ (the price at t when the state process is y_t), which, in some models, shall be determined in equilibrium. We assume $S_i(t, y_t)$ is as differentiable as needed to apply Itô's lemma, such that

$$\frac{dS_{it}}{S_{it}} = \bar{\mu}_i(t, y_t)\, dt + \sigma_i(t, y_t)\, dW_t, \tag{5.77}$$

where

$$\bar{\mu}_i S_i = \frac{\partial S_i}{\partial t} + \frac{\partial S_i}{\partial y}\varphi + \frac{1}{2}\mathrm{Tr}\left(\Sigma\Sigma^\top \frac{\partial^2 S_i}{\partial y \partial y^\top}\right), \quad \sigma_i S_{it} = \frac{\partial S_i}{\partial y}\Sigma,$$

and remaining notation is straightforward. We also assume that there is a safe asset in zero net supply, i.e., an asset with a price S_{0t} satisfying Eq. (5.77) with $\bar{\mu}_0(t, y) \equiv r(t, y)$ (the short-term rate) and $\sigma_0(t, y) \equiv \mathbf{0}_m$.

Thus, Eq. (5.77) is a specific version of the Eq. (5.66) in Section 5.4.1. In particular, σ is the $m \times d$ matrix with entry $\sigma_{ij}(t, y)$, i.e., the i-th asset price exposure to the j-th source of risk in Eq. (5.76). We assume that $\mathrm{rank}(\sigma) = m \leq d$ for each τ, just as in Section 5.4.1, such that the value V_t of a self-financed portfolio satisfies Eq. (5.67). Finally, we assume that $D_{it} = D_i(t, y_t)$ for a given function D_i.

5.5.2.1 Markets with non-productive assets

While the price system in (5.77) can lead to incomplete markets, we also consider a case in which markets could be (but are not necessarily) *completed* by "non-productive" assets in zero-net supply, interpreted as derivative instruments.[18] Let σ_F denote the volatility matrix of these contingent claims, defined similarly as σ. A condition under which these non-productive contingent claims complete the market is that the augmented volatility matrix $\hat{\sigma} = \left[\sigma^\top \;\; \sigma_F^\top\right]^\top$ is square and has full-rank. This condition is easy to interpret: rewrite the augmented price system as

$$\begin{pmatrix} \frac{dS_{1t}}{S_{1t}} \\ \dots \\ \frac{dS_{mt}}{S_{mt}} \\ \frac{dF_{1t}}{F_{1t}} \\ \dots \\ \frac{dF_{nt}}{F_{nt}} \end{pmatrix} = \begin{pmatrix} \bar{\mu}_{1t} \\ \dots \\ \bar{\mu}_{mt} \\ \bar{\mu}^F_{1t} \\ \dots \\ \bar{\mu}^F_{nt} \end{pmatrix} dt + \hat{\sigma} dW_t, \tag{5.78}$$

where F_{it} denotes the time-t price of the i-the contingent claim. Then, the condition $\mathrm{rank}(\hat{\sigma}) = d$ means that observing the asset prices is equivalent to observing the sources of uncertainty in the market.

The notion that derivative assets can complete the market was first introduced by Ross (1976). Naturally, the presence of derivative securities does not guarantee that markets are

[18] We formulate a model that, according to parameter restrictions, can be interpreted in the spirit of the discrete time Lucas' (1978) tree model, and also in terms of the Cox, Ingersoll and Ross model (1985a).

complete; further conditions are required, as discussed in the example of a stochastic volatility model (see Eqs. (5.96) below).

Let us consider a frictionless economy with a representative agent,[19] and derive the equilibrium restrictions on (5.78). The agent maximizes his intertemporal utility under a budget constraint that generalizes Eq. (5.67); this new budget constraint reflects the agent's increased portfolio possibilities arising through non-productive assets. Let us denote his value function with $\mathcal{J}(t,V_t,y_t) \equiv \max_c E_t(\int_t^T u(s,c_s)ds)$. Under regularity conditions, the decision problem can be summarized by a Bellman equation, $0 = \max_{c,\pi} G(t,V,y;c,\pi,\pi_F)$, where

$$\begin{aligned} G(t,V,y;c,\pi,\pi_F) \equiv & \, u(t,c) + \mathcal{J}_t(t,V,y) + \mathcal{J}_V(t,V,y)\left(\pi^\top(\mu - \mathbf{1}_m r) + \pi_F^\top\left(\mu^F - \mathbf{1}_n r\right) + rV - c\right) \\ & + \frac{1}{2}\mathcal{J}_{VV}(t,V,y)\pi_a^\top\hat{\sigma}\hat{\sigma}^\top\pi_a + \pi_a^\top\hat{\sigma}\Sigma^\top\mathcal{J}_{Vy}(t,V,y) + \frac{1}{2}\mathrm{Tr}\left(\Sigma\Sigma^\top\mathcal{J}_{yy}(t,V,y)\right), \end{aligned} \tag{5.79}$$

and where μ^F is a n-dimensional vector with generic element μ_j^F, and μ_j^F is the cum-dividend expected return of the j-th contingent claim (and a similar notation regards μ, as usual); π_F is the n-dimensional vector process of wealth invested in the contingent claims; finally, $\pi_a = \left[\pi^\top \;\; \pi_F^\top\right]^\top$.

The first order conditions are $0 = u_c - \mathcal{J}_V$, and

$$\left\{ \begin{array}{rcl} \mathbf{0}_m & = & (\mu - \mathbf{1}_m r)\mathcal{J}_V + (\sigma\sigma^\top\pi + \sigma\sigma_F^\top\pi_F)\mathcal{J}_{VV} + \sigma\Sigma^\top\mathcal{J}_{Vy} \\ \mathbf{0}_m & = & (\mu^F - \mathbf{1}_n r)\mathcal{J}_V + (\sigma_F\sigma_F^\top\pi_F + \sigma_F\sigma^\top\pi)\mathcal{J}_{VV} + \sigma_F\Sigma^\top\mathcal{J}_{Vy} \end{array} \right. \tag{5.80}$$

We now elaborate on these conditions and characterize portfolio choices and their asset pricing implications.

5.5.2.2 Hedging demands and separation theorems

We focus on portfolio selection regarding productive assets, leaving some discussion regarding non-productive assets below. By the first of Eqs. (5.80), the assets demand is

$$\frac{\pi}{V} = \left(\sigma\sigma^\top\right)^{-1}(\mu - \mathbf{1}_m r)\frac{\mathcal{J}_V}{-V\mathcal{J}_{VV}} + \left(\sigma\sigma^\top\right)^{-1}\sigma\Sigma^\top\frac{\mathcal{J}_{Vy}}{-V\mathcal{J}_{VV}}. \tag{5.81}$$

Eq. (5.81) shows that our agent aims to a portfolio that has two components. The first is proportional to the "tangent portfolio" and the second is the "hedging portfolio," a term we use for reasons explained below. The notion of a tangent portfolio is known since Chapter 1: in the static mean-variance context, it is the portfolio on the efficient frontier that allows an investor to construct any desired asset allocation consistent with his risk-return trade-off profile. Precisely, any efficient risk-return complex (in the mean-variance sense) can be achieved by investing proportions of wealth in a risk-free asset *and* in the tangent portfolio. If the riskless asset is in zero-net supply, then, all agents hold the tangent portfolio, which for this reason is also known as the market portfolio.

In the dynamic market of this section, the tangent portfolio also collapses to a market portfolio when the price system encompasses the whole information provided by the state variables. In this case, indeed, the value function ceases to directly depend on y, which makes $\mathcal{J}_{Vy} = 0$;

[19] It is known since at least Huang (1987) that in the diffusive case of this chapter, and under conditions, there exists a representative agent economy supporting the prices in the initial economy when markets are complete. This result is based on the welfare theorems of Chapter 2, and will be further reviewed in Part II (see Chapter 9).

and pre-multiplying Eq. (5.81) by $\mathbf{1}_m^\top$, assuming $\mathcal{J}_{Vy} = 0$, using the equilibrium condition that $\mathbf{1}_m^\top \pi = V$ (safe asset is in zero net supply) and, finally, rearranging terms, then leaves the following restriction on the safe interest rate[20]

$$(D - Cr)^{-1} = \frac{\mathcal{J}_V}{-V\mathcal{J}_{VV}}, \tag{5.82}$$

where $D = \mathbf{1}_m^\top \left(\sigma\sigma^\top\right)^{-1} \mu$ and $C = \mathbf{1}_m^\top \left(\sigma\sigma^\top\right)^{-1} \mathbf{1}_m$, consistently with similar notation introduced in Chapter 1. By replacing Eq. (5.82) into Eq. (5.81) leaves, for $\mathcal{J}_{Vy} = 0$, the (continuous time diffusive equivalent to the) market portfolio in the mean-variance context of Chapter 1, as claimed (see Eq. (1.16)).

It is indeed a market portfolio: in equilibrium, $\pi_{it} = \bar{\theta}_i S_{it}$, where $\bar{\theta}_i$ is the number of outstanding securities, and because we assume the safe asset is in zero net supply, our representative agent then holds the value-weighted portfolio of all existing securities in positive supply, $V_t = \sum_{i=1}^m \bar{\theta}_i S_{it}$. Note, finally, that there is one more instance in which the tangent portfolio collapses to the market portfolio, arising when the indirect utility function is separable in V and y, such that $\mathcal{J}_{Vy} = 0$. This case occurs, for example, when the representative investor has instantaneous logarithmic utility, a case we shall return to below.

When the markets do not encompass all the information provided by the state variables, the value function depends directly on y, and the asset demand also includes the second component in Eq. (5.81), which we refer to as *hedging demand*. This second component captures investors' concerns of mitigating their risks beyond standard mean-variance diversification. It reflects a demand for d portfolios, where each portfolio maximizes the covariance of wealth with the state variable y_i for a given variance of wealth (i.e., each portfolio maximizes the correlation between wealth and y_i). Let us show these claims.

Let Σ_i denote the vector of exposures of factor y_i to the Brownian motions in Eq. (5.77), i.e., the i-th row of Σ, and define

$$p_i^* = \arg\max_p cov_{p,y_i} \quad \text{s.t.} \quad p^\top \sigma\sigma^\top p = \sigma_p^2,$$

where $cov_{p,y_i} \equiv p^\top \sigma \Sigma_j^\top$ denotes the instantaneous covariance between wealth and the state variable y_i when wealth is invested in p, and σ_p^2 is a given level of variance. We have

$$p_i^* = \frac{1}{2\ell_i} \left(\sigma\sigma^\top\right)^{-1} \sigma\Sigma_i^\top,$$

where ℓ_i is a Lagrange multiplier. The second component in Eq. (5.81) is a weighted average of these d portfolios, with weights equal to $\frac{\mathcal{J}_{Vy_i}}{-V\mathcal{J}_{VV}} 2\ell_i$.

Eq. (5.81) is, thus, a "$(d+2)$-separation theorem" that provides us with interesting generalizations of the static CAPM of Chapter 1. It tells us that investors can achieve any asset allocation consistent with their risk-return trade-off profiles by means of $d+2$ "mutual funds" that invest into (i) the riskless security, (ii) the mean-variance portfolio and, finally, (iii) d portfolios that have the highest correlation with the d state variables.[21] How much is invested in each of these funds depends on the investors' attitude to risk.

We now discuss some examples and, then, proceed with the asset pricing implications of the model.

[20] That is, we determine the restriction in (5.82) while simply requiring that the market for the safe asset clears, just as in the static model (see Section 1.3.4 in Chapter 1).

[21] Note that to simplify the presentation, the mean-variance and the d hedging protfolios are not re-scaled to sum up to one.

5.5.2.3 One example: stochastic volatility and asset allocation puzzles

The hedging demand can be understood from a different angle. Note that

$$p_i^* \propto \arg\min_p \left\|\Sigma_i - p^\top \sigma\right\|^2 = \left(\sigma\sigma^\top\right)^{-1} \sigma\Sigma_i^\top. \tag{5.83}$$

That is, p_i^* minimizes the "gap" (in the mean-square sense) between the dynamics of optimally invested wealth and those of the state variables. By attempting at mimicking the factors dynamics, the hedging demand provides the best (mean-square) hedge against fluctuations in the state variables. How much of these hedging demand is generated (and in which sign) depends on the investor's attitude to risk.

There is one case in which the hedging demand completely fills these gaps, which arises when markets are complete. Markets may be complete because there are exactly as many productive assets as state variables. In this case, $p_i^* \propto \left(\sigma^\top\right)^{-1} \Sigma_i^\top$, such that the correlation between the hedging portfolio returns and the state variable y_i is just one. Therefore, hedging demands arise even when markets are complete.

But markets may be complete even with relatively few productive assets, namely, when (i) the number of productive and non-productive assets equals d and (ii) the augmented volatility matrix $\hat{\sigma}$ in Eq. (5.78) has full rank, as previously explained. Consider the following example, and a market for a single security with price equal to S. Assume that W_t has two components, and that asset returns exhibit *stochastic volatility*, in that

$$\frac{dS_t}{S_t} = \bar{\mu}\left(S_t, y_t\right) dt + s\left(y_t\right) dW_{1t}, \tag{5.84}$$

for some functions $\bar{\mu}\left(\cdot,\cdot\right)$ and $s\left(\cdot\right)$, and where y_t is solution to Eq. (5.76).

Part III (e.g., Chapter 11) contains surveys of these models with stochastic volatility, which seem to capture a few stylized features of market behavior and that, for this reason, are the workhorse in standard market practice. In this chapter, the volatility function $s\left(\cdot\right)$ is given by Mother Nature, an assumption further discussed below. Instead, the expected return $\bar{\mu}\left(\cdot,\cdot\right)$ is determined in the equilibrium of the model (see Eq. (5.96) below).

In this example, the market is incomplete, because there is only one asset against two standard Brownian motions. However, markets can be completed by a second asset, a non-productive asset such as a European option written on the productive security. We shall return to this issue briefly below. Chapter 11 discusses more precise conditions under which derivatives can complete markets in this context (see Section 11.5). We now proceed to explain some interesting features of this model that help rationalize standard advisory pratice regarding long-term portfolio choice.

Chapter 1 explains that mean-variance theory generates efficient portfolios that clash with advisory pratice (see Section 1.2). One "asset allocation puzzle" is that less-risk averse investors are advised to increase their portfolio holdings of more risky assets (see Section 1.2). It is a puzzle from a mean-variance perspective because, in a mean-variance context, investors decide their exposure to risk while choosing appropriate proportions of the tangent portfolio and the risk-free asset. This puzzle (and a possible solution to it) can be illustrated within the context of the simple model in (5.84) (even if this is a single-asset model!). In term of Eq. (5.78), we have that $m = 1$, $d = 2$ and $\sigma = [s \;\; 0]$ and in terms of Eq. (5.76), we have that y_t is a single state variable with $\Sigma = [\Sigma_1 \;\; \Sigma_2]$. Therefore, by Eq. (5.81), the demand for the risky security is

$$\frac{\pi}{V} = \frac{\mu - r}{s^2} \frac{\mathcal{J}_V}{-V\mathcal{J}_{VV}} + \frac{\Sigma_1}{s} \frac{\mathcal{J}_{Vy}}{-V\mathcal{J}_{VV}}. \tag{5.85}$$

How does time-varying risk affect portfolio choice? The first term in (5.85) reflects standard mean-variance trade-offs, and the second is the hedging component, which is proportional to p_i^* in Eq. (5.83). [Provide succinct discussion of Campbell and Viceira (2002) and Chacko and Viceira (2005)].

[In progress]

5.5.3 Arrow-Debreu densities and restrictions on expected returns

5.5.3.1 The martingale restriction

Finally, we study the equilibrium restrictions on the asset expected returns. In equilibrium, $\pi_t = \text{Cap}_t$ and $\pi_{Ft} = \mathbf{0}_n$, where Cap is the m-dimensional vector of market capitalizations, i.e., $\text{Cap}_{it} = \bar{\theta}_i S_{it}$ for each i. Eqs. (5.80) then reduce to

$$\begin{pmatrix} \mu - \mathbf{1}_m r \\ \mu^F - \mathbf{1}_n r \end{pmatrix} = \hat{\sigma}\lambda, \quad \lambda \equiv -\left(\sigma^\top \text{Cap}\frac{\mathcal{J}_{VV}}{\mathcal{J}_V} + \Sigma^\top \frac{\mathcal{J}_{Vy}}{\mathcal{J}_V}\right). \tag{5.86}$$

The unit risk premiums, λ, relate to the curvature of the indirect utility function and the volatility of asset returns and the state variables, as we should expect: the curvature reflects the investor's attitude towards risks and the volatility determines the size of these risks. Importantly, Eq. (5.86) shows that the agent's optimizing behaviors (and equilibrium) lead to a single risk-neutral probability in $\mathcal{Q}$, regardless of whether asset markets are complete (see Eqs. (5.71)).

Eqs. (5.86), then, amount to "martingale restrictions" on asset prices, thereby providing a theory of asset prices for both productive and non-productive assets (such as options). For example, the price of the non-productive assets is the solution to partial differential equations subject to appropriate boundary conditions, similarly as in the models of Sections 5.1 and 5.2.[22] Naturally, these prices are not preference-free: even when markets are completed by their very same assets ($m + n = d$), no security is redundant in these markets.

The next section explains how to determine the last element of these markets: the safe interest rate, r. Prior to this, we provide additional discussion and illustrate how these martingale restrictions lead to famous generalizations of the CAPM.

5.5.3.2 Multibeta CAPM

We describe expected excess returns based on a representation that has the same flavor as the CAPM. Accordingly, we rely on derivations that are similar to those in Appendix 2 of Chapter 1. Note that Eq. (5.81) provides the weights of the market portfolio, which are split into the tangent (or mean-variance efficient) and the hedging portfolios. We now elaborate on the previous martingale restrictions, by applying them to both the market and hedging portfolios. Appendix 3 provides precise analytical details of the following derivations.

First, note that the first of Eqs. (5.86) is

$$\mu - \mathbf{1}_m r = \text{cov}_{S,M}\frac{-V\mathcal{J}_{VV}}{\mathcal{J}_V} + \sigma\Sigma^\top \frac{-\mathcal{J}_{Vy}}{\mathcal{J}_V}, \tag{5.87}$$

where $\text{cov}_{S,M}$ is the instantaneous covariance between the productive assets returns and wealth invested in the market portfolio.

[22] As in Sections 4.1 and 4.2, these partial differential equations arise because, by Itô's lemma, μ^h and σ_h involve partial derivatives of asset prices with respect to state variables, just as μ and σ in Eq. (5.77).

Next, pre-multiply Eq. (5.87) by the market portfolio, obtaining

$$\mu_M - r = \text{var}_M \frac{-V\mathcal{J}_{VV}}{\mathcal{J}_V} + \text{cov}_{M,y} \frac{-\mathcal{J}_{Vy}}{\mathcal{J}_V}, \tag{5.88}$$

where $\mu_M \equiv \frac{\pi_M^\top}{V}\mu$, and $\pi_M^\top$ is as in Eq. (5.81) and equal to Cap in equilibrium (such that $\pi_M^\top \mathbf{1}_m = V$), that is, μ_M is the market expected return; finally, var_M denotes the instantaneous variance of wealth invested in the market portfolio.

Furthermore, pre-multiply Eq. (5.87) by the d hedging portfolios, defined as

$$\frac{\pi_{H_i}}{V} \equiv \frac{\left(\sigma\sigma^\top\right)^{-1}\sigma\Sigma_i^\top}{\mathbf{1}_m^\top\left(\sigma\sigma^\top\right)^{-1}\sigma\Sigma_i^\top}, \quad i = 1, \cdots, d. \tag{5.89}$$

We have, for each of the expected excess returns on the hedging portfolios, that

$$\mu_{H_i} - r = \text{cov}_{H_i,M} \frac{-V\mathcal{J}_{VV}}{\mathcal{J}_V} + \text{cov}_{H_i,y} \frac{-\mathcal{J}_{Vy}}{\mathcal{J}_V}, \quad i = 1, \cdots, d. \tag{5.90}$$

with notation similar to that in Eq. (5.88).

Eqs. (5.88) and (5.90) can be regarded to form a system of $d+1$ linear equations, which can be solved for the $d+1$ unknowns, $\frac{-V\mathcal{J}_{VV}}{\mathcal{J}_V}$ and $\frac{-\mathcal{J}_{Vy_i}}{\mathcal{J}_V}$. Replacing the solutions into Eq. (5.87) leaves the following restriction on each asset expected excess returns

$$\mu_i - r = \beta_{i,M}\left(\mu_M - r\right) + \sum_{j=1}^{d} \beta_{i,H_j}(\mu_{H_j} - r), \quad i = 1, \cdots, m, \tag{5.91}$$

where the coefficients $\beta_{i,M}$ and β_{i,H_j} are the coefficient estimates of the multiple regression of the excess returns of security i onto the excess returns on the market and the hedging portfolios.

Note that if the investor has log-utility, $\mathcal{J}_{Vy} = 0$, or in the absence of stochastic opportunity sets, Eqs. (5.87)-(5.88) immediately deliver

$$\mu - \mathbf{1}_m r = \frac{\text{cov}_{S,M}}{\text{var}_M}\left(\mu_M - r\right). \tag{5.92}$$

Eq. (5.92) is the continuous time equivalent to the static single beta CAPM of Chapter 1.

5.5.3.3 Consumption CAPM

In equilibrium, instantaneous consumption equals the sum of the instantaneous dividend rates promised by all the securities. But we are assuming that these dividend rates depend on the exogenous state process y. Therefore, in equilibrium, consumption is driven by this state process. Note, also, that consumption may depend on y even when the dividend rates are independent of y but utility is state dependent. Part II contains many detailed discussions of these cases (see Chapters 8 and 9). Our objective, now, is to derive a martingale restriction expressed in terms of equilibrium consumption, consistently with the introductory Lucas' tree examples of Section 5.2.6.

Consider the first order conditions, $u_c = \mathcal{J}_V$, which can be expanded with respect to both V and y, leaving

$$\frac{\mathcal{J}_{VV}}{\mathcal{J}_V} = \frac{u_{cc}}{u_c}c_V, \quad \frac{\mathcal{J}_{Vy}}{\mathcal{J}_V} = \frac{u_{cc}}{u_c}c_y,$$

where c_V and c_y denote the partial derivatives of the equilibrium consumption with respect to V and y, respectively. Based on the previous relations, the risk-premium in the martingale restrictions in Eqs. (5.86) can be re-stated as

$$\lambda = -\frac{cu_{cc}}{u_c}\left(\sigma^\top \text{Cap}\frac{c_V}{c} + \Sigma^\top \frac{c_y}{c}\right) \equiv -\frac{cu_{cc}}{u_c}\text{vol}_c^\top, \tag{5.93}$$

where, by Itô's lemma, vol_c can be recognized as the vector of consumption growth exposures to the primitive sources of risk,

$$\frac{dc}{c} = (\cdots)\, dt + \text{vol}_c dW.$$

These conclusions generalize those obtained in the introductory examples of Section 5.2.6. Finally, there is an alternative way to express the martingale restriction. Consider the first of Eqs. (5.86)

$$\mu - \mathbf{1}_m r = -\frac{cu_{cc}}{u_c}\sigma\text{vol}_c^\top \equiv -\frac{cu_{cc}}{u_c}\text{cov}_{S,c}, \tag{5.94}$$

where we have used Eq (5.93), and $\text{cov}_{S,c}$ denotes the vector of the m instantantaneous covariances of the securities returns with consumption. Pre-multiplying Eq. (5.94) by an arbitrary portfolio weighting $\frac{\pi_p^\top}{V}$ yields, $\mu_p - r = -\frac{cu_{cc}}{u_c}\text{cov}_{p,c}$, which we can replace back to Eq. (5.94), leaving

$$\mu - \mathbf{1}_m r = \frac{\text{cov}_{S,c}}{\text{cov}_{p,c}}\left(\mu_p - r\right). \tag{5.95}$$

Eq. (5.95) is known as the Consumption CAPM, or CCAPM. Note that the portfolio p is not the market portfolio. But if there is a portfolio p that is pefectly correlated with changes in consumption, Eq. (5.95) would imply that

$$\mu - \mathbf{1}_m r = \frac{\text{cov}_{S,c}}{\text{var}_c}\left(\mu_c - r\right),$$

where μ_c and var_c denote the average and the variance of consumption growth.

5.5.3.4 Markets completed by options and equilibrium option prices

Consider the previous example of a market with a single asset and random volatility (see Eq. (5.84)) and, to simplify, set $\theta_1 \equiv 1$. We now assume that there is a European option available for trading, and that this option completes the market, an assumption that holds through (see Chapter 11). The martingale restrictions in Eq. (5.86) collapse to

$$\begin{pmatrix} \mu - r \\ \mu^F - r \end{pmatrix} = \hat{\sigma}\lambda, \quad \hat{\sigma} = \begin{pmatrix} s & 0 \\ \sigma_1^F & \sigma_2^F \end{pmatrix}, \quad \lambda = \begin{pmatrix} s\frac{-V\mathcal{J}_{VV}}{\mathcal{J}_V} + \Sigma_1\frac{-V\mathcal{J}_{Vy}}{\mathcal{J}_V} \\ \Sigma_2\frac{-V\mathcal{J}_{Vy}}{\mathcal{J}_V} \end{pmatrix}, \tag{5.96}$$

where we have used the equilibrium condition $V = S$.

The first equation in (5.96) provides a restriction regarding the expected excess returns on the primitive asset, $\mu - r$; the next section explains how the safe interest rate r is determined (see Eq. (5.97) below). To summarize, in equilibrium, $V_t = S_t$, where S_t is solution to Eq. (5.84), with $\bar{\mu}\left(S_t, y_t\right) = \mu\left(S_t, y_t\right) - \frac{D_t}{S_t}$, and $\mu\left(S_t, y_t\right)$ restricted as just explained. The second condition is an option pricing theory. Chapter 11 explains it is a partial differential equation that the price of an option solves, subject to the appropriate boundary condition (see Section 11.5).

The risk premium reflect the representative investor attitudes to risk. If the discounted utility function exhibits CRRA, i.e., $u(t,c) = e^{-\rho t}\frac{c^{1-\gamma}}{1-\gamma}$, with ρ and γ constants, then, in equilibrium, $\mathcal{J}(V,y,t) = e^{-\rho t}g(y,t)\frac{V^{1-\gamma}}{1-\gamma}$, for some function g. Note that g is independent of y in the logarithmic utility case, $\gamma = 1$, such that $\mathcal{J}_{Vy} = 0$: the representative investor does not require compensation for bearing volatility risk,[23] an assumption that does not seem to receive empirical support.

When does the option complete markets? Note that the economic rationale for market completeness is critical. The equilibrium in this section relies on the assumption of a representative agent, an assumption that holds when markets are complete and under additional conditions identified by Huang (1987). In the market of this section, markets are complete once $\hat{\sigma}$ is invertible, as explained earlier. Assuming that s is strictly positive, $\hat{\sigma}$ is invertible if $\sigma_2^F \neq 0$ a.s. But then note that by Itô's lemma, $\sigma_2^F = \frac{F_y}{F}\Sigma_2$, where F denotes the option price and F_y is the partial of F with respect to the volatility state variable, y. Assuming $\Sigma_2 \neq 0$ (for otherwise volatility would not make markets incomplete), markets are complete when the option price is monotonic in y. Naturally, the option price is endogenous and a monotonicity property cannot be taken as an assumption, but Bajeux and Rochet (1996) and Romano and Touzi (1997) show that such a property holds under mild conditions, as further discussed in Chapter 11.

5.5.3.5 Asset volatility and general equilibrium

One aspect of the previous model with stochastic volatility is the assumption that the asset return volatility $s(y_t)$ is given by Mother Nature. While this assumption leads to tractable models with random volatility, an important task in financial economics is to elaborate theorerical explanations for asset volatility that are consistent with market behavior. These explanations necessitate a further elaboration of the framework in this section, and is provided in Part II (see Chapters 8 and 9). We now proceed to explain how interest rates are determined in the equilibrium of the model in this section.

5.5.4 Interest rates

5.5.4.1 I: consumption-based

We know already that interest rates and risk-premiums link to the marginal utility of consumption (see Section 5.2). How do these facts relate to the predictions in this section? Recall that by the first order conditions, an optimal consumption path satisfies $u_c = \mathcal{J}_V$: marginal utility of consumption equals marginal indirect utility of wealth. We wish to find the *equilibrium* dynamics of the marginal indirect utility of wealth, say $\mathcal{J}_V^*$. First, note that, in equilibrium,

$$d\ln\mathcal{J}_V^* = \left(\frac{L\mathcal{J}_V^*}{\mathcal{J}_V^*} - \frac{1}{2}\|\lambda\|^2\right)dt - \lambda dW,$$

[23] These results are known since at least Lewis (2000, p. 228 and 232). Pham and Touzi (1996) first showed that the volatility risk-premium (i.e., the second component in (5.96)) is zero when a representative investor has logarithmic utility.

where λ is just as in Eq. (5.86). Next, we determine the expected growth rate in the marginal utility of wealth. Let us differentiate the Bellman's equation,

$$\begin{aligned}0 &= G_V\left(t,V,y;c,\pi,\pi_F\right)\\ &= \mathcal{J}_{Vt}\left(t,V,y\right)+\mathcal{J}_{VV}\left(t,V,y\right)\left(\pi^\top\left(\mu-\mathbf{1}_m r\right)+\pi_F^\top\left(\mu^F-\mathbf{1}_n r\right)+rV-c\right)+r\mathcal{J}_V\left(t,V,y\right)\\ &+\frac{1}{2}\mathcal{J}_{VVV}(t,V,y)\pi_a^\top\hat{\sigma}\hat{\sigma}^\top\pi_a+\pi_a^\top\hat{\sigma}\Sigma^\top\mathcal{J}_{VVy}\left(t,V,y\right)+\frac{1}{2}\mathrm{Tr}\left(\Sigma\Sigma^\top\mathcal{J}_{Vyy}\left(t,V,y\right)\right),\end{aligned}$$

where G is defined in Eq. (5.79). Evaluating the previous expression in equilibrium, and re-arranging, leaves[24]

$$r=-\frac{L\mathcal{J}_V^*\left(t,V,y\right)}{\mathcal{J}_V^*\left(t,V,y\right)}=-\frac{Lu_c\left(t,c\right)}{u\left(t,c\right)}. \tag{5.97}$$

That is, the interest rate equilibrium dynamics link to those of consumption. For example, if the dividends of the productive assets are exogenously specified (the assumption in this section), then, the equilibrium dynamics of consumption relate to those of the sum of all the dividends, and are determined by expanding the R.H.S. of the last equality in (5.97). We shall return to these topics in Section 5.5.2, which are the basis of much of the analysis in Part II.

5.5.4.2 II: with exogenously given productive activities

In a famous contribution, Cox, Ingersoll and Ross (1985a) develop an asset pricing framework that is a variant of that in this section, because productive assets are "linear activities," with average growth μ_t and volatilities σ_t given by Mother Nature. These production activities do not distribute payoffs during their life. They are called linear because investing S_{i0} in the i-th of them at time-zero, delivers S_{it} at time-t.

We could derive the equilibrium interest rate in this framework by repeating the arguments leading to Eq. (5.82), but it is instructive to consider an alternative formulation. Let us evaluate the first set of the first-order conditions (5.86) at the equilibrium, pre-multiply the result by $\pi^\top$, and use the additional equilibrium condition $\pi^\top\mathbf{1}_m=V$, leaving

$$r=\frac{\pi^\top}{V}\mu+var\left(\frac{dV}{V}\right)\frac{V\mathcal{J}_{VV}}{\mathcal{J}_V}+cov\left(\frac{dV}{V},dy\right)\frac{\mathcal{J}_{Vy}}{\mathcal{J}_V}, \tag{5.98}$$

where the equilibrium dynamics of wealth are given by

$$dV_t=\left(\pi_t^\top\mu_t-c_t\right)dt+\pi_t^\top\sigma_t dW_t, \tag{5.99}$$

and $\pi_t=\mathrm{Cap}_t$ now denotes the vector of linear activities values.

Eq. (5.98) says that the safe interest rate equal the expected return of wealth plus additional factors reflecting risk-aversion. Intuitively, the higher the investor's risk-aversion, the lower the risky asset demand, which in equilibrium leads to lower interest rates needed to clear the market for the safe asset.

This model can be used to price the entire term structure of interest rates, and may be analytically convenient under some simplifying assumptions. Cox, Ingersoll and Ross (1985b) make such additional assumptions and derive a celebrated theory of the term structure of interest rates. Let us just consider an economy in which the investor can access a single production

[24] In particular, $L\mathcal{J}_V^*=\mathcal{J}_{Vt}^*+\mathcal{J}_{VV}^*\left(\pi^\top\mu-c\right)+\frac{1}{2}\mathcal{J}_{VVV}^*\mathrm{Cap}^\top\sigma\sigma^\top\mathrm{Cap}+\mathrm{Cap}^\top\sigma\Sigma^\top\mathcal{J}_{VVy}^*+\frac{1}{2}\mathrm{Tr}(\Sigma\Sigma^\top\mathcal{J}_{Vyy}^*)$.

activity, such that $V = \pi_1 = \theta_1 S_1$, with $\theta_1 = 1$. We assume that (i) $\mu_t = \hat{\mu} y_t$ and $\sigma_t \sigma_t^\top = \Omega y_t$, for two constants $\hat{\mu}$ and Ω, and that (ii) y_t in Eq. (5.76) is a single state variable, solution to

$$dy_t = \phi\left(\bar{y} - y_t\right) dt + v\sqrt{y_t} dW_t, \tag{5.100}$$

for some d-dimensional vector of constants v, and two constants ϕ and $\bar{y}$. Under these assumptions, $\pi = V$, and the equilibrium interest rate in Eq. (5.98) reduces to

$$r_t = \left(\hat{\mu} + \Omega \frac{V \mathcal{J}_{VV}}{\mathcal{J}_V}\right) y_t + \sigma_t v^\top \frac{\mathcal{J}_{Vy}}{\mathcal{J}_V} \sqrt{y_t}. \tag{5.101}$$

This expression for the equilibrium interest rate further simplifies when the indirect utility of the representative investor is separable, i.e., $\mathcal{J}_{Vy} = 0$. For example, this case occurs when the investor has logarithmic instantaneous utility, and Eq. (5.101) becomes

$$r_t = \left(\hat{\mu} - \Omega\right) y_t. \tag{5.102}$$

Assuming that $\hat{\mu} > \Omega$, and additional regularity conditions, the short-term rate is, then, strictly positive, and inherits the "square-root dynamics" from y_t in Eq. (5.100). Finally, in this economy, the risk premium in Eq. (5.86) is $\lambda_t = \bar{\sigma}^\top \sqrt{y_t}$, where $\bar{\sigma}$ is some vector of constants such that $\|\bar{\sigma}\|^2 = \Omega$; by Eq. (5.101), it can be expressed as a function of the short-term rate, as follows

$$\lambda_t = \bar{\sigma}^\top \left(\hat{\mu} - \Omega\right)^{-1/2} \sqrt{r_t}. \tag{5.103}$$

Chapter 13 examines this model in some more detail, but it also reviews another model along these lines, developed by Longstaff and Schwartz (1992), which aims to richer dynamics for the short-term rate than those in Eq. (5.102). The only more general assumption made by Longstaff and Schwartz is that the single linear activity is solution to

$$\frac{dS_1}{S_1} = \left(\hat{\mu}_1 y_{1t} + \hat{\mu}_2 y_{2t}\right) dt + \bar{\sigma}\sqrt{y_{2t}} dW_t,$$

where $\hat{\mu}_i$ are two constants, $\bar{\sigma}$ is a vector of constants with the same properties as that in the Cox, Ingersoll and Ross (1985b) model, and

$$dy_{it} = \phi_i\left(\bar{y}_i - y_{it}\right) dt + v_i\sqrt{y_{it}} dW_t, \quad i = 1, 2. \tag{5.104}$$

Under these assumptions, the short-term rate in Eq. (5.98) and the risk-premium in Eq. (5.86) are

$$r_t = \hat{\mu}_1 y_{1t} + \left(\hat{\mu}_2 - \Omega\right) y_{2t}, \quad \lambda_t = \bar{\sigma}^\top \sqrt{y_{2t}}. \tag{5.105}$$

That is, the equilibrium interest rate follows a two-factor model. This model is a useful benchmark to think about in a world where bond return volatility is random, as discussed in Chapter 13.

5.6 Martingale methods

5.6.1 Intertemporal budget constraints

We assume markets are complete, $m = d$, and consider market imperfections in the next section. As usual, our agent maximizes expected utility over consumption flows and terminal wealth under the constraint in Eq. (5.68):

$$J(t, V_t) = \max_{(\pi, c, v)} E_t\left(U(V_T) + \int_t^T u(c_\tau) d\tau\right), \quad \text{s.t. Eq. (5.68) holds.} \tag{4.P1}$$

This optimization problem can be solved relying on the notion of Arrow-Debreu state prices, similarly as in Chapter 2. The first task is to derive a budget constraint that parallels that in Chapter 2, and applied to a two period economy. For reference, in Chapter 2, we explained that in the presence of complete markets, the budget constraint can be written compactly as:

$$0 = c_0 - w_0 + E\left(m \cdot \left(c^1 - w^1\right)\right), \tag{5.106}$$

where c and w denote consumption and endowments, and m is the stochastic discount factor. It is instructive to re-iterate the steps to achieve Eq. (5.106). First, we have the two budget constraints at time zero and one, $c_0 - w_0 = -S\theta$, and $c^1 - w^1 = V\theta$, where notation is as in Chapter 2. By absence of arbitrage, $S = \phi^\top V$, such that $c_0 - w_0 = -\phi^\top (c^1 - w^1)$, and Eq. (5.106) follows by the definition of the Arrow-Debreu state prices,

$$\phi_s = m_s \cdot P_s, \quad m_s \equiv (1+r)^{-1} \eta_s, \quad \eta_s = \frac{Q_s}{P_s},$$

and by "taking the *sum* over all the states of nature."

The same logic would also obviously apply in the continuous time case. First, we define Arrow-Debreu state price *densities*,

$$\phi_{t,T} \equiv m_{t,T} \cdot dP, \quad m_{t,T} = S_{0T}^{-1} \eta_T, \quad \eta_T = \left.\frac{dQ}{dP}\right|_{\mathbb{F}_T}. \tag{5.107}$$

As in the finite state case, we want to evaluate the present value of future consumption and the current. For a given $\omega \in \Omega$, the value of future and present consumption is,

$$C_{t,T}(\omega) \equiv \phi_{t,T}(\omega) V_T(\omega) + \int_t^T \phi_{t,u}(\omega) c_u(\omega) du.$$

In terms of the two-period model, $C_{t,T}(\omega)$ is the counterpart to $c_0 + \phi_s (c_s - w_s)$, i.e. the value of present consumption (the numéraire) and consumption in state s. Next, we "take the *integral* over all states of nature," obtaining,

$$E_t\left(m_{t,T} V_T + \int_t^T m_{t,u} c_u du\right) = x, \tag{5.108}$$

where the first equality follows by Eq. (5.107), and the second by the budget constraint.

Note that the original problem, one with an infinite number of trajectory constraints, is now reduced to one with a single constraint, just as in the two period market model. In the Appendix, we also show that Eq. (5.108) is equivalent to one in which the expectation is taken under the risk-neutral probability, as follows:

$$x \equiv V_t = \mathbb{E}_t\left(\frac{V_T}{S_{0T}} + \int_t^T \frac{c_u}{S_{0u}} du\right) = E_t\left(m_{t,T} V_T + \int_t^T m_{t,u} c_u du\right). \tag{5.109}$$

Eq. (5.109) is the finite horizon counterpart of Eq. (5.47) of Section 5.3.4.2, derived through a different route—by rewriting the constraint under the risk-neutral probability and "integrating out." The approach in this section provides the economic intuition underlying that change of probability.

5.6.2 The optimization problem

So the program is,

$$J(t,x) = \max_{(c,v)} E\left(e^{-\rho(T-t)}U(V_T) + \int_t^T u(\tau, c_\tau)d\tau\right), \quad \text{s.t. Eq. (5.109)},$$

where v denotes state-contingent consumption arising from terminal wealth, V_T.

Because of the emphasis on the equivalent martingale measure, this approach to dynamic portfolio optimization is known as the *martingale method*. The critical assumption underlying it is that market are incomplete, such that there is a unique Arrow-Debreu density process underlying the optimization problem. However, this method could conveniently be extended to deal with the presence of portfolio constraints and incomplete markets in particular (see in Section 5.7).

Consider the Lagrangean:

$$\max_{(c,v)} E_t\left(\int_t^T \left(u\left(\tau, c_\tau\right) - \kappa \cdot m_{t,\tau} \cdot c_\tau\right) d\tau + U(v) - \kappa \cdot m_{t,T} \cdot v + \kappa \cdot x\right),$$

where κ is a Lagrange multiplier, and by Eqs. (5.70) and (5.107),

$$m_{t,\tau} = \exp\left(-\int_t^\tau \left(r_u + \frac{1}{2}\left\|\lambda_u\right\|^2\right) du - \int_t^\tau \lambda_u^\top dW_u\right). \tag{5.110}$$

The first order conditions are

$$\text{for } \tau \in [t,T), \quad u_c\left(\tau, c_\tau\right) = \kappa \cdot m_{t,\tau}, \quad \text{and } U'\left(V_T^{x,\pi,c}\right) = \kappa \cdot m_{t,T}. \tag{5.111}$$

To determine the portfolio-consumption policy, note that for a generic $\mathbb{F}_T$-measurable random variable $\hat{v}$, the martingale,

$$M_\tau \equiv \mathbb{E}_t\left(S_{0T}^{-1} \cdot \hat{v} + \int_t^T S_{0u}^{-1}\hat{c}_u du\right) = E_t\left(m_{t,T} \cdot \hat{v} + \int_t^T m_{t,u} \cdot \hat{c}_u du\right).$$

By the predictable representation theorem, there exists a vector ϕ such that.

$$M_\tau = M_t + \int_t^\tau \phi_u^\top dW_u,$$

and by Itô's lemma,

$$m_{0,t}V_\tau + \int_t^\tau m_{t,u} \cdot c_u du = x + \int_t^\tau m_{t,u} \cdot \left(\pi^\top \sigma - V\lambda^\top\right)_u dW_u.$$

By identifying,

$$\pi_\tau^\top = \left(V_\tau \lambda_\tau^\top + \frac{\phi_\tau^\top}{m_{t,\tau}}\right)\sigma_\tau^{-1}, \tag{5.112}$$

where V_τ can be determined from the constraint in Eq. (5.109) after setting $t \equiv \tau$ and $V_T = \hat{v}$.

5.6.3 Marginal utility of income

The Lagrange multiplier κ in Eq. (5.111) carries the interpretation of marginal utility of income, similarly as for some derivations in Chapter 2. We provide a proof of this quite intuitive "envelope property" by relying on an infinite horizon framework and instantaneous utility at t equal to $e^{-\rho t}u(c_t)$, as this is a framework often utilized in Part II of the Lectures. We, then, consider the following program, a special case of program [4.P1],

$$U(V_0) \equiv \max_{(c_t)_{t\geq 0}} E_0\left(\int_0^\infty e^{-\rho t}u(c_t)\,dt\right), \quad \text{s.t. } E_0\left(\int_0^\infty \frac{\xi_t}{\xi_0}c_t dt\right) = V_0, \quad \xi_0 \equiv 1. \qquad [4.P2]$$

where V_0 denotes initial wealth, and the notation,

$$\frac{\xi_{t+\Delta}}{\xi_t} \equiv m_{t,t+\Delta},$$

generalizes that in Section 5.2. Optimality conditions are,

$$\hat{c}_t : e^{-\rho t}u'(\hat{c}_t) = \kappa\xi_t \Longrightarrow u'(\hat{c}_0) = \kappa, \qquad (5.113)$$

as usual. Furthermore, by assuming enough regularity conditions, we can differentiate the intertemporal utility in [4.P2],

$$U'(V_0) = E_0\left(\int_0^\infty e^{-\rho t}u'(\hat{c}_t)\frac{\partial \hat{c}_t}{\partial V_0}dt\right) = \kappa E_0\left(\int_0^\infty \xi_t\frac{\partial \hat{c}_t}{\partial V_0}dt\right) = \kappa, \qquad (5.114)$$

where the second equality holds by the first of the optimality conditions (5.113), and the third by differentiating L.H.S. and R.H.S. of the budget constraint in [4.P2] with respect to V_0, as originally claimed.

Next, we determine κ, thereby pinning down $\hat{c}_0$ in the second of the optimality conditions (5.113). Rather than performing this task within a general framework, we assume the agent has CRRA equal to γ, such that the optimality conditions (5.113) become $\hat{c}_t = \kappa^{-\frac{1}{\gamma}}\left(e^{\rho t}\xi_t\right)^{-\frac{1}{\gamma}}$. By replacing optimal consumption into the budget constraint of [4.P2], and rearranging terms,

$$\kappa = V_0^{-\gamma}\left(\int_0^\infty e^{-\frac{\rho}{\gamma}t}E_0\left(\xi_t^{1-\frac{1}{\gamma}}\right)dt\right)^\gamma, \qquad (5.115)$$

and by integrating out Eq. (5.114),

$$U(V_0) = E_0\left(\int_0^\infty e^{-\frac{\rho}{\gamma}t}\xi_t^{1-\frac{1}{\gamma}}dt\right)^\gamma \times \frac{V_0^{1-\gamma}}{1-\gamma}.$$

Finally, by the optimality conditions (5.113), initial consumption has to satisfy,

$$\hat{c}_0 = \kappa^{-\frac{1}{\gamma}} = E_0\left(\int_0^\infty e^{-\frac{\rho}{\gamma}t}\xi_t^{1-\frac{1}{\gamma}}dt\right)^{-1}V_0 \equiv \text{const.}^{-1}V_0, \qquad (5.116)$$

where the second equality follows by the expression of κ in Eq. (5.115). Assume, for instance, log-utility, $\gamma = 1$, in which case $\hat{c}_0 = \rho V_0$ as in some of the models in Chapter 3.

This example shows what we would have to expect in general equilibrium. In equilibrium, optimal consumption equals the dividends of the trees. Assuming one tree to illustrate, we have that the initial endowment has to satisfy the following constraint,

$$V_0 = \text{const.}D_0,$$

where const. is the same constant defined in Eq. (5.116).

5.6.4 Example: log-utility

Let $U(v) = \ln v$, $u(x) = \ln x$ and set $\rho \equiv 0$. The optimality conditions (5.111) reduce to $\frac{1}{\hat{c}_\tau} = \kappa \cdot m_{t,\tau}$ and $\frac{1}{\hat{v}} = \kappa \cdot m_{\tau,T}$ and replacing these conditions into the budget constraint leaves the solution for the Lagrange multiplier: $\kappa = \frac{T+1}{x}$, such that $\hat{c}_t = \frac{x}{T+1}\frac{1}{m_{t,\tau}}$ and $\hat{v} = \frac{x}{T+1}\frac{1}{m_{t,T}}$. Regarding the portfolio process, we have that

$$M_\tau = E_t\left(m_{\tau,T}\cdot\hat{v} + \int_t^T m_{t,u}\hat{c}_u du\right) = x,$$

which shows that $\phi = 0$ in the representation of Eq. (5.112). Then, by replacing $\phi = 0$ into (5.112),

$$\pi_\tau^\top = V_\tau^{x,\pi,\hat{c}}\lambda_\tau\sigma_\tau^{-1}.$$

We determine $V^{x,\pi,\hat{c}}$ using $\hat{c}$ and Eq. (5.109),

$$V_\tau^{x,\pi,\hat{c}} = \frac{x}{T+1}E_t\left(\frac{m_{\tau,T}}{m_{t,T}} + \int_\tau^T \frac{m_{\tau,u}}{m_{t,u}}du\right) = \frac{x}{m_{t,\tau}}\frac{T+1-(\tau-t)}{T+1},$$

where we used the "semigroup property" that for $a \leq b \leq c$, $m_{a,b}\cdot m_{b,c} = m_{a,c}$. The solution is

$$\pi_\tau = \frac{x}{m_{t,\tau}}\frac{T+1-(\tau-t)}{T+1}\sigma_\tau^{-1}\lambda_\tau = \frac{x}{m_{t,\tau}}\frac{T+1-(\tau-t)}{T+1}\left(\sigma_\tau\sigma_\tau^\top\right)^{-1}(\mu_\tau - \mathbf{1}_m r_\tau),$$

where we used the expression for the risk-premium vector, $\mu - \mathbf{1}_m r = \sigma\lambda$.

5.6.5 Equilibrium

In a complete markets setting, an equilibrium is: (i) a consumption-portfolio policy satisfying Eqs. (5.111)-(5.112), and (ii) market clearing: for all $\tau \in [t,T)$,

$$c_\tau = D_\tau \equiv \sum_{i=1}^m D_{i\tau}, \quad \theta_{0\tau} = 0, \quad \pi_\tau = S_\tau. \tag{5.117}$$

We now derive equilibrium allocations and Arrow-Debreu state price densities. First, note that the dividend process, D, satisfies:

$$\frac{dD_\tau}{D_\tau} = g_\tau d\tau + \sigma_{0\tau}dW_\tau,$$

where $gD \equiv \sum_{i=1}^m g_i D_i$ and $\sigma_0 D \equiv \sum_{i=1}^m \sigma_{0i}D_i$.

We have:

$$\begin{aligned} d\ln u_c(\tau, D_\tau) &= d\ln u_c(\tau, c_\tau) \\ &= d\ln m_{t,\tau} \\ &= -\left(r_\tau + \frac{1}{2}\|\lambda_\tau\|^2\right)dt - \lambda_\tau^\top dW_\tau, \end{aligned} \tag{5.118}$$

where the first equality holds in an equilibrium, the second equality follows by the first order conditions in (5.111), and the third equality is true by the definition of $m_{t,\tau}$ in Eq. (5.110).

Finally, by Itô's lemma, $\ln u_c(\tau, D_\tau)$ is solution to:

$$d\ln u_c = \left(\frac{u_{\tau c}}{u_c} + gD\frac{u_{cc}}{u_c} + \frac{1}{2}\sigma_0^2 D^2\left(\frac{u_{ccc}}{u_c} - \left(\frac{u_{cc}}{u_c}\right)^2\right)\right)dt + \frac{u_{cc}}{u_c}D\sigma_0 dW. \tag{5.119}$$

By identifying drifts and diffusion terms in Eqs. (5.118)-(5.119), we obtain, after a few simplifications, the expression for the equilibrium short term rate and the prices of risk:

$$\begin{aligned} r_\tau &= -\left(\frac{u_{\tau c}(\tau, D_\tau)}{u_c(\tau, D_\tau)} + g_\tau D_\tau \frac{u_{cc}(\tau, D_\tau)}{u_c(\tau, D_\tau)} + \frac{1}{2}\sigma_{0\tau}^2 D_\tau^2 \frac{u_{ccc}(\tau, D_\tau)}{u_c(\tau, D_\tau)}\right) \\ \lambda_\tau^\top &= -\frac{u_{cc}(\tau, D_\tau)}{u_c(\tau, D_\tau)}\sigma_{0\tau} D_\tau. \end{aligned}$$

For example, consider the CRRA utility function, if $u(\tau, c) = e^{-(\tau - t)\rho}\left(c^{1-\gamma} - 1\right)/(1-\gamma)$, and $m = 1$. Then,

$$r_\tau = \rho + \gamma g_\tau - \frac{1}{2}\gamma(\gamma + 1)\sigma_{0\tau}^2, \quad \lambda_\tau = \gamma\sigma_{0\tau}.$$

Appendix 2 performs Walras's consistency tests regarding Eq. (5.117).

5.6.6 Continuous time Consumption-CAPM

[In progress]

We summarize,

$$S_{i\tau} = \mathbb{E}_t\left(\frac{S_{0\tau}}{S_{0T}}S_{iT} + \int_\tau^T \frac{S_{0\tau}}{S_{0u}}D_{iu}du\right) = E_t\left(\frac{m_{t,T}}{m_{t,\tau}}S_{iT} + \int_\tau^T \frac{m_{t,s}}{m_{t,\tau}}D_{iu}du\right),$$

where the second equality holds by the same arguments leading to Eq. (5.109). Replacing the first order condition in (5.111), and the equilibrium conditions in Eq. (**??**), we obtain the continuous time version of the Consumption-CAPM,

$$S_{i\tau} = E_t\left(\frac{u'(\underline{S}_T)}{u'(D_\tau)}S_{iT} + \int_\tau^T \frac{u'(D_s)}{u'(D_\tau)}D_{is}ds\right), \quad i = 0, 1, \cdots, m.$$

For example, the price of a pure discount bond, is,

$$b_\tau \equiv E_t\left(\frac{u'(\underline{S}_T)}{u'(D_\tau)}\right) = E_t\left(\frac{m_{t,T}}{m_{t,\tau}}\right),$$

where $m_{t,\tau}$ is as in Eq. (5.110).

5.6.7 Partial hedging in incomplete markets: introduction

[In progress]

5.7 Inaction: the economics of American options

5.7.1 Early exercise premiums: an introductory example

Consider two gambles on outcomes of rolling a dice. The first gamble is very simple: we draw a dice and receive a payoff equal to $\pi \in \{1, \cdots, 6\}$, such that its fair value from the perspective

of a risk-neutral bettor is $v_0 \equiv \frac{1}{6}\sum_{j=1}^{6} j = 3.5$.[25] The second gamble is more complex: we are now allowed to draw the dice three times and decide to earn the payoff $\pi \in \{1, \cdots, 6\}$ and, then, stop the game, at any of these draws. The design of the second gamble clearly aims to increase the chances of better outcomes, and is expected to be valued more than the first—the right to "exercise the option" to restart the game two times more has a positive value. For example, in case the outcome in any of the draws is $j = 1$, we would for sure like to draw the dice again and, in case the outcome is $j = 6$, we would for sure like to stop drawing the dice. More generally, now, we wish to find a state-contingent rule that maps states to a decision to stop the game.

We assume that we want to maximize the expected payoff, $\sup_{\tau\in\{1,2,3\}} \mathbb{E}(\pi_\tau)$, where π_τ denotes the payoff obtained on trial τ. We proceed recursively. Each time we draw the dice, we stop, provided the outcome is favourable enough, in that it is higher than the outcome we would expect from the next draw. Mathematically, the value of the game in each state is $\max\{j, \bar{p}'\}$, where $j \in \{1, \cdots, 6\}$ and $\bar{p}'$ denotes the expected payoff from drawing the dice another time.

The next table provides the values of the second gamble on the first and second draws, for each state. Let us consider Panel B. If the outcome after the second draw is 1 or 2 or 3, we draw the dice again, because the expected payoff from a third (a final) draw, $\bar{p}_3$ say, is obviously $\bar{p}_3 = v_0$, and is higher than that we would currently obtain, $\bar{p}_3 > 3$. However, we stop this gamble after the second draw, provided we observe an outcome equal to 4 or 5 or 6, as any of these outcomes is higher than $\bar{p}_3$. By a similar reasoning, the value of the gamble at the first draw in state j is, $p_1(j) = \max\{j, \bar{p}_2\}$, where $\bar{p}_2 = \frac{1}{6}(3.5 \times 3 + 4 + 5 + 6) = 4.25$, leading to Panel A of the table. The value of the gamble is, then, the expected payoff from Panel A, $\bar{p}_1 = \frac{1}{6}(4.25 \times 4 + 5 + 6) = 4.6667$.

Panel A: Value after the 1^{st} draw

state j	value at j, $p_1(j) = \max\{j, \bar{p}_2\}$	decision
1	4.25	draw again
2	4.25	draw again
3	4.25	draw again
4	4.25	draw again
5	5	stop
6	6	stop

Panel B: Value after the 2^{nd} draw

state j	value at j, $p_2(j) = \max\{j, \bar{p}_3\}$	decision
1	3.5	draw again
2	3.5	draw again
3	3.5	draw again
4	4	stop
5	5	stop
6	6	stop

The value of the first gamble is less than that of the second, $v_0 < \bar{p}_1$. We may interpret the second gamble as an "American option," which we can exercize at any time before its expiration, say after having seen any of the first two draws. The first gamble is, instead, interpreted as a "European option," with a payoff that only links to the third draw. The difference between the values of the two gambles is a premium we need to pay for the optionality to exercise early, such that in general, the value of the gamble after t draws is:

$$p_t(j) = \max\{j, \mathbb{E}(p_{t+1})\}, \quad p_T = \pi \in \{1, \cdots, 6\},$$

where we consider the unconditional expectation $\mathbb{E}(\cdot)$ due to the fact that the draws of the dice are independent and identically distributed.

[25] This gamble is, thus, the risk-neutral counterpart to the gamble described in Appendix 3 of Chapter 1.

Note that the "exercise boundaries" are $j \geq 5$ after the first draw and $j \geq 4$ after the second, meaning that the region where it is optimal to stop widens as the game goes through—the value to wait obviously decreases as the expiration approaches. Moreover, these boundaries are obviously endogenous, in that we calculate them as a part of the decision process. Finally, the boundaries are time-varying in this *free-boundary* problem, although in the next section, we shall consider infinite horizon decision problems where the free-boundary to be selected is constant.

5.7.2 Gambles and securities again

Chapter 2 explains a distinction between gambles and securities is that securities are traded while gambles are not, and that the value of gambles typically depends on supply and demand. For example, the evaluation in the previous section relies on the assumption that the bettor is risk-neutral. Alternatively, we could consider a risk-averse bettor, for example, one with logarithmic utility. We assume that this individual wants to maximize his expected utility, $\sup_{\tau\in\{1,2,3\}} \mathbb{E}\left(\ln \pi_\tau\right)$, even if we there might be other decision making models. His value function at time t and state j is equal to

$$\nu_t\left(j\right) = \max\left\{\ln j, E_t\left(\nu_{t+1}\right)\right\}.$$

The value function to be expected from a third draw is $\bar{\nu}_3 \equiv \frac{1}{6}\sum_{j=1}^{6} \ln j = 1.0965$, and because $\ln 2 < \bar{\nu}_3$ and $\ln 3 > \bar{\nu}_3$, the decision after the second draw is to draw again only when the draw delivers particularly poor outcomes, $j = 1$ or 2. Likewise, the value expected from drawing the dice a second time is $\bar{\nu}_2 \equiv \frac{1}{6}\left(2 \times 1.0965 + \sum_{j=3}^{6} \ln j\right) = 1.3465$, such that, and given that $\ln 3 < \bar{\nu}_2$ and $\ln 4 > \bar{\nu}_2$, the decision after the first draw is to stop as soon as the state is particularly favourable, i.e., when j is at least as large as 4. Therefore, the value of this gamble from the perspective of the log-utility bettor is $\bar{\nu}_1 = \frac{1}{6}\left(3 \times 1.3465 + \sum_{j=4}^{6} \ln j\right) = 1.4712$.

Note how more cautious this risk-averse individual is, compared to the risk-neutral. The risk-averse individual wants to continue this game only when the draw outcome is larger than three. We can quantify how much (insurance) risk-premium (see Section 1.3.3 in Chapter 1) is contained in the value of the gamble. First, the certainty equivalent of the gamble is $\hat{\nu} : \ln \hat{\nu} = 1.4712$, i.e., $\hat{\nu} = 4.3545$. It is lower than the value $\bar{p}_1 = 4.6667$ given by risk-neutral gambler. The difference is a risk-premium. Still, the risk-averse individual acknoweledges this gamble carries a strictly positive early exercise premium equal to $\hat{\nu} - \hat{\nu}_0 = 1.3608$, where $\hat{\nu}_0 = 2.9937$ is the certainty equivalent for the one-draw case.

Panel A: Value after the 1^{st} draw

state j	value at j, $\nu_1\left(j\right) = \max\{\ln j, \bar{\nu}_2\}$	decision
1	1.3465	draw again
2	1.3465	draw again
3	1.3465	draw again
4	$\ln 4$	stop
5	$\ln 5$	stop
6	$\ln 6$	stop

Panel B: Value after the 2^{nd} draw

state j	value at j, $\nu_2\left(j\right) = \max\{\ln j, \bar{\nu}_3\}$	decision
1	1.0965	draw again
2	1.0965	draw again
3	$\ln 3$	stop
4	$\ln 4$	stop
5	$\ln 5$	stop
6	$\ln 6$	stop

As explained, the reason a gamble like this is so relying on factors such as risk-aversion is that the asset we consider ("throwing a dice") is not traded and, as such, does not satisfy the martingale restriction. The next section develops an evaluation framework for a traded asset with a payoff relating to another traded asset, such that standard risk-neutral pricing applies.

5.7.3 Real options theory

We built upon the intuition developed through the example of the previous section, and derive a continuous-time approximation to the solution of the resulting problem, by hinging upon a series of heuristic arguments. We consider an "American option," i.e. an option that we can exercise at any time before the expiry date, T. Once the option is exercized, it yields a payoff equal to a function of the underlying asset price, say $\psi(S_t)$. Let $\mathcal{C}_t$ be the price of an American option as of time t. In discrete time, we have:

$$\mathcal{C}_t = \max\left\{\psi(S_t), e^{-r\Delta t}\mathbb{E}_t\left[\mathcal{C}_{t+\Delta t}\right]\right\}.$$

We assume that the nature of the option, summarized by the payoff $\psi(S_t)$, is such that there are two regions, a stopping region and a continuation region, defined as follows:

(i) *Stopping region*, where time-to-maturity and the price of the asset underlying the option are such that it is optimal to exercise, $\mathcal{C}_t = \max\left\{\psi(S_t), e^{-r\Delta t}\mathbb{E}_t\left[\mathcal{C}_{t+\Delta t}\right]\right\} = \psi(S_t)$, in which case, of course, $\mathcal{C}_t \geq e^{-r\Delta t}\mathbb{E}_t\left[\mathcal{C}_{t+\Delta t}\right]$. By rearranging terms

$$0 \geq e^{-r\Delta t}\frac{\mathbb{E}_t\left[\mathcal{C}_{t+\Delta t}\right]-\mathcal{C}_t}{\Delta t} - \frac{1-e^{-r\Delta t}}{\Delta t}\mathcal{C}_t. \tag{5.120}$$

The expected return on the option under the risk-neutral probability is less than that on a bank deposit, which further clarifies why it is optimal to exercise early. Naturally, the fact the option is yielding less than the safe interest rate is not an arbitrage. We could simply not short the derivative, as no one else is willing to buy it, as it is not optimal to do so.

(ii) *Continuation region*, where time-to-maturity and the price of the asset underlying the option are such that it is optimal to wait, $\mathcal{C}_t = \max\left\{\psi(S_t), e^{-r\Delta t}\mathbb{E}_t\left[\mathcal{C}_{t+\Delta t}\right]\right\} = e^{-r\Delta t}\mathbb{E}_t\left[\mathcal{C}_{t+\Delta t}\right]$, or

$$0 = e^{-r\Delta t}\frac{\mathbb{E}_t\left[\mathcal{C}_{t+\Delta t}\right]-\mathcal{C}_t}{\Delta t} - \frac{1-e^{-r\Delta t}}{\Delta t}\mathcal{C}_t. \tag{5.121}$$

The expected return on the option under the risk-neutral probability is the same as that on a bank deposit.

Note that the existence of these two regions is not guaranteed. For example, we shall see that it is never optimal to exercise early American calls written on assets that do not distribute dividends. When the two regions are, instead, well-defined, they define an exercise "envelope," a function of the asset price underlying the option and time-to-maturity. It is a "free boundary" problem: we need to find a boundary that triggers some action, in this case, exercising the option, and the boundary is free in that it is not given in advance as in the case of, say, the barrier options of the following section.

This problem can be quite complex, but sometimes, simplifies for those derivatives with an infinite expiry date, T. This simplification arises as in this case, the option price and, hence,

the envelope, only depends on the underlying asset price. Under this assumption, and the assumption that the price of the asset underlying the option is a geometric Brownian motion with volatility parameter σ, we have that the option price satisfies, in the limit $\Delta t \to 0$:

$$\text{Stopping region:} \qquad \mathcal{L}[\mathcal{C}] - r\mathcal{C} \leq 0 \text{ and } \mathcal{C} = \psi(S) \tag{5.122}$$

$$\text{Continuation region:} \qquad \mathcal{L}[\mathcal{C}] - r\mathcal{C} = 0 \tag{5.123}$$

where $\mathcal{L}[\mathcal{C}] = \frac{1}{2}\sigma^2 S^2 \mathcal{C}'' + rS\mathcal{C}'$. Eqs. (5.122)-(5.123) are the continuous time counterparts to Eqs. (5.120)-(5.121). To these two equations, we need to add a number of conditions, discussed in the two examples in the subsections below.

We can understand this evaluation of this asset under an equivalent angle, one cast in terms of an optimal stopping time problem,

$$\mathcal{C} = \sup_{\tau \geq 0} E\left(e^{-r\tau}\psi(S_\tau)\right). \tag{5.124}$$

That is, we would like to find the time at which it is best to undertake an action, in this case, exercizing the option. This time links to the free-exercise boundary S^*, say, as follows:

$$\tau^* = \inf_{\tau \geq 0}\{\tau : S_\tau = S^*\},$$

where S^* is determined endogenously. It is a "real option" problem, as it links nicely to the theory of evaluation of firms undertaking real economic activities as we shall say in Section 5.7. We now describe this problem in detail in standard cases. Finally, note that the free boundary is constant because the model is stationary and maturity is $T = \infty$. The free boundary may well be time dependent in case any of these assumptions does not hold. For example, the free boundary in the dice example of the previous section does well depend on time.

5.7.4 Perpetual puts

Consider a perpetual put, where $\psi(S) = (K - S)^+$, and the price p is, accordingly, a function of the underlying asset price S only. This price satisfies Eqs. (5.122)-(5.123), with some additional conditions and qualifications. First, we assume, and later verify, that there exists a value for the asset price, the free boundary, S_* say, such that, it is optimal to exercise the option whenever $S < S_*$. In other terms, Eqs. (5.122)-(5.123) can be written as:

$$\text{Stopping region } (S \leq S_*): \; p(S_*) = K - S_* \tag{5.125}$$

$$\text{Continuation region } (S > S_*): \; \mathcal{L}[p] - rp = 0 \tag{5.126}$$

where K is the strike price of the option. Eq. (5.125) is a "value-matching" condition. It ensures that the pricing function p is continuous as we move from the continuation region towards the stopping region.

Second, we require the following boundary condition:

$$\lim_{S \to \infty} p(S) = 0. \tag{5.127}$$

That is, as the asset price gets large, the value of the put option needs to approach zero, as the probability the derivative is ever exercised becomes negligible.

Finally, the pricing function, $p(S)$, satisfies the following "smooth-pasting" condition, obtained after taking the derivative in Eq. (5.125),

$$p'(S_*) = -1. \tag{5.128}$$

We conjecture that in the continuation region, the pricing function p that solves Eq. (5.126) has the form $p(S) = AS^{\gamma}$, for two constants A and γ. Plugging this guess into Eq. (5.126) reveals that actually, the pricing function satisfying it, has the following form:

$$p(S) = A_+ S^{\gamma_+} + A_- S^{\gamma_-}, \tag{5.129}$$

where A_+ and A_- are two constants, to be pinned down, $\gamma_+ = 1$ and $\gamma_- = -\frac{2r}{\sigma^2}$. To satisfy the boundary condition in Eq. (5.127), we need that $A_+ = 0$, which leaves $p(S) = A_- S^{\gamma_-}$. Evaluating this function at S_*, as in Eq. (5.125), and using the smooth pasting condition in Eq. (5.128), yields:

$$\begin{cases} p(S_*) = A_- S_*^{\gamma_-} = K - S_* \\ p'(S_*) = \gamma_- A_- S_*^{\gamma_- - 1} = -1 \end{cases} \tag{5.130}$$

The endogenous variables of this system are the two constants A_- and S_*. We have:

$$S_* = \frac{r}{r + \frac{1}{2}\sigma^2} K, \tag{5.131}$$

and $A_- = (K - S_*) S_*^{-\gamma_-}$, such that

$$p(S) = (K - S_*) \left(\frac{S}{S_*}\right)^{\gamma_-}.$$

A few comments are in order. First, Eq. (5.131) shows that the value to wait increases with σ^2. Second, when the short-term rate is zero, $S_* = 0$, meaning it is never optimal to exercise, and the option is worthless. Intuitively, in the stopping region, the expected return on the option under the risk-neutral probability is less than that on a bank deposit. When $r = 0$, this expected return is negative, which destroys the time-value of money argument underpinning early exercise.

5.7.5 Perpetual calls

As anticipated, not any payoff gives rise to well-defined stopping and continuation regions, such as those in Eqs. (5.122)-(5.123). For call options, where $\psi(S) = (S - K)^+$, it is never optimal to exercise early, when the underlying assets do not pay dividends. To illustrate, we literally follow the arguments in the previous subsection, and find that the call price, $c(S)$, satisfies,

$$\text{Stopping region } (S \geq S_*): \; c(S_*) = S_* - K \tag{5.132}$$

$$\text{Continuation region } (S < S_*): \; \mathcal{L}[c] - rc = 0 \tag{5.133}$$

The solution to Eq. (5.133) has the same functional form as that in Eq. (5.129), with the same values of γ_- and γ_+. However, due to the obvious conjectures about the location of the stopping and continuation regions in Eqs. (5.132)-(5.133), it satisfies the boundary condition $\lim_{S \to 0} c(S) = 0$, rather than $\lim_{S \to \infty} c(S) = 0$, as the put price does in Eq. (5.127). Therefore, and because $\gamma_+ = 1$, we must have that $c(S) = A_+ S$, or

$$c(S_*) = A_+ S_* = S_* - K,$$

where the second equality follows by the value matching condition in Eq. (5.132). Solving for S_* yields,

$$S_* = \frac{K}{1-A}.$$

As for the smooth-pasting condition we have that $c'(S_*) = A_+ = 1$, such that $S_* = \infty$. It is never optimal to exercise. In other words, there are no solutions to the counterparts to the two Eqs. (5.130). The call option price fails to simultaneously satisfy the value matching and smooth pasting condition.

5.7.5.1 With dividends

If the underlying asset distributes any payouts over the life of the option, say due to storage costs, or even dividends, the problem has, instead, a well-defined solution. Assume that under the risk-neutral probability, the underlying asset price satisfies,

$$\frac{dS_t}{S_t} = (r-q)\,dt + \sigma \hat{W}_t,$$

where q is the constant payout ratio, $q \equiv \frac{D_t}{S_t}$, with D_t denoting the instantaneous payout (e.g., a dividend), and as usual, $\hat{W}_t$ denotes a Brownian motion under the risk-neutral probability.

For this problem, Eqs. (5.122)-(5.123) reduce to,

$$\text{Stopping region } (S \geq S_*): \; c(S_*) = S_* - K \tag{5.134}$$

$$\text{Continuation region } (S < S_*): \; \mathcal{L}_q[c] - rc = 0 \tag{5.135}$$

where now, the infinitesimal generator is $\mathcal{L}_q[c] = \frac{1}{2}\sigma^2 S^2 c'' + (r-q)Sc'$. We can show the solution to Eq. (5.135) is,

$$c(S) = A_+ S^{\gamma_+} + A_- S^{\gamma_-},$$

where, $\gamma^+ = \frac{1}{\sigma^2}\left(\alpha + \sqrt{\alpha^2 + 2\sigma^2 r}\right)$, $\gamma_- = \frac{1}{\sigma^2}\left(\alpha - \sqrt{\alpha^2 + 2\sigma^2 r}\right)$, and $\alpha \equiv q - r + \frac{1}{2}\sigma^2$. Clearly, we have that $\gamma^- < 0$, and $\gamma^+ > 0$, such that the conjectures about the location of the stopping and continuation regions in Eqs. (5.134)-(5.135) deliver the boundary condition $\lim_{S\to 0} c(S) = 0$, and then,

$$c(S) = A_+ S^{\gamma_+}. \tag{5.136}$$

The value matching condition and the smooth pasting conditions equivalent to the two Eqs. (5.130) are, now

$$\begin{cases} c(S_*) = A_+ S_*^{\gamma+} = S_* - K \\ c'(S_*) = \gamma_+ A_+ S_*^{\gamma_+ - 1} = 1 \end{cases}$$

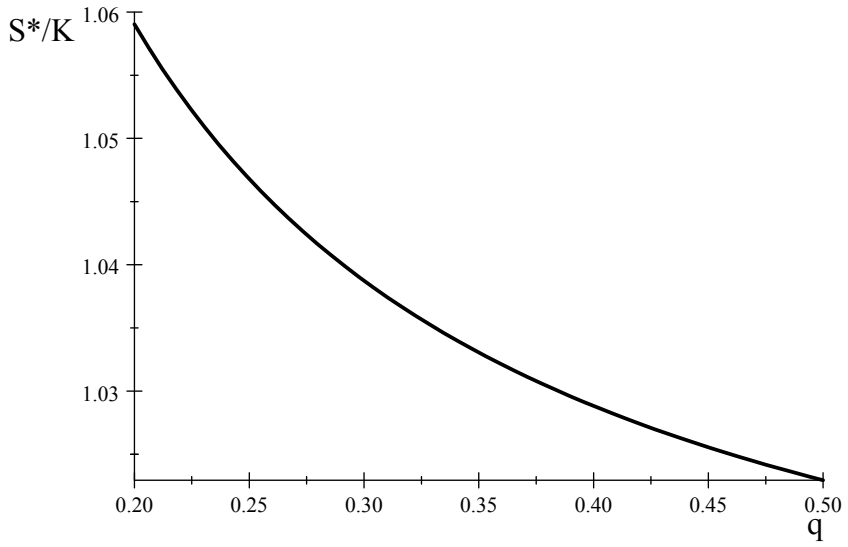
The solution to this system is,

$$A_+ = (S_* - K)\, S_*^{-\gamma^+}, \quad S_* = \frac{\gamma^+}{\gamma^+ - 1} K$$

such that the call option price in Eq. (5.136) can be written as:

$$c(S) = (S_* - K)\left(\frac{S}{S_*}\right)^{\gamma^+}.$$

The following picture depicts the triggering ratio, $\frac{S^*}{K}$, after which exercise is triggered as a function of q, when $r = 1\%$ and $\sigma = 15\%$.



The triggering ratio blows up as the payout ratio shrinks to zero. In general, the expected optimal stopping time is inversely related to the payout ratio q.

5.7.5.2 American calls and incomplete markets

American call options might be worth even in the absence of dividends paid out by the underlying over their life, in the presence of incomplete markets. Consider the following example of a perpetual American call option written on the instantaneous variance (not an average expected volatility) of an asset return. Suppose this variance, v_t say, is solution to,

$$dv_t = \kappa(\theta - v_t)\,dt + \sigma v_t dW_t,$$

where κ, θ and σ are parameters, and W_t is a Brownian motion under the risk-neutral probability. It is easy to show that the American option price is solution to,

$$\begin{aligned} \text{Stopping region } (v \;&\geq\; v_*)\text{:}\;\; c(v_*) = v_* - \bar{v} \\ \text{Continuation region } (v \;&<\; v_*)\text{:}\;\; \mathcal{L}[c] - rc = 0 \end{aligned}$$

where $\bar{v}$ is the strike price of the option, and

$$\mathcal{L}[c] = \kappa(\theta - v)\frac{\partial c}{\partial v} + \frac{1}{2}\sigma^2 v^2 \frac{\partial^2 c}{\partial v^2}. \tag{5.137}$$

The smooth pasting condition is,

$$c'(v^*) = 1,$$

and can be satisfied, because the fact v is not traded makes the drift of v show up in the infinitesimal generator of Eq. (5.137)—the fact v is not traded makes its drift under the risk-neutral probability play a role similar to dividends and leads to a positive economic value for c in this context.

5.8 Further topics on real options and controlled Brownian motions

5.8.1 Irreversible investments and the decision to invest

We discuss a classical model of the optimal time at which a firm should start to sell its output due to McDonald and Siegel (1986). Assume this firm has access to a technology such that at each instant of time τ, an output x_τ could be produced (and sold), which is solution to,

$$\frac{dx_\tau}{x_\tau} = \mu d\tau + \sigma dW_\tau,$$

where μ and σ are constants and W_τ is a standard Brownian motion. The idea underlying the model is that the firm begins to produce only once x_τ has reached a certain threshold, and that the production cannot be interrupted once it starts—the investment is irreversible.

Note that if the firm decides to begin selling output for x_τ at τ, its NPV is $\frac{x_\tau}{r-\mu} - c$, where r is the constant discounting rate, which needs to be larger than output growth to ensure finite NPV, $r > \mu$, and c is an initial constant cost incurred while starting the production. Therefore, the value of the firm that holds the potentiality to start produce at some future time τ when current potential output is x_0, is,

$$G(x_0) = \sup_{\tau \geq 0} E\left(e^{-r\tau}\psi(x_\tau)\right), \quad \psi(x) \equiv \frac{x}{r-\mu} - c.$$

That is, the value of the firm is formally the same as that of a perpetual American option in (5.124). Accordingly, we conjecture that the optimal stopping time is determined as the first time x_τ crosses a free-boundary from below, x^o say. In the continuation region,

$$0 = \mathcal{L}G(x) - rG(x) \quad \text{for } x < x^o, \tag{5.138}$$

where $\mathcal{L}G = \mu x G' + \frac{1}{2}\sigma^2 x^2 G''$. At the free-boundary,

$$G(x^o) = \frac{x^o}{r-\mu} - c. \tag{5.139}$$

The solution to Eq. (5.138) is, $G(x) = A_1 x^{\gamma_1} + A_2 x^{\gamma_2}$, where it is easy to show that $\gamma_1 < 0$ and $\gamma_2 > 1$ assuming as we are that that $r > \mu$, and A_1 and A_2 are constants to be determined. Note that it must be that $A_1 = 0$, for otherwise the value function would increase disproportionately as the output potential becomes arbitrarily small. Therefore, the economically viable solution to Eq. (5.138) is the bounded value function,

$$G(x) = A_2 x^{\gamma_2} \quad \text{for } x < x^o. \tag{5.140}$$

The free-boundary x^o and the constant A_2 can now be determined by requiring that the two following conditions hold at the threshold x^o: (i) the value matching condition, i.e. the value function in Eq. (5.140) is the same as that in (5.139), and (ii) the smooth pasting condition, the derivatives agree. The result is:

$$x^o = \frac{\gamma_2(r-\mu)}{\gamma_2 - 1}c, \quad A_2 = \frac{(x^o)^{1-\gamma_2}}{\gamma_2(r-\mu)}.$$

That is, the firm should start produce immediately, unless it had to pay a cost c in which case the threshold to start produce increase proportionally with c as the previous expression for x^o suggests.

5.8.2 A model of determination of exchange rates in target zones

We provide one instance of policy intervention that can "force" asset prices to experience less long-term volatility than implied by economic fundamentals: an agreement between central banks to do "whatever it takes" to maintain the exchange rate within a given band. The main assumption underlying the model is that monetary intervention are fully credible and that interventions occur marginally, in a way that the fundamentals behave as processes with reflecting barriers, just as in the dividend example given in Section 5.2.6.4. Below, we shall discuss issues arising when the policy is not fully credible, and the subsequent possibility of speculative attacks.

Consider, first, the classical model of exchange rate determination, in which the exchange rate is tied down to the consumer price differentials between two countries through the purchasing power parity equation, $s_t = p_t - p_t^*$, where s_t denotes the log-spot exchange rate, p_t is the domestic log-price level at time t, and p_t^* is its foreign counterpart. Money demand increases with log-output y_t and decreases with the nominal interest rate i_t in a way that in equilibrium, log-money supply is $m_t = p_t + ay_t - \alpha i_t + v_t$, where a and α are constants, and v_t is some stochastic disturbance—α is thus the semi-elasticity of money demand to to interest rates. An analogous, "starred" equation holds in the foreign country. Finally, the covered interest rate parity holds,

$$i_t = i_t^* + \frac{E\left(ds_t\right)}{dt}.$$

We can put all together and conclude that the exchange rate is solution to,

$$s_t = x_t + \alpha \frac{E\left(ds_t\right)}{dt}, \tag{5.141}$$

where $x_t \equiv m_t - m_t^* - a\left(y_t - y_t^*\right) - \left(v_t - v_t^*\right)$. Assume that were the exchange rate free to fluctuate, the fundamentals, x_t, would follow an arithmetic Brownian motion

$$dx_t = \mu dt + \sigma dW_t,$$

where μ and σ are constants, and W_t is a Brownian motion. It is immediate to see that this assumption regarding the fundamentals, Eq. (5.141) and Itô's lemma, all imply that the equilibrium exchange rate, $s_t = s\left(x_t\right)$, is solution to the following differential equation:

$$s\left(x\right) = x + \alpha\left(\mu s'\left(x\right) + \frac{1}{2}\sigma^2 s''\left(x\right)\right), \tag{5.142}$$

which is solved by,

$$s_{\mathrm{F}}\left(x_t\right) \equiv \alpha\mu + x_t.$$

Note that exchange rate volatility does not play any role in this context. In fact, there are many additional exchange rate functions on top of $s_{\mathrm{F}}\left(x_t\right)$, which satisfy Eq. (5.142),

$$s\left(x_t\right) = s_{\mathrm{F}}\left(x_t\right) + B_1 e^{\beta_1 x_t} + B_2 e^{\beta_2 x_t}, \tag{5.143}$$

where $\beta_{1/2} = \frac{1}{\alpha\sigma^2}(-\alpha\mu \mp \sqrt{\alpha^2\mu^2 + 2\alpha\sigma^2})$, and B_1 and B_2 are two arbitrary constants. However, $s_{\mathrm{F}}\left(x_t\right)$ is the only solution that we can interpret as one of expected fundamentals consistent with Eq. (5.141), in that,

$$s\left(x_t\right) = E_t\left(\int_t^\infty e^{-\alpha^{-1}(\tau - t)}\left(\alpha^{-1}x_\tau\right)d\tau\right) = \alpha\mu + x_t \equiv s_{\mathrm{F}}\left(x_t\right).$$

How would exchange rates behave, once central banks are credible enough to modify the fundamentals and make the exchange rates lie within a given band? Krugman (1991) considers a model in which the exchange rate is maintained withing a target zone, say $s_t \in [\underline{s}, \bar{s}]$ for two given constants $\underline{s}$ and $\bar{s}$. The model's main assumption is that central banks control the fundamentals through "infinitesimal interventions," by injecting or withdrawing monetary base $m_t - m_t^*$, as soon as the fundamentals become too weak or too strong for the current exchange rate to be maintained within the band $[\underline{s}, \bar{s}]$. The fundamentals are then solution to,

$$dx_t = \mu dt + \sigma dW_t + dU_t - dL_t,$$

where U_t and L_t are continuous processes that only increase when the fundamentals reach hit some critical values $\underline{x}$ (U) and $\bar{x}$ (L) that ask for intervention, which will be determined below.

By arguments entirely analogous to those utilized while pricing assets with dividends that have reflecting barriers (see Eq. (5.55) in Section 5.2.6.4), we now have that the exchange rate is solution to,

$$ds(x_t) = \left(\mu s'(x_t) + \frac{1}{2}\sigma^2 s''(x_t)\right) dt + s'(x_t)\,\sigma dW_t + s'(x_t)\,dU_t - s'(x_t)\,dL_t.$$

Therefore, the exchange rate is still solution to the differential equation (5.142), and its general solution is therefore given by Eq. (5.143). We can now determine the two constants, B_1 and B_2. First, note that analogous to the reflecting barrier case in Section 5.2.6.4 (see Eq. (5.56)), to rule out arbitrage, we need to ensure that the following smooth pasting conditions hold at the critical values $\underline{x}$ and $\bar{x}$:

$$s'(\underline{x}) = 0 \quad \text{and} \quad s'(\bar{x}) = 0. \tag{5.144}$$

Moreover, at the boundaries, we need to ensure that

$$s(\underline{x}) = \underline{s} \quad \text{and} \quad s'(\bar{x}) = \bar{s}. \tag{5.145}$$

Eqs. (5.144)-(5.145) are four independent equations with four unknowns: the two constants of the exchange rate function B_1 and B_2, and the two critical values of the fundamentals that trigger intervention, $\underline{x}$ and $\bar{x}$, which are all implicitly determined.

[In progress: discuss the literature on speculative attacks & policy credibility]

5.8.3 Liquidity constraints and optimal dividend policy

Jeanblanc-Picqué and Shirayev (1995) consider a model in which a firm maximizes its value through a dividend policy that accounts for periods of inaction. Inaction occurs because of liquidity constraints. While it will be assumed that the firm capital has positive growth, there may be bad times in which it would not be optimal for the firm to distribute any dividends as this might trigger bankruptcy. This issue would not arise if shareholders were able to inject new funds or the firm could borrow from a bank or issue new securities.

Consider the simple frictionless case, in which the cash generated by the firm is,

$$dk_\tau = \mu d\tau + \sigma dW_\tau, \tag{5.146}$$

where μ and σ are parameters, and W_τ is a standard Brownian motion. The net present value of the firm that has initial cash equal to k_o is,

$$\text{NPV}(k_0) \equiv \max\left\{0, k_0 + \frac{\mu}{r}\right\}, \tag{5.147}$$

where r is the constant discounting rate.

Next, consider the case in which the firm is liquidity constrained. It will now be optimal to distribute dividends only after cash reaches a threshold, with the firm's capital being solution to,

$$dk_\tau = \mu d\tau + \sigma dW_\tau - dD_\tau, \tag{5.148}$$

where D_τ is a cumulative dividend process, which only increases once capital k_τ has reached a sufficiently high level k^o, to be determined below. It is assumed that $k_0 > 0$, that bankruptcy occurs whenever k_τ hits zero, and that the value of the firm is maximized by risk-neutral managers as:

$$V(k_0) = \sup E\left(\int_0^\infty e^{-r\tau} dD_\tau\right), \quad \text{s.t. Eq. (5.148).} \tag{5.149}$$

The evaluation in (5.148) quite differs from that in Eq. (5.147), as we now show, depending as it is on the dividend policy.

5.8.3.1 Bounded dividend policy

Assume, initially, that the dividend satisfies,

$$dD_\tau = \omega(k_\tau)\, d\tau, \quad \text{with} \;\; 0 \leq \omega(k) \leq \bar{d} < \infty. \tag{5.150}$$

The limiting case, $\bar{d} = \infty$, is considered below. By Eqs. (5.148)-(5.149), and (5.150), the problem of the firm is then,

$$V(k_0) = \sup E\left(\int_0^\infty e^{-r\tau} \omega(k_\tau)\, d\tau\right), \quad \text{s.t.} \;\; dk_\tau = (\mu - \omega(k_\tau))\, d\tau + \sigma dW_\tau, \tag{5.151}$$

such that the Bellman equation for the problem is,

$$\begin{aligned} 0 &= \sup_{0 \leq \omega(k) \leq \bar{d}} [LV(k) - rV(k) + \omega(k)] \\ &= \mu V'(K) + \frac{1}{2}\sigma^2 V''(k) - rV(k) + \sup_{0 \leq \omega(k) \leq \bar{d}} [\omega(k)(1 - V'(k))], \end{aligned} \tag{5.152}$$

where L denotes the usual infinitesimal generator of the constraint in (5.151).

The solution is,

$$\omega(k) = \begin{cases} 0, & \forall k : V'(k) > 1 \text{ (inaction)} \\ \bar{d}, & \forall k : V'(k) \leq 1 \text{ (dividend distribution)} \end{cases} \tag{5.153}$$

The details of the problem suggest the following interpretation of the solution: the firm should pay off as soon as the level of capital reaches a threshold level, at which the payout takes place at the maximal speed, $\bar{d}$. This interpretation needs to be corroborated by verifying that the marginal value of capital, $V'(k)$, is indeed decreasing in k. In this case, the interpretation of the optimal policy in (5.153) is quite neat: the firm pays out whenever the marginal value of capital is low (less than unity).

Substituting Eqs. (5.153) into Eq. (5.152), and conjecturing that V is concave (decreasing returns to capital) over the inaction region, yields that the firm does not distribute any dividends until k reaches a threshold k^o (say),

$$\begin{cases} 0 = \mu V'(k) + \frac{1}{2}\sigma^2 V''(k) - rV(k), & k < k^o \text{ (inaction region)} \\ 0 = \frac{1}{\bar{d}}\left(\mu V'(k) + \frac{1}{2}\sigma^2 V''(k) - rV(k)\right) + (1 - V'(k)), & k \geq k^o \text{ (payout region)} \end{cases} \tag{5.154}$$

It will be verified that V is indeed concave in the inaction region. It is straightforward to verify that the solutions to the two equations in (5.154), $V_I(k)$ and $V_P(k)$ say, take the following form:

$$V_I(k) = A_{I1}e^{\alpha_1 k} + A_{I2}e^{\alpha_2 k}, \quad \text{and} \quad V_P(k) = A_{P1}e^{\beta_1 k} + A_{P2}e^{\beta_2 k} + \frac{\bar{d}}{r},$$

for some coefficients A_{Ij} and A_{Pj} to be determined, and $\alpha_1 < 0$, $\alpha_2 > 0$, $\beta_1 < 0$ and $\beta_2 > 0$.

The bankruptcy condition stipulates that $V(0) = 0$, such that $A_{I1} = -A_{I2}$, and then,

$$V_I(k) = A_{I1}\left(e^{\alpha_1 k} - e^{\alpha_2 k}\right), \quad \text{for } k < k^o. \tag{5.155}$$

As for $V_P(k)$, note that because $\omega(k)$ is bounded by $\bar{d}$, then by Eq. (5.151), V_P must be positive and bounded by $\frac{\bar{d}}{r}$, which requires $A_{P2} = 0$. Therefore,

$$V_P(k) = A_{P1}e^{\beta_1 k} + \frac{\bar{d}}{r}, \quad \text{for } k \geq k^o. \tag{5.156}$$

To summarize, the value of the firm is given by Eq. (5.155) and Eq. (5.156), with three constants to be determined: the two coefficients, A_{I1}, P_{I1}, and the free-boundary, k^o. The constants are determined by imposing (i) $V_I(k^o) = V_P(k^o)$ (value matching), (ii) $V_I'(k^o) = V_P'(k^o)$ (smooth pasting), (iii) $V_I''(k^o) = V_P''(k^o)$ (super-contact). The super-contact condition makes the second order derivative continuous as required for an application of Itô's lemma to a function of a variable that takes values over the real line, $V(k)$. Finally, it can be verified that the thusly determined coefficients imply that V_I is positive, increasing and concave for all $k < k^o$, thereby validating the conjectures underlying the two equations in (5.154). The difference between $V(k)$ in Eqs. (5.155)-(5.156) and NPV(k_0) in Eq. (5.147) is a cost accounting for the financial frictions faced by the firm.

5.8.3.2 Reflected Brownian motions

Next, we consider the limiting case in which $\bar{d} = \infty$. In a subsection below, we argue that the firm's capital would behave as a reflected Brownian motion in this case: when capital is below the threshold k^o, it is an arithmetic Brownian motion as in Eq. (5.146), but as soon as capital reaches k^o, it is then reflected back to lower values, with the capital in excess distributed as a dividend. It is also assumed that if the firm had to begin with cash larger than the threshold, the part exceeding k^o would be paid off as a dividend. Heuristically, Eqs. (5.154) would hold in this case when,

$$0 = \mu V'(k) + \frac{1}{2}\sigma^2 V''(k) - rV(k) \quad \text{for } k < k^o, \qquad \text{and } V'(k^o) = 1, \tag{5.157}$$

with the bankruptcy condition, $V(0) = 0$, and

$$V(k) = k - k^o + V(k^o) \quad \text{for } k \geq k^o,$$

with the interpretation that were the firm to start with capital k_t higher than the threshold, the difference $k_t - k^o$ would be immediately paid out as a dividend. The super-contact condition is therefore, $V''(k^o) = 0$.

The solution to Eq. (5.157) that satisfies the bankruptcy condition, is,

$$V(k) = \max\left\{0, \frac{e^{\alpha_2 k} - e^{\alpha_1 k}}{\alpha_2 e^{\alpha_2 k^o} - \alpha_1 e^{\alpha_1 k^o}}\right\}, \tag{5.158}$$

where $\alpha_1 < 0$ and $\alpha_2 > 0$, just as in Eq. (5.155). The threshold, k^o is determined through the super-contact condition. It is: $k^o = \frac{\ln(\alpha_1^2) - \ln(\alpha_2^2)}{\alpha_2 - \alpha_1}$.

Similarly as in the previous subsection, the difference between $V(k)$ in Eq. (5.158) and $v(k)$ in Eq. (5.147) is a cost accounting for the financial frictions faced by the firm. We now provide a few technical but heuristic details regarding the connection of this problem with reflected Brownian motions.

5.8.3.3 Heuristic technical details on reflected Brownian motions

While details are provided below to fill the reader with additional insights regarding the nature of reflected Brownian motions, a standard and rigorous source of details is in Karatzas and Shreve (1991, p. 210-212). Define the *occupancy time* of a random process x_τ in a set A up to time t as:

$$m(A,t) = \int_0^t \mathbb{I}_{x_\tau \in A} d\tau,$$

where $\mathbb{I}.$ denotes the indicator function. It measures how much time x_τ spends in set A up to time t for a given realization on $[0,t]$. Below, it will be argued that there exists a function $\ell(x,t)$, called *local time* of x_τ at a given point x, such that,

$$m(A,t) = \int_A \ell(x,t)\, dx, \quad \ell(x,t) = \lim_{\epsilon \downarrow 0} \frac{1}{2\epsilon} m\left((x-\epsilon, x+\epsilon), t\right). \tag{5.159}$$

Note the following property. We have, for any fixed x^o, and using the definitions of ℓ and m, and that of the Dirac's function,

$$\begin{aligned}
\ell(x^o,t) &= \int \delta(x - x^o)\, \ell(x,t)\, dx \\
&= \int \delta(x - x^o) \lim_{\epsilon \downarrow 0} \frac{1}{2\epsilon} \int_0^t \mathbb{I}_{x_\tau \in (x-\epsilon, x+\epsilon)} d\tau dx \\
&= \int_0^t \int \delta(x - x^o)\, \delta(x_\tau - x)\, dx d\tau \\
&= \int_0^t \delta(x_\tau - x^o)\, d\tau.
\end{aligned} \tag{5.160}$$

Note, also, that

$$\begin{aligned}
m(A,t) &= \int_0^t \mathbb{I}_{x_\tau \in A} d\tau \\
&= \int_0^t \int \delta(x - x_\tau)\, \mathbb{I}_{x \in A} dx d\tau \\
&= \int \mathbb{I}_{x \in A} \int_0^t \delta(x - x_\tau)\, d\tau dx \\
&= \int_A \ell(x,t)\, dx,
\end{aligned}$$

where the last equality follows by Eq. (5.160). This is Eq. (5.159).

Next, consider the solution for ω in Eq. (5.153), which in the limiting case, $\bar{d} = \infty$, is $\omega(k_t) = \delta(k_t - k^o)$, where $\delta(\cdot)$ is Dirac's delta. To simplify the presentation, set $\mu = 0$ and

$\sigma = 1$, such that the firm's capital in Eq. (5.148) now becomes,

$$k_t = k_0 - \int_0^t \delta\left(k_\tau - k^o\right) d\tau + W_t = k_0 - \ell\left(k^o, t\right) + W_t, \tag{5.161}$$

where the second equality follows by the property of local time in (5.160).

We argue that Eq. (5.161) describes a Brownian motion reflected at k^o: each time k_τ hits k^o from below, it is reflected back. Consider the function $|x_\tau - x^o| \equiv (x_\tau - x^o)^+ + (x_\tau - x^o)^-$, where x_τ is a Brownian motion, and apply heuristically Itô's lemma,

$$-|x_t - x^o| = -|x_0 - x^o| - \int_0^t \text{sign}\left(x_\tau - x^o\right) dx_\tau - \int_0^t \delta\left(x_\tau - x^o\right) d\tau. \tag{5.162}$$

Eq. (5.162) is known as Tanaka's formula for Brownian local time (see, also, the Appendix). It is possible to show that the second term on the R.H.S. of this equation is a Brownian motion starting from zero (Karatzas and Shreve, 1991, p. 209), say $\hat{W}_t$ Moreover, the third term is local time, by Eq. (5.160). Therefore, Eq. (5.162) is,

$$-|x_t - x^o| = -|x_0 - x^o| + \hat{W}_t - \ell\left(x^o, t\right),$$

such that the two processes in (5.161) and (5.162) have the same distribution.

5.9 Portfolio constraints

This section analyzes portfolio choices in contexts of market imperfections such as market incompleteness or short sale constraints, or others. We keep on hinging upon the setup in Section 5.4, and we fix $m = d$ although then we allow for frictions, by restricting the vector of normalized portfolio shares in the risky assets, $p_t \equiv \pi_t / V_t^{x,\pi,c}$, to lie in a closed convex set $K \in \mathbb{R}^d$.

We follow the approach put forward by Cvitanić and Karatzas (1992), which consists in "embedding" the *constrained* portfolio choice of the investor into a set of *unconstrained* portfolio optimization problems. Under regularity conditions, in this set of unconstrained problems, there exists one, the solution to which is that of the original constrained portfolio problem. Therefore, the constrained portfolio problem is solved, once we solve for the unconstrained, which we can do through the martingale methods in Section 5.4. This approach is closely related to the minimax probability methods of He and Pearson (1991), briefly illustrated within the discrete-time framework with incomplete markets of Chapter 2. It is a systematic approach to implement solutions to consumption and portfolio policies.in a context of constrained portfolio selection, and generalizes results from He and Pearson (1991), which only relate to contexts with incomplete markets.

5.9.1 Technical background

The starting point is the definition of the *support function*,

$$\zeta(\nu) = \sup_{p \in K}(-p^\top \nu), \quad \nu \in \mathbb{R}^d, \tag{5.163}$$

and its *effective domain*,

$$\tilde{K} = \{\nu \in \mathbb{R}^d : \zeta(\nu) < \infty\}. \tag{5.164}$$

The set $\tilde{K}$ in Eq. (5.164) is also known as the *barrier cone* of $-K$, where K is the set of admissible portfolio choices.

Examples of the support function ζ in Eq. (5.163) are the *unconstrained case*: $K = \mathbb{R}^d$, in which case $\tilde{K} = \{0\}$ and $\zeta = 0$ on $\tilde{K}$; *prohibition of short-selling*: $K = [0, \infty)^d$, in which case $\tilde{K} = K$ and $\zeta = 0$ on $\tilde{K}$, or: *incomplete markets*: $K = \{p \in \mathbb{R}^d : p_{m+1} = \cdots = p_d = 0\}$ (i.e. the first M assets can only be traded), in which case $\tilde{K} = \{\nu \in \mathbb{R}^d : \nu_1 = \cdots = \nu_m = 0\}$ and $\zeta = 0$ on $\tilde{K}$, which is the case covered by He and Pearson (1991).

We illustrate these cases in the $\mathbb{R}^2$ case, as in Figures 4.1 and 4.2 below. Given a cone K, interpreted as before as the set of admissible portfolio choices, consider the *polar cone* of $-K$, defined as $\tilde{K}_- = \{\nu \in \mathbb{R}^d : p^\top \nu \geq 0, \forall p \in K\}$, illustrated in Figure 4.1 in the case $d = 2$. We emphasize $\tilde{K}_-$ is the polar cone of $-K$, not the barrier cone of $-K$ defined in Eq. (5.164).

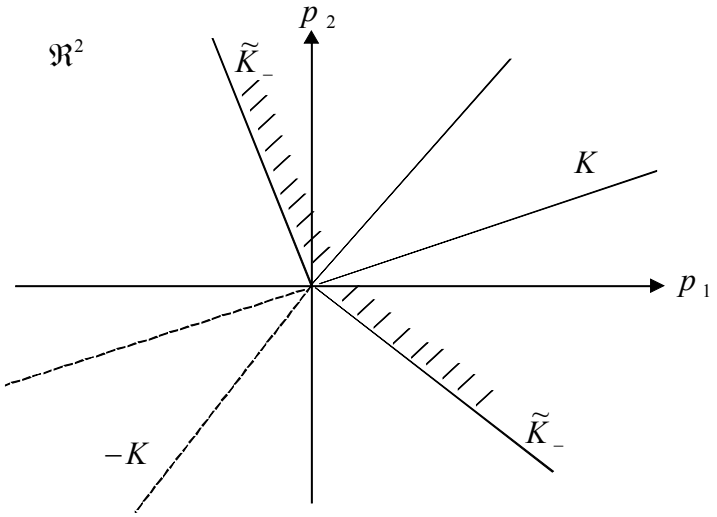


FIGURE 4.1

Consider, then, again, *prohibition of short-selling*, in which case $K = [0, \infty)^2$. It is easy to see that in this case, $\tilde{K}_- = K$ and that

$$\zeta(\nu) = \sup_{p \in [0,\infty)^2} (-p^\top \nu) = \begin{cases} 0, & \text{if } \nu \in \tilde{K}_- = [0, \infty)^2 \\ \infty, & \text{otherwise} \end{cases}$$

All in all, $\zeta(\nu)$ is finite only on $\tilde{K}_-$, such that the effective domain is $\tilde{K} = \tilde{K}_-$.

As for the *incomplete markets* case, assume the second asset is missing for sake of illustration. We have that $K = \{p \in \mathbb{R}^2 : p_2 = 0\}$, such that,

$$\begin{aligned} \tilde{K}_- &= \{\nu \in \mathbb{R}^2 : p_1\nu_1 + p_2\nu_2 \geq 0, \ \forall p_1 \in \mathbb{R} \text{ and } p_2 = 0\} \\ &= \{\nu \in \mathbb{R}^2 : p_1\nu_1 \geq 0, \ \forall p_1 \in \mathbb{R}\} \\ &= \{\nu \in \mathbb{R}^2 : \nu_1 = 0\} \end{aligned}$$

Panel C in Figure 4.2 illustrates this incomplete markets case. Panel A and B are examples of less severe forms of constraints, where markets are still complete. Panels A, B and C impose progressively tougher constraints, to the extent of the incomplete markets case in Panel C.

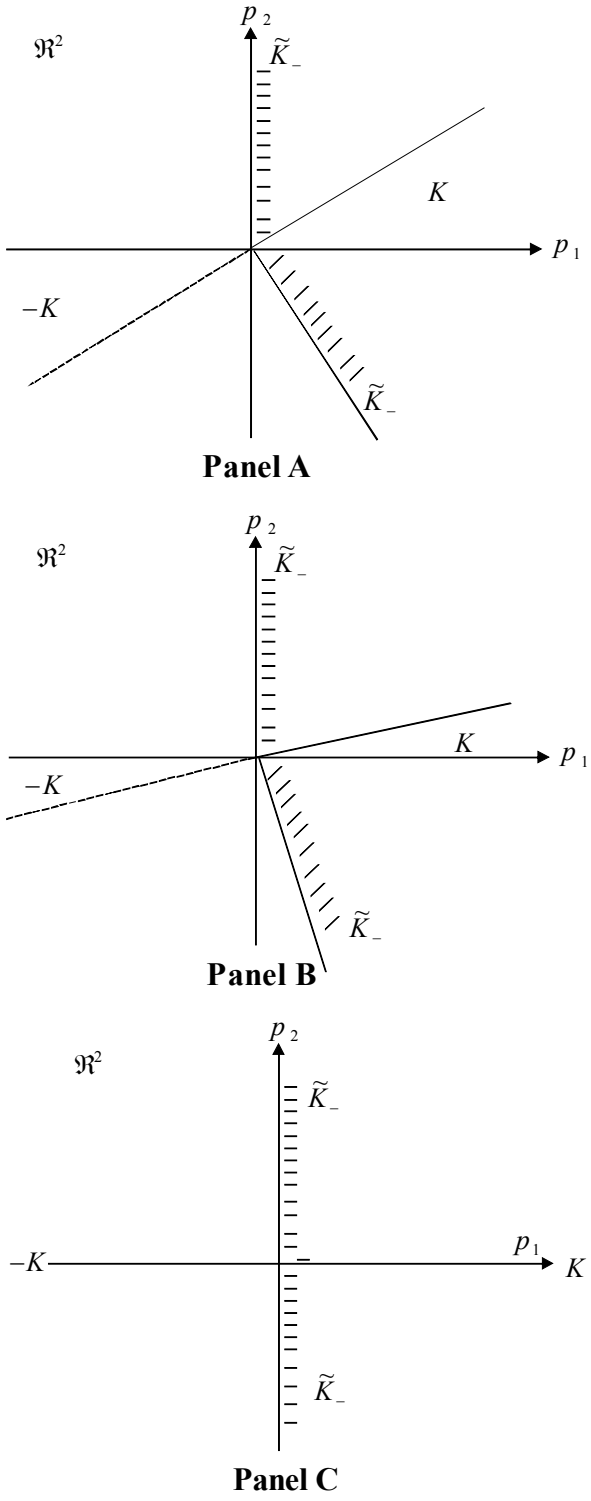


FIGURE 4.2: Portfolio constraints and incomplete markets.

As for the function $\zeta(\nu)$ in this context, we have that:

$$\zeta(\nu) = \sup_{p_1 \in \mathbb{R}} (-p_1 \nu_1) = \begin{cases} 0, & \text{if } \nu \in \tilde{K}_- \text{ (i.e. } \nu_1 = 0) \\ \infty, & \text{otherwise} \end{cases}$$

Again, $\zeta(\nu)$ is finite only on $\tilde{K}_-$, such that the effective domain is $\tilde{K} = \tilde{K}_-$.

5.9.2 Artificial markets

The role of the support function ζ in Eq. (5.163) is to tilt the dynamics of the price system in Section 5.4, as follows:

$$\frac{dS_{0t}}{S_{0t}} = r_t^\nu dt, \quad \frac{dS_{it}}{S_{it}} = \hat{a}_{it}^\nu dt + \sigma_{it} dW_t, \quad i = 1, \cdots, d, \tag{5.165}$$

where:

$$r^\nu \equiv r + \zeta(\nu), \quad \hat{a}_i^\nu \equiv \hat{a}_i + \nu + \zeta(\nu), \tag{5.166}$$

and $\hat{a}_i$ is as in Section 5.4.

The main result is as follows. Denote with $\text{Val}(x; K)$ the value of the problem faced by an investor facing a portfolio constraint $K \in \mathbb{R}^d$, when his initial wealth is x. Let $\text{Val}_\nu(x)$ be the corresponding value of the problem faced by an unconstrained investor in the market (5.165). Clearly, this value is just $\text{Val}_0(x)$ for the markets in Sections 5.4 and 5.5. Moreover, for each $\nu \in \mathbb{R}^d$, the unconstrained program the investor faces in the market (5.165), can be solved through martingale methods, using the unique risk-neutral probability Q^ν, equivalent to P, with Radon-Nikodym derivative equal to,

$$\eta_T^\nu \equiv \left.\frac{dQ^\nu}{dP}\right|_{\mathbb{F}_T} = \eta_T^0 \exp\left(-\int_0^T \left(\sigma_t^{-1}\nu_t\right)^\top dW_t - \frac{1}{2}\int_0^T \left\|\sigma_t^{-1}\nu_t\right\|^2 dt\right), \tag{5.167}$$

where η_T^0 is as in Eq. (5.70). As mentioned, He and Pearson (1991) is a special instance of this setting, which obtains once we assume incomplete markets, in which case the support function is $\zeta = 0$, as explained.

Under regularity conditions, we have that:

$$\text{Val}(x; K) = \inf_{\nu \in \tilde{K}} \left(\text{Val}_\nu(x)\right), \tag{5.168}$$

and optimal consumption and portfolio choices for this unconstrained problem are exactly those chosen by the investor constrained to have $p \in K$. Appendix 4 provides a heuristic derivation of Eq. (5.168).

In the context of log-utility functions, we have that,

$$\hat{\nu} = \arg\min_{\nu \in \tilde{K}} \left(2\zeta(\nu) + \left\|\lambda + \sigma^{-1}\nu\right\|^2\right), \tag{5.169}$$

where $\lambda = \sigma^{-1}(a - \mathbf{1}_d r)$. Part II illustrates applications of this result.

[Eq. (5.169) follows by applying martingale methods in a market where the Brownian motion is W_ν. It's actually very simple—see Cvitanić and Karatzas (1991, p. 790 + 797) + a portfolio $= \left(\sigma\sigma^\top\right)^{-1}(\nu + \hat{a} - \mathbf{1}_d r)$.

For example, in the incomplete markets case, we have that $\zeta = 0$, and the solution is then $\hat{\nu} = 0$ as originally explained by He and Pearson (1991). This is the case with only one asset—the multi-asset case is more complex and dealt with by Cvitanić and Karatzas (1991, p. 797).]

[In progress]

5.10 Jumps

Brownian motions are well suited to model the price behavior of liquid assets that are not subject to default risk. There is, however, a fair amount of interest in understanding asset prices as "discontinuous" processess. To illustrate, there may be events through which agents may mutate: for example, a liquidity shock (modeled as the realization of a "rare event") may make agents willing to sell the assets they hold, with important equilibrium dynamics (see e.g., Chapter 10). Fixed income instruments may actually well undergo liquidity dry-ups, or default, and price discontinuities may arise (see e.g., Chapter 14). This section is an introduction to a class of processes (Poisson processes), which is particularly useful in modeling all these issues.

5.10.1 Poisson jumps

Let (t,T) be a given time interval, and consider events in this interval with the following properties:

(i) The number of events arrivals on all disjoint intervals of (t,T) are independent.

(ii) For any two disjoint but equal intervals in (t,T), the probability of events' arrivals is the same on both intervals.

(iii) The probability that more than one events arrive simultaneously in any interval is zero.

We rely on the previous properties and pin down the functional form of $P_k(\tau - t)$, the probability that k events arrive during $\tau - t$. First, note that $P_k(\tau - t)$ satisfies

$$P_0(\tau + d\tau - t) = P_0(\tau - t) P_0(d\tau), \tag{5.170}$$

where

$$P_0(0) = 1, \quad P_k(0) = 0 \text{ for } k \geq 1. \tag{5.171}$$

Eq. (5.170) and the first condition in (5.171) are satisfied by $P_0(\tau) = e^{-v\tau}$, for some constant v and we take $v > 0$ to ensure that $P_0 \in [0,1]$. Furthermore, we have that

$$\begin{cases} P_1(\tau + d\tau - t) &= P_0(\tau - t) P_1(d\tau) + P_1(\tau - t) P_0(d\tau) \\ &\vdots \\ P_k(\tau + d\tau - t) &= P_{k-1}(\tau - t) P_1(d\tau) + P_k(\tau - t) P_0(d\tau) \\ &\vdots \end{cases} \tag{5.172}$$

The first equation in (5.172) can be rearranged as follows

$$\frac{P_1(\tau + d\tau - t) - P_1(\tau - t)}{d\tau} = -\frac{1 - P_0(d\tau)}{d\tau} P_1(\tau - t) + \frac{P_1(d\tau)}{d\tau} P_0(\tau - t).$$

For small $d\tau$, $P_1(d\tau) \approx 1 - P_0(d\tau)$ and $P_0(d\tau) = 1 - v d\tau + O(d\tau^2) \approx 1 - v d\tau$. Therefore, $P_1'(\tau - t) = -v P_1(\tau - t) + v P_0(\tau - t)$. By a similar reasoning,

$$P_k'(\tau - t) = -v P_k(\tau - t) + v P_{k-1}(\tau - t).$$

The solution to this equation is

$$P_k(\tau - t) = \frac{v^k (\tau - t)^k}{k!} e^{-v(\tau - t)}.$$

5.10.2 Interpretation

A Poisson model is one of *rare events*. Moreover, we have

$$E\left(\text{event arrival in } d\tau\right) = P_1\left(d\tau\right) = vd\tau,$$

and we usually refer to v as the *intensity* of event arrivals.

Let us provide additional intuition. Consider the expression for the probability of k "arrivals" in n trials predicted by a binomial distribution

$$P_{n,k} = \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!\,(n-k)!} p^k q^{n-k}, \qquad p, q > 0,\ p+q = 1,$$

where p is the probability of arrival in each trial. We want to model the probability p as a function of n, with the feature that $\lim_{n\to\infty} p(n) = 0$, which makes each arrival "rare," so to speak. One choice is $p\left(n\right) = \frac{a}{n}$, for some constant $a > 0$. Under this assumption, we have

$$\begin{aligned}
P_{n,k} &= \frac{n!}{k!\,(n-k)!} p(n)^k \left(1-p(n)\right)^{n-k} \\
&= \frac{n!}{k!\,(n-k)!} \left(\frac{a}{n}\right)^k \left(1-\frac{a}{n}\right)^{n-k} \\
&= \frac{n!}{k!\,(n-k)!} \left(\frac{a}{n}\right)^k \left(1-\frac{a}{n}\right)^{n} \left(1-\frac{a}{n}\right)^{-k} \\
&= \frac{n!}{n^k\,(n-k)!} \frac{a^k}{k!} \left(1-\frac{a}{n}\right)^{n} \left(1-\frac{a}{n}\right)^{-k} \\
&= \underbrace{\frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n}}_{k \text{ times}} \frac{a^k}{k!} \left(1-\frac{a}{n}\right)^{n} \left(1-\frac{a}{n}\right)^{-k}.
\end{aligned}$$

Taking the limits,

$$\lim_{n\to\infty} P_{n,k} \equiv P_k = \frac{a^k}{k!} e^{-a}.$$

Next, we split the interval $(\tau - t)$ into n subintervals of length $\frac{\tau - t}{n}$, and then make the probability of one arrival in each sub-interval proportional to each sub-interval length, as illustrated in Figure 4.3,

$$p(n) = v\frac{\tau - t}{n} \equiv \frac{a}{n}, \quad a \equiv v(\tau - t).$$

The Poisson model in the previous section is thus as that we consider here, with $n \to \infty$, which is continuous-time, as each sub-interval in Figure 5.1 shrinks to $d\tau$. The probability of one arrival in $d\tau$ is $vd\tau$, and this is also the expected number of events in $d\tau$:

$$\begin{aligned}
&E\left(\#\text{ arrivals in } d\tau\right) \\
&= \Pr\left(\text{one arrival in } d\tau\right) \times \text{one arrival} + \Pr\left(\text{zero arrivals in } d\tau\right) \times \text{zero arrivals} \\
&= \Pr\left(\text{one arrival in } d\tau\right) \times 1 + \Pr\left(\text{zero arrivals in } d\tau\right) \times 0 \\
&= vd\tau.
\end{aligned}$$

This heuristic construction helps us simulating a Poisson process. We may just simulate a Uniform random variable $U\left(0,1\right)$, with the continuous-time process being approximated by Y, where

$$Y = \begin{cases} 0 & \text{if} \quad 0 \le U < 1 - vh \\ 1 & \text{if} \quad 1 - vh \le U < 1 \end{cases}$$

and h is a discretization interval.

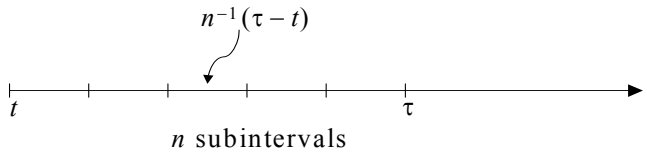


FIGURE 4.3. Heuristic construction of a Poisson process from a binomial distribution.

5.10.3 Properties and related distributions

We check that P_k is a probability. We have

$$\sum_{k=0}^{\infty} P_k = e^{-a} \sum_{k=0}^{\infty} \frac{a^k}{k!} = 1,$$

where the last equality follows by a McLaurin expansion of e^a. The mean of P_k is

$$\text{Mean} = \sum_{k=0}^{\infty} k \cdot P_k = e^{-a} \sum_{k=0}^{\infty} k \cdot \frac{a^k}{k!} = a. \tag{5.173}$$

A related distribution is the *exponential* (or Erlang) distribution. Remember, the probability of zero arrivals in $\tau - t$ predicted by the Poisson model is $P_0(\tau - t) = e^{-v(\tau - t)}$, from which it follows that:

$$G(\tau - t) \equiv 1 - P_0(\tau - t) = 1 - e^{-v(\tau - t)}$$

is the probability of at least one arrival in $\tau - t$. The function G can be also interpreted as the probability the first arrival occurred before τ, starting from t. The density function of G is:

$$g(\tau - t) = \frac{\partial}{\partial \tau} G(\tau - t) = v e^{-v(\tau - t)}.$$

The first two moments of the exponential distribution are:

$$\text{Mean} = \int_0^{\infty} x v e^{-vx} dx = v^{-1}, \quad \text{Variance} = \int_0^{\infty} \left(x - v^{-1}\right)^2 v e^{-vx} dx = v^{-2}.$$

The expected time of the first arrival occurred before τ starting from t equals v^{-1}. More generally, v^{-1} can be interpreted as the average time from an arrival to another.[26]

A more general distribution than the exponential is the Gamma distribution with density:

$$g_\gamma(\tau - t) = v e^{-v(\tau - t)} \frac{(v(\tau - t))^{\gamma - 1}}{(\gamma - 1)!}.$$

The exponential distribution obtains when $\gamma = 1$.

[26]Suppose arrivals are generated by Poisson processes, and consider the random variable "time interval elapsing from one arrival to next one." Let τ' be the instant at which the last arrival occurred. Then, the probability the time $\tau - \tau'$ which will elapse from the last arrival to the next is less than Δ is the same as the probability that during the time interval $\tau - \tau'$, there is at least one arrival.

5.10.4 Asset pricing implications

This section is a short introduction to modeling asset prices as being driven by Brownian motions and jumps processes. We model jumps by interpreting the "arrivals" in the previous sections as those events upon which a certain random variable experiences a jump of size $\mathcal{S}$, where $\mathcal{S}$ is another random variable with a fixed probability p. A simple model is:

$$dS_\tau = b(S_\tau)d\tau + \sigma(S_\tau)dW_\tau + \ell(S_\tau)\cdot\mathcal{S}\cdot dZ_\tau, \tag{5.174}$$

where b, σ, ℓ are given functions (with $\sigma > 0$), W is a standard Brownian motion, and Z is a Poisson process with intensity equal to v, i.e.

(i) $\Pr(Z_t) = 0$.

(ii) $\forall t \leq \tau_0 < \tau_1 < \cdots < \tau_N < \infty$, $Z(\tau_0)$ and $Z(\tau_k) - Z(\tau_{k-1})$ are independent for each $k = 1, \cdots, N$.

(iii) $\forall \tau > t$, $Z_\tau - Z_t$ is a random variable with Poisson distribution and expected value $v(\tau - t)$, i.e.:

$$\Pr(Z_\tau - Z_t = k) = \frac{v^k(\tau - t)^k}{k!}e^{-v\cdot(\tau-t)}.$$

Using the property in Eq. (5.173), we have that

$$E(Z_T - Z_t) = v(T - t)$$

In this framework, k is the number of jumps over the time interval $\tau - t$.[27] From this, we have that $\Pr(Z_\tau - Z_t = 1) = v(\tau - t)e^{-v\cdot(\tau-t)}$ and for $\tau - t$ small,

$$\Pr(dZ_\tau = 1) \equiv \Pr(Z_\tau - Z_t|_{\tau\to t} = 1) = v(\tau - t)e^{-v\cdot(\tau-t)}\big|_{\tau\to t} \simeq vd\tau.$$

More generally, the process $(Z_\tau - v(\tau - t))_{\tau\geq t}$ is a martingale.

Armed with these preliminary facts, we can provide a heuristic derivation of Itô's lemma for jump-diffusion processes. Consider any function f with enough regularity conditions, a rational function of time and S in Eq. (5.174), i.e. $f(\tau) \equiv f(S_\tau, \tau)$. Consider the following expansion of f:

$$\begin{aligned} df(\tau) = {} & Lf(S_\tau, \tau)d\tau + f_S(S_\tau, \tau)\sigma(S_t)dW_\tau \\ & + [f(S_\tau + \ell(S_\tau)\cdot\mathcal{S}, \tau) - f(S_\tau, \tau)]\cdot dZ_\tau. \end{aligned}$$

The first two terms in are the usual Itô's lemma terms, with L denoting the usual infinitesimal generator for diffusions. The third term accounts for jumps. If there are no jumps from time τ_- to time τ (where $d\tau = \tau - \tau_-$), then $dZ_\tau = 0$. If there is a jump then $dZ_\tau = 1$, and in this case f, as a "rational" function, needs also instantaneously jump to $f(S_\tau + \ell(S_\tau)\cdot\mathcal{S}, \tau)$. The jump will be exactly $f(S_\tau + \ell(S_\tau)\cdot\mathcal{S}, \tau) - f(S_\tau, \tau)$, where $\mathcal{S}$ is another random variable with

[27] For simplicity, we take v to be constant. If v is a deterministic function of time, we have that

$$\Pr(Z(\tau) - Z(t) = k) = \frac{\left(\int_t^\tau v(u)du\right)^k}{k!}\exp\left(-\int_t^\tau v(u)du\right), \quad k = 0, 1, \cdots$$

and there is also the possibility to model v as a function of the state: $v = v(q)$, for example. Cox processes.

a fixed probability measure. Clearly, if $f(S,\tau) = S$, we are back to the initial jump-diffusion model in Eq. (5.174).

To derive the infinitesimal generator for jumps-diffusion, $L^J f$ say, note that:

$$\begin{aligned} E(df) &= (Lf)\,d\tau + E\left[\left(f(S+\ell\mathcal{S},\tau) - f(S,\tau)\right)\cdot dZ_\tau\right] \\ &= (Lf)\,d\tau + E\left[\left(f(S+\ell\mathcal{S},\tau) - f(S,\tau)\right)\cdot v\cdot d\tau\right], \end{aligned}$$

or

$$L^J f = Lf + v\cdot\int_{\text{supp}(\mathcal{S})}\left[f(S+\ell\mathcal{S},\tau) - f(S,\tau)\right]p(d\mathcal{S}),$$

where $\text{supp}(\mathcal{S})$ denotes the support of $\mathcal{S}$. Therefore, the infinitesimal generator for jumps-diffusion is, simply, $L^J f$.

5.10.5 An option pricing formula

Merton (1976, JFE), Bates (1988, working paper), Naik and Lee (1990, RFS) are the seminal papers.

Obviously we cannot hedge. Make the argument. Use Itô's lemma and show that the strategy cannot hedge the jump component.

5.11 Continuous time Markov chains

Needed to model credit risk.

5.12 Appendix 1: An introduction to stochastic calculus for finance

5.12.1 Stochastic integrals

5.12.1.1 Motivation

Given is a Brownian motion $W_t \equiv W_t(\omega)$, $t \geq 0$, and the associated natural filtration $\mathbb{F}_t$. This appendix aims to explore the sense to be given to the "integral,"

$$I_t(\omega) \equiv \int_0^t f(s) dW_s(\omega), \tag{5A.1}$$

where f is a given function or, more generally,

$$I_t(\omega) \equiv \int_0^t g(s;\omega) dW_s(\omega),$$

where g is now a progressively $\mathbb{F}_t$-measurable function.

The motivation is so apparent from how useful many of the models in these lectures are, which build upon Brownian motions. Let us illustrate. Given a probability space (Ω, F, P) on which W is Brownian motion, and some $T < \infty$, let us write dW for the increment of W over an infinitesimal amount of time. In a heuristic sense, dW_t equals $W_{t+\Delta t} - W_t$ as $\Delta t \to 0$. We may think of the "increment" dW_t as being normally distributed: $dW_t \sim N(0, dt)$, and use this as a mean to build up a richer process X, say, solution to:

$$dX_t(\omega) = \mu_t(\omega) dt + \sigma_t(\omega) dW_t, \tag{5A.2}$$

for some $\mu_t(\cdot)$ and $\sigma_t(\cdot)$ to be defined later.

These processes are known as Itô's processes, as further explained below. The intuition on $\mu_t(\cdot)$ and $\sigma_t(\cdot)$ is as follows. Heuristically, we have that $E[dX_t(\omega)] = E[\mu_t(\omega)] dt + \sigma E(dW_t) = E[\mu_t(\omega)] dt$, such that $\mu_t(\cdot)$ is related to the instantaneous expected changes of dX. So this model is richer than Brownian motions because μ can be different from identically zero.

Think of X as an asset price process. No risk-averse individuals are likely to invest in the asset should its expected capital gain be zero, $E(dX) = 0$. Eq. (5A.2) also allows to thinl about the asset return variance to be expected over the next infinitesimal amount of time, being equal to $var(dX_t) = E[dX_t - E(dX_t)]^2 = E(\sigma dW_t)^2$, which turns out to equal $\sigma^2 dt$.

The "process" μ is called the *drift* and the "process" σ is called the *diffusion coefficient*, or the *volatility* of X. Clearly, the drift μ determines the trend, and the volatility determines the noisiness of X around that trend. Both drift and diffusion coefficients need to be adapted processes, as we shall explain. One example of drift and diffusion coefficients. Assume that $\mu \equiv W_t$ and $\sigma \equiv 0$. In this case, we have that: $dX_t = W_t dt$, which shows that X_t still fluctuates randomly. Here, μ is a stochastic process and so is X_t. Its infinitesimal variations can be predicted. But its further development cannot. We would say that X_t is *locally riskless* in this example.

Let us proceed with a more delicate example, relating to *strategies and trading gains*. Suppose that a stock price is just a Brownian motion. Assume it does not distribute dividends over some time-horizon of interest, and that we hold θ_t units of it at time t. What are our trading gains from 0 to t? Later, we shall argue that the expression, $\int_0^t \theta_s dW_s(\omega)$, is the answer to this question: intuitively, this expression is the sum of the instantaneous capital gains on the assets, multiplied by the units of the asset that are held. This expression is what is known as *stochastic integral*.

Why are we insisting in modeling asset prices through Brownian motions? As we shall see, Brownian motions are wild in some sense, i.e. they are of unbounded variation on any interval. So why don't we go for smoother processes? The answer is that "smoother" processes would give rise to arbitrage opportunities. Harrison, Pitbladdo and Schaefer (1984) show that in continuous time, asset prices must be "wild." Intuitively, if stock-prices are continuous in time and have finite variation, we could predict them over the immediate future, thus cashing-in the capital gains.

A Brownian motion is actually nowhere differentiable. Therefore, Eq. (5A.2) should be only understood as a shorthand for,

$$X_t(\omega) = X_0 + \int_0^t \mu_s(\omega)\, ds + \int_0^t \sigma_s(\omega)\, dW_s(\omega).$$

The question then is what does the "stochastic integral" $\int_0^t \sigma_s(\omega)\, dW_s(\omega)$ mean? In standard calculus, the integral can be defined from its differential. To anticipate, in stochastic calculus this is no longer the case, in that the stochastic integral is the real thing.

We now provide short reviews of the ordinary Riemann integral and the Riemann-Stieltjes integral, explaining that these two approaches to pathwise integration generically fail to provide a solid foundation to the "expression" $I_t(\omega)$ in Eq. (5A.1). To anticipate, the main issue relates to unboundedness of Brownian motions:

$$\forall \omega \in \Omega,\ \sup_{\tau} \sum_{i=1}^{n} \left| W_{t_i}(\omega) - W_{t_{i-1}}(\omega) \right| = \infty,$$

where the supremum is taken over all partitions of $[0, T]$. We shall state conditions on "how bounded" the integrator and integrands in $I_t(\omega)$ should be to make Riemann-Stieltjes theory go through. Unfortunately, these conditions are restrictive within the context of interest in finance. For example, Riemann-Stieltjes theory works when one takes $f(t) = 1$, or $f(t) = t$ in Eq. (5A.1). However, this theory doesn't hold in more general context.

5.12.1.2 Riemann

Given is $x \mapsto f(x)$, $x \in (0, 1)$. Consider two standard definitions. First, a *partition*, $\tau_n : 0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = 1$ and $\Delta_i = t_i - t_{i-1},\ i = 1, \cdots, n$. Second, an *intermediate partition*, σ_n : any collection of values y_i satisfying $t_{i-1} \leq y_i < t_i$, $i = 1, \cdots, n$. Then, for a given partition τ_n and intermediate partition σ_n, the Riemann sum is defined as:

$$S_n(\tau_n, \sigma_n) \equiv \sum_{i=1}^{n} f(y_i)\Delta_i.$$

It's a "weighted average" of the values taken by $f(\cdot)$.

Next, let $\text{Mesh}(\tau_n) \equiv \max_{i=1,\cdots,n} \Delta_i$, and consider $\text{Mesh}(\tau_n) \to 0$ by sending $n \to \infty$. If the limit, $\lim_{n\to\infty} S_n(\tau_n, \sigma_n)$, exists, and is independent of τ_n and σ_n, then it is called the *Riemann integral* of f on $(0, 1)$ and it is written:

$$\int_0^1 f(t)dt.$$

Two properties are worth mentioning:

1. *Linearity*: Given two constants c_1 and c_2, $\int_0^1 (c_1 f_1(t) + c_2 f_2(t))dt = c_1 \int_0^1 f_1(t)dt + c_2 \int_0^1 f_2(t)dt$.
2. *Linearity on adjacent intervals*: $\int_0^1 f(t)dt = \int_0^a f(t)dt + \int_a^1 f(t)dt$ for every $a \in (0, 1)$.

5.12.1.3 Riemann-Stieltjes

The main idea is to "integrate one function f with respect to another function g." One standard example relates to the computation of the expectation of a random variable with distribution function g. Heuristically, we have that:

$$\int_0^1 t dg(t) \approx \sum_i t_i \left[g(t_i) - g(t_{i-1}) \right].$$

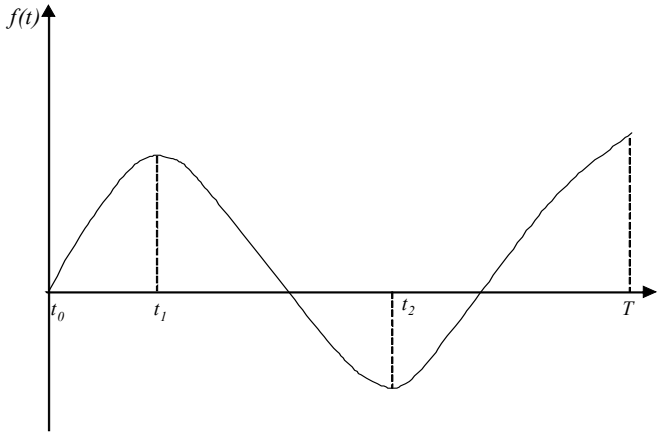
More generally, let us be given two functions f and g, and consider, again, the definitions of τ_n, σ_n given earlier. Let $\Delta g_i = g(t_i) - g(t_{i-1}),\ i = 1, \cdots, n$. The Riemann-Stieltjes sum is defined as:

$$S_n(\tau_n, \sigma_n) = \sum_{i=1}^{n} f(y_i)\Delta g_i.$$

Clearly the Riemann sum is a special case obtained with the identity function $g(t) = t$. Similarly as when proceeding with definition of the Riemann sum, the *Riemann-Stieltjes integral* of f with respect to g on $(0, 1)$ is the limit, $\lim_{n\to\infty} S_n(\tau_n, \sigma_n)$, provided it exists and is independent of τ_n and σ_n. It is written:

$$\int_0^1 f(t)\, dg_t.$$

The crucial issue now is how we can use Riemann-Stieltjes theory to define integrals of functions against Brownian motions: can interpret $\int_0^1 f_t dW_t(\omega)$ as a Riemann-Stieltjes integral, path by path, i.e. $\forall\omega \in \Omega$? The answer is in the negative, except in very special cases. Indeed, a natural example of an integral of functions with respect to Brownian motion is $I_t(\omega) \equiv \int_0^t f(s)dW_s(\omega)$. But what does this representation mean? We know that a ω-W_t path is not-differentiable. However, the main point, here, does not even relate to differentiability, but to a quite peculiar property, known as *unboundedness of Brownian motions*. To introduce this property, consider a certain function f, such as that depicted in the next picture.



Consider, then, its "first variation," defined as:

$$\begin{aligned}[f(t_1) - f(t_0)] - [f(t_2) - f(t_1)] + [f(T) - f(t_2)] &= \int_0^{t_1} f'(t)\, dt + \int_{t_1}^{t_2} \left(-f'(t)\right) dt + \int_{t_2}^{T} f'(t)\, dt \\ &= \int_{t_0}^{T} \left|f'(t)\right| dt.\end{aligned}$$

We can see that this first variation is a measure of the total amount of up and down motion of the path of the function f. We can formalize this reasoning as follows. Let f be a function of a real variable. Its variation in an interval $[a, b]$ is defined as

$$V_f([a, b]) = \sup \sum_{i=1}^{n} \left|f(t_i^{(n)}) - f(t_{i-1}^{(n)})\right|,$$

where the supremum is taken over the partitions $a \equiv t_0^{(n)} < t_1^{(n)} < \cdots < t_n^{(n)} \equiv b$. By the triangle inequality, $|f(x) - f(y)| = |f(x) - f(z) + f(z) - f(y)| \leq |f(x) - f(z)| + |f(z) - f(y)|$, the sums in $V_f([a, b])$ can only increase as we add more and more into the partition, such that,

$$V_f([a, b]) = \lim_{\text{mesh}\downarrow 0} \sum_{i=1}^{n} \left|f(t_i^{(n)}) - f(t_{i-1}^{(n)})\right|.$$

Next, consider the following definition.

DEFINITION 5A.1. *A real function h on $(0,1)$ has bounded p-variation, $p>0$, if*

$$\sup_{\tau}\sum_{i=1}^{n}|h(t_i)-h(t_{i-1})|^p<\infty,$$

where the supremum is taken over all partitions of $(0,1)$.

We have:

THEOREM 5A.2. *The Riemann-Stieltjes integral, $\int_0^1 f(t)dg(t)$, exists under the following conditions:*

(i) *The functions f and g don't have discontinuities at the same points.*

(ii) *f has bounded p-variation and g has bounded q-variation, with $\frac{1}{p}+\frac{1}{q}>1$, that is, f,g satisfy $\sup_\tau\sum_{i=1}^n|f(t_i)-f(t_{i-1})|^p<\infty$ and $\sup_\tau\sum_{i=1}^n|g(t_i)-g(t_{i-1})|^q<\infty$ with $\frac{1}{p}+\frac{1}{q}>1$.*

Now, it is well-known that almost every ω-W_t path has bounded p-variation for $p\geq 2$. And, as expected, unbounded p-variation for $p<2$, as further argued below. Consider, then, the integral, $\int_0^1 f_t dW_t(\omega)$, and suppose f is differentiable with bounded derivatives. By the mean value theorem, there exists a $K>0$ such that: $|f(t)-f(s)|\leq K(t-s)$ for $s<t$. Therefore, $\sup_\tau\sum_{i=1}^n|f(t_i)-f(t_{i-1})|$ $\leq K\sum_{i=1}^n(t_i-t_{i-1})=K$. That is, f has bounded p-variation, with $p=1$. By Theorem 5A.2, we now have that for almost every ω-W_t path, the Riemann-Stieltjes integral of f with respect to Brownian motions,

$$I_t(\omega)\equiv\int_0^t f(s)\,dW_s(\omega),$$

exists for every deterministic function f which is differentiable with bounded first-order derivative. For example, $f(t)=1$, or $f(t)=t$. We aren't done. Consider $f_t(\omega)=W_t(\omega)$ and, then:

$$I(W)(\omega)=\int_0^1 W_t(\omega)\,dW_t(\omega).$$

Let $p=2+\epsilon$, for some $\epsilon>0$. Hence $p=q=2+\epsilon$, and so $\frac{1}{p}+\frac{1}{q}=\frac{2}{2+\epsilon}<1$. The Riemann-Stieltjes theory doesn't work even with this simple example. This is where the theory of Itô's stochastic integrals comes in.

5.12.1.4 A digression on unboundedness of Brownian motions

Why do Brownian motions display unbounded variation? Consider the "Brownian tree" below.

$$\begin{array}{lll} & \nearrow^{1/2} & W_{\Delta t}=W_0+\Delta h \\ W_0 & & \\ & \searrow^{1/2} & W_{\Delta t}=W_0-\Delta h \end{array}$$

Time is Δt and space is Δh. In the Brownian tree, we must have,

$$\Delta h=\sqrt{\Delta t}. \tag{5A.3}$$

Indeed, and heuristically, we have that $var(\Delta W) = (\Delta h)^2$, which matched to $var(\Delta W) = \Delta t$, leaves precisely Eq. (5A.3). Therefore, $E(|\Delta W|) = \Delta h = \sqrt{\Delta t}$. Next let us chop a time interval of length t in $n \equiv \frac{t}{\Delta t}$ parts. The total expected length traveled by a Brownian motion is,

$$\frac{t}{\Delta t}\Delta h = \frac{t}{\Delta t}\sqrt{\Delta t} \to \infty \text{ as } \Delta t \to 0.$$

A more substantive proof is one for example of Corollary 2.5 p. 25 in Revuz and Yor (1999). A sketch of this proof proceeds as follows. We have:

$$\sum_i \left(W_{t_i} - W_{t_{i-1}}\right)^2 \leq \max_i \left|W_{t_i} - W_{t_{i-1}}\right| \cdot \sum_i \left|W_{t_i} - W_{t_{i-1}}\right|.$$

Moreover, $\max_i \left|W_{t_i} - W_{t_{i-1}}\right|$ converges to zero $\forall \omega \in \Omega$ because $W_\cdot$ is continuous, and by the Heine-Cantor theorem, continuous functions are uniformly continuous on finite intervals. Then, suppose that $W_{t_\cdot}$ has bounded variation, which would imply that

$$\forall \omega \in \Omega, \quad \sum_i \left(W_{t_i} - W_{t_{i-1}}\right)^2 \to 0, \quad \text{Mesh} \downarrow 0,$$

which is impossible, as we shall now argue. Indeed, in the next section, we shall establish that $L_n \equiv \sum_i \left(W_{t_i} - W_{t_{i-1}}\right)^2 \overset{q.m}{\to} t$, implying that $\text{p}\lim_n L_n = t$. Therefore, there exists a sequence $n_k : L_{n_k} \to t$ for all $\omega \in \Omega$. (Convergence in probability does not imply almost sure convergence, yet it implies that there exists a suitable subsequence n_k s.t. $\exists$ a.s. convergence, which is what we just need here.)

5.12.1.5 Itô

Let us begin with a first example, which can help grasp the nature of the issues under study. Consider

$$I(W)(\omega) = \int_0^1 W_t(\omega) dW_t(\omega).$$

Consider, then, the following Riemann-Stieltjes sum:

$$S_n = \sum_{i=1}^n W_{t_{i-1}} \Delta_i W, \quad \Delta_i W = W_{t_i} - W_{t_{i-1}}, \quad t_n = t.$$

where the intermediate partition relies on the left-end points $y_i = t_{i-1}$, $i = 1, \cdots, n$. Simple computations leave:

$$S_n = \frac{1}{2}\left[W_t^2 - Q_n(t)\right], \quad Q_n(t) \equiv \sum_{i=1}^n (\Delta_i W)^2.$$

The quantity $Q_n(t)$ is known as *Quadratic Variation*, a quite useful concept in financial econometrics. We have,

$$E[Q_n(t)] = \sum_{i=1}^n E(\Delta_i W)^2 = \sum_{i=1}^n \Delta_i = t.$$

Moreover, $var[(\Delta_i W)^2] = \Delta_i^2 var[(\frac{\Delta_i W}{\sqrt{\Delta_i}})^2] = 2\Delta_i^2$, where the last equality follows because $\frac{\Delta_i W}{\sqrt{\Delta_i}} \sim N(0,1)$, which implies that $\left(\frac{\Delta_i W}{\sqrt{\Delta_i}}\right)^2 \sim \chi^2(1)$. Hence,

$$var(Q_n(t)) = \sum_{i=1}^n var\left(\Delta_i W^2\right) = 2\sum_{i=1}^n \Delta_i^2 \leq 2\sum_{i=1}^n \text{Mesh}(\tau_n) \cdot \Delta_i = 2t \cdot \text{Mesh}(\tau_n) \to 0.$$

But $var\left(Q_n\left(t\right)\right) = E\left(Q_n(t) - E\left(Q_n\left(t\right)\right)\right)^2 = E\left(Q_n\left(t\right) - t\right)^2$. Therefore,

$$var\left(Q_n\left(t\right)\right) = E\left(Q_n(t) - t\right)^2 \to 0 \;\; t\text{-pointwise.}$$

This type of convergence is called *convergence in quadratic mean of* $Q_n(t)$ *to* t and is written $Q_n(t) \overset{q.m.}{\to} t$, as explained in more detail in the appendixes of Chapter 6. By the celebrated Chebyshev's inequality, convergence in quadratic mean implies convergence in probability:

$$\forall \delta > 0, \; \Pr\left\{|Q_n(t) - t| > \delta\right\} \leq \frac{E\left(Q_n(t) - t\right)^2}{\delta^2}.$$

Issues related to uniform convergence issues will be dealt with later.

To sumup, $\int_0^{\cdot} W_s(\omega) dW_s(\omega)$ doesn't exist as a Riemann-Stieltjes integral. Nevertheless, the previous facts suggest that a good definition of it could hinge upon the notion of a mean square limit, viz

$$S_n = \sum_{i=1}^{n} W_{t_{i-1}} \Delta_i W = \frac{1}{2}\left(W_t^2 - Q_n(t)\right) \overset{q.m.}{\to} \frac{1}{2}\left(W_t^2 - t\right),$$

or, as we shall explain,

$$S_n = \sum_{i=1}^{n} W_{t_{i-1}} \Delta_i W \overset{q.m.}{\to} \int_0^t W_s dW_s,$$

where $\int_0^t W_s dW_s = \frac{1}{2}\left(W_t^2 - t\right)$ has the Itô's sense.

Clearly, $\int_0^t W_s dW_s$ does not satisfy the usual Riemann-Stieltjes rule of integration. (For any smooth function f such that $f\left(0\right) = 0$, the Riemann-Stieltjes integral $\int_0^t f(u) df(u) = \frac{1}{2} f^2(t)$.) This doesn't work here because we have yet to see what the chain-rule for functions of ω-W_t is. This will lead us to the celebrated Itô's lemma, which shall confirm that $\int_0^t W_s dW_s = \frac{1}{2}\left(W_t^2 - t\right)$. This example vividly illustrated that standard integration methods fails. In fact, the timing of the integrands is quite critical. For example, in Riemann integration, the integrand can be evaluated at any point in the interval. If we apply this to the kind of integrals we are studying here we obtain, $\lim \sum_i f\left(W_{t_{i-1}}\right)\left(W_{t_i} - W_{t_{i-1}}\right)$ (for the left boundary) and $\lim \sum_i f\left(W_{t_i}\right)\left(W_{t_i} - W_{t_{i-1}}\right)$ (for the right boundary). But the two limits do not agree. The expectation of the first is zero (by the law of iterated expectations), while the expectation of the second is not necessarily zero. Finally, Riemann integration theory differs from the integration theory underlying the previous example because of the mode of convergence utilized in the two theories.

A short digression is in order. The so-called Stratonovich stochastic integral selects as points of the intermediate partion the central ones:

$$\tilde{S}_n = \sum_{i=1}^{n} f\left(W_{y_i}\right) \Delta_i W, \quad y_i = \frac{1}{2}\left(t_{i-1} + t_i\right).$$

For the Stratonovich integral, the usual Riemann-Stieltjes rule applies, yet the Stratonovich stochastic integral isn't Riemann-Stieltjes.

5.12.1.6 The Itô's stochastic integral for *simple* processes

Let $\mathcal{F}$ be the P-augmentation of the filtration of W. Consider $[0,T]$ and partitions $\tau_n : 0 = t_0 < t_1 < \cdots < t_n = T$, and the following definition:

DEFINITION 5A.3 (SIMPLE PROCESSES). *The process* $C = (C_t,\ t \in [0,T])$ *is* simple *if*

(i) *There exists a partition* τ_n *and a sequence of r.v.* Z_i, $i = 1, \cdots, n$, *s.t*

$$C_t = \begin{cases} Z_n, & \text{if } t = T \\ Z_i, & \text{if } t_{i-1} \leq t < t_i, \;\; i = 1, \cdots, n \end{cases}$$

(ii) *The sequence* (Z_i) *is* $\mathbb{F}_{t_{i-1}}$*-adapted,* $i = 1, \cdots, n$.

(iii) $E(Z_i^2) < \infty$ *all* i *(*L^2*).*

As an example, consider $C_t = W_{t_{n-1}}$, if $t = T$, and $C_t = W_{t_{i-1}}$, if $t_{i-1} \leq t < t_i$, $i = 1, \cdots, n$. Next, we have:

DEFINITION 5A.4. *The Itô's stochastic integral of a simple process* C *is,*

$$\begin{aligned}\int_0^T C_s dW_s &= \sum_{i=1}^{n} C_{t_{i-1}} \left(W_{t_i} - W_{t_{i-1}}\right) = \sum_{i=1}^{n} Z_i \left(W_{t_i} - W_{t_{i-1}}\right), \text{ on } [0,T] \\ \int_0^t C_s dW_s &= \sum_{i=1}^{k-1} C_{t_{i-1}} \left(W_{t_i} - W_{t_{i-1}}\right) + Z_k \left(W_t - W_{t_{k-1}}\right), \ t \in [t_{k-1}, t_k].\end{aligned}$$

with the notation $\sum_{i=1}^{0} m_i \equiv 0$.

It is a Riemann-Stieltjes sum of C with respect to Brownian motions evaluated at *left-end* points. Finally, we proceed with listing a set of useful properties.

PROPERTY 5A.5. $I_t(C) = \int_0^t C_s dW_s, t \in [0,T]$ *is a* $\mathbb{F}_t$*-martingale and has expectation equal to zero.*

Proof. Let us check that $I_t(C)$ is a $\mathbb{F}_t$-martingale. We have to check three conditions: (i) $E\left|I_t\left(C\right)\right| < \infty$, all $t \in [0,T]$; (ii) $I_t\left(C\right)$ is $\mathbb{F}_t$-adapted; (iii) $E\left[I_t\left(C\right)|\,\mathbb{F}_s\right] = I_s\left(C\right)$, $s < t$. Condition (i) follows by the isometry property to be introduced below. Condition (ii) is trivial. To show (iii), suppose, initially, that $s, t \in [t_{k-1}, t_k]$, $s < t$. We have:

$$\begin{aligned}I_t\left(C\right) &= \sum_{i=1}^{k-1} Z_i \left(W_{t_i} - W_{t_{i-1}}\right) + Z_k \left(W_t - W_{t_{k-1}}\right) \\ &= \sum_{i=1}^{k-1} Z_i \left(W_{t_i} - W_{t_{i-1}}\right) + Z_k \left(W_s - W_{t_{k-1}}\right) + Z_k \left(W_t - W_s\right) \\ &= I_s\left(C\right) + Z_k \left(W_t - W_s\right)\end{aligned}$$

$$\begin{aligned}E\left[I_t\left(C\right)|\,\mathbb{F}_s\right] &= E\left[I_s\left(C\right)|\,\mathbb{F}_s\right] + E\left[Z_k\left(W_t - W_s\right)|\,\mathbb{F}_s\right] \\ &= I_s\left(C\right) + Z_k E\left[\left(W_t - W_s\right)|\,\mathbb{F}_s\right] = I_s\left(C\right).\end{aligned}$$

The case $s \in [t_{l-1}, t_l]$ and $t \in [t_{k-1}, t_k]$, $l < k$ is proven similarly. Finally, $I_t(C)$ has zero expectation because it starts from the origin by the definition: $I_0(C) = 0 \Rightarrow E\left(I_t(C)\right) = 0$ all t. That is, $\forall t$, $E\left[I_t\left(C\right)\right] = E\left[I_0\left(C\right)\right] = I_0\left(C\right) = 0$.

PROPERTY 5A.6 (ISOMETRY). $E\left(\int_0^t C_s dW_s\right)^2 = \int_0^t E\left(C_s^2\right) ds$, *for all* $t \in [0,T]$.

Proof. Without loss of generality, set $t = t_k$. We have:

$$\begin{aligned} E\left(\int_0^t C_s dW_s\right)^2 &= E\left(\sum_{i=1}^k C_{t_{i-1}}\left(W_{t_i} - W_{t_{i-1}}\right)\right)^2 \\ &= E\left(\sum_{i=1}^k \sum_{j=1}^k C_{t_{i-1}}\left(W_{t_i} - W_{t_{i-1}}\right) C_{t_{j-1}}\left(W_{t_j} - W_{t_{j-1}}\right)\right) \\ &= E\left(\sum_{i=1}^k C_{t_{i-1}}^2\left(W_{t_i} - W_{t_{i-1}}\right)^2\right), \end{aligned}$$

where the last equality follows because $\left(W_{t_i} - W_{t_{i-1}}\right)$ and $\left(W_{t_j} - W_{t_{j-1}}\right)$ are independent, for all $i \neq j$. Then,

$$\begin{aligned} E\left(\int_0^t C_s dW_s\right)^2 &= E\left(\sum_{i=1}^k C_{t_{i-1}}^2\left(W_{t_i} - W_{t_{i-1}}\right)^2\right) \\ &= E\left(\sum_{i=1}^k E\left(C_{t_{i-1}}^2\left(W_{t_i} - W_{t_{i-1}}\right)^2 \middle| \mathbb{F}_{t_{i-1}}\right)\right) \\ &= E\left(\sum_{i=1}^k E\left(C_{t_{i-1}}^2 \middle| \mathbb{F}_{t_{i-1}}\right)(t_i - t_{i-1})\right) \\ &= \sum_{i=1}^k E\left(C_{t_{i-1}}^2\right)(t_i - t_{i-1}) \\ &= \int_0^t E\left(C_s^2\right) ds. \end{aligned}$$

Property 5A.7 (Linearity and linearity on adjacent intervals).

Property 5A.8. *$I_t(C)$ has continuous ω-paths.*

5.12.1.7 The general Itô's stochastic integral

We now consider a more general class of integrand $\mathcal{F}$-adapted processes C_t, $t \in [0,T]$ satisfying $\int_0^T E\left(C_s^2\right) ds < \infty$, and $\in L^2\left(P \otimes dt\right)$, which is obviously satisfied by simple processes, although now we are now moving to continuous time. Clearly, $\mathcal{H}^2$ is a closed linear subspace of $L^2\left(P \otimes dt\right)$. So let $\|\cdot\|_{L^2(P\otimes dt)}$ be the norm of $L^2\left(P \otimes dt\right)$. Let $\mathcal{H}_0^2$ be the subset of $\mathcal{H}^2$ consisting of all simple processes. We now outline how to construct the stochastic integral, in four steps.

Step 1: ($\mathcal{H}_0^2$ is dense in $\mathcal{H}^2$). It is possible to show that for any $C \in \mathcal{H}^2$, there exists a sequence of simple processes $C^{(n)}$ s.t $\left\|C - C^{(n)}\right\|_{L^2(P\otimes dt)} \to 0$, i.e. $\int_0^T E(C_s - C_s^{(n)})^2 ds \to 0$.

Step 2: By step 1, $\left\{C^{(n)}\right\}$ is a Cauchy sequence in $L^2\left(P \otimes dt\right)$. By the isometry property of the Itô's integral for simple processes

$$\left\|I_T(C^{(n)}) - I_T(C^{(n')})\right\|_{L^2(P)} = \left\|C^{(n)} - C^{(n')}\right\|_{L^2(P\otimes dt)}.$$

Therefore, $I_T\left(C^{(n)}\right)$ is a Cauchy sequence in $L^2\left(P\right)$. Now it is well-known that $L^2\left(P\right)$ is complete, and so $I_T\left(C^{(n)}\right)$ must converge to some element of $L^2\left(P\right)$, denoted as $I_T(C)$.

Step 3: This limit is called the *Itô's stochastic integral* of C, and is written as

$$I_T(C) = \int_0^T C_s dW_s.$$

Finally, the limit is well-defined: if there is another $C_*^{(n)}$: $\left\|C - C_*^{(n)}\right\|_{L^2(P\otimes dt)} \to 0$, then $\lim I_T(C_*^{(n)}) = \lim I_T(C^{(n)}) = I_T(C)$ in the $L^2(P)$ norm.

Step 4: (Itô's integral as a process) We wish to create a whole "continuum" of Itô's integrals at a single glance. Step 3 is not enough because we need uniform convergence on $[0,T]$. To show that it's feasible lies beyond the aim of these introductory lectures. The final result is that for any $C \in \mathcal{H}^2$, there exists a process $(I_t)_{t\in[0,T]}$, a continuous $\mathbb{F}_t$-martingale satisfying

$$I_t = \int_0^t C_s dW_s, \quad t \in [0,T], \quad P \otimes dt\text{-a.s.}$$

To summarize, introduce the two spaces

$$\mathcal{L}^n = \left\{\theta \in \mathcal{L} : \int_0^T |\theta_t|^n \, dt < \infty \text{ a.s.}\right\}, \; n = 1, 2, \qquad \mathcal{H}^2 = \left\{\theta \in \mathcal{L}^2 : E\left(\int_0^T |\theta_t|^2 \, dt\right) < \infty \text{ a.s.}\right\}$$

where $\mathcal{L}$ denotes the set of all adapted processes. Let $\theta \in \mathcal{H}^2$. The stochastic integral $I_t(\theta) = \int_0^t \theta_s dW_s$ satisfies the following properties: (i) Continuous sample paths, and $I_t(\theta)$ is a $\mathbb{F}_t$-martingale; (ii) Expectation equal to zero; (iii) Itô's isometry on $\mathcal{H}^2$, i.e. $E(\int_0^t \theta_s dW_s)^2 = \int_0^t E\left(\theta_s^2\right) ds < \infty$, $t \in [0,T]$, hence $E[\left|\int_0^t C_s dW_s\right|]^2 \leq E[\int_0^t C_s dW_s]^2 = \int_0^t E(C_s^2) ds < \infty$; (iv) Linearity and linearity on adjacent intervals.

A few remarks are in order. If $\theta \in \mathcal{H}^2$, then X solution to $dX_t = \theta_t dW_t$ is a martingale. If $\theta \notin \mathcal{H}^2$, but $\in \mathcal{L}^2$, X is, instead, called a *local martingale*. The converse is the *Martingale Representation Theorem*. This theorem states that if X is a $\mathbb{F}_t$-martingale, then there exists a $\theta \in \mathcal{H}^2 : dX_t = \theta_t dW_t$. This result is utilized in the main text of this chapter, when it helps us tell whether we live in a world with complete or incomplete markets. Moreover, in continuous-time finance, θ is often a portfolio strategy. It must be in $\mathcal{H}^2$ to avoid doubling strategies, which are a kind of arbitrage opportunities (at least in absence of frictions such as short-selling constraints). Assume, for example, that an asset price is W, and that this asset does not distribute dividends from 0 to T. Then dW is the instantaneous gain from holding one unit of this asset. The condition $\theta \in \mathcal{H}^2$ implies that these strategies cannot become arbitrarily large according to the $\mathcal{H}^2$ criterion. Moreover, the previous properties of $I_t(\theta) = \int_0^t \theta_s dW_s$ suggest that the "cumulative" gain process $G_t = G_0 + I_t(\theta)$ is a martingale (not only a "local" martingale). Therefore, no investor expects to make profits from investing in this asset.

5.12.1.8 Itô's lemma: Introduction

We develop, heuristically, a basic version of Itô's lemma, with its most general version stated further in this appendix. Let $f : \mathbb{R} \mapsto \mathbb{R}$ be twice continuously differentiable. We have:

$$f(W_t) = f(W_0) + \int_0^t f'(W_s) \, dW_s + \frac{1}{2}\int_0^t f''(W_s) \, ds, \tag{5A.4}$$

where the first integral is an Itô's stochastic integral, and second one is a Riemann's one. For example, let $f(B) = B^2$. Then,

$$W_t^2 = 2\int_0^t W_s dW_s + \int_0^t ds \Leftrightarrow \int_0^t W_s dW_s = \frac{1}{2}\left(W_t^2 - t\right). \tag{5A.5}$$

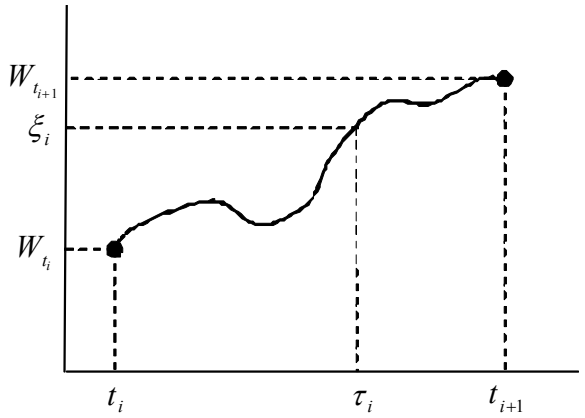
To provide a sketchy proof of Eq. (5A.4), note that:

$$f(W_t) - f(W_0) = \sum_{i=0}^{k-1} \left[f\left(W_{t_{i+1}}\right) - f\left(W_{t_i}\right) \right].$$

By Taylor,

$$f\left(W_{t_{i+1}}\right) - f\left(W_{t_i}\right) = f'\left(W_{t_i}\right)\left(W_{t_{i+1}} - W_{t_i}\right) + \frac{1}{2} f''\left(\xi_i\right)\left(W_{t_{i+1}} - W_{t_i}\right)^2,$$

where $\min\left(W_{t_i}, W_{t_{i+1}}\right) < \xi_i < \max\left(W_{t_i}, W_{t_{i+1}}\right)$, as in the figure below. Because W is continuous, $\xi_i(\omega) = W_{\tau_i(\omega)}$ for some $\tau_i(\omega) : t_i \leq \tau_i(\omega) \leq t_{i+1}$.



Therefore,

$$f\left(W_t\right) - f\left(W_0\right) = \sum_{i=0}^{k-1} f'\left(W_{t_i}\right)\left(W_{t_{i+1}} - W_{t_i}\right) + \frac{1}{2}\sum_{i=0}^{k-1} f''\left(W_{\tau_i}\right)\left(W_{t_{i+1}} - W_{t_i}\right)^2.$$

We have

$$\sum_{i=0}^{k-1} f''\left(W_{\tau_i}\right)\left(W_{t_{i+1}} - W_{t_i}\right)^2 \approx \sum_{i=0}^{k-1} f''\left(W_{\tau_i}\right)\left(t_{i+1} - t_i\right).$$

Finally,

$$\begin{aligned} \textstyle\sum_i f'\left(W_{t_i}\right)\left(W_{t_{i+1}} - W_{t_i}\right) &\rightarrow \int f'\left(W_s\right) dW_s \\ \textstyle\sum_i f''\left(W_{\tau_i}\right)\left(t_{i+1} - t_i\right) &\rightarrow \int f''\left(W_s\right) ds \end{aligned}$$

More technical details in order of descending difficulty can be found in Karatzas and Shreve (1991), Arnold (1974), Steele (2001) and Mikosch (1998).

Let us reconsider the example in Eq. (5A.4). By the stochastic integral theorem, is a martingale. This is confirmed by Eq. (5A.4). According to Eq. (5A.4),

$$\int_0^t W_s dW_s = \frac{1}{2}\left(W_t^2 - t\right)$$

and $\left(W_t^2 - t\right)$ is indeed a martingale for $E\left(W_t^2\right) = t$ all t.

5.12.2 Stochastic differential equations

5.12.2.1 Background

Consider the differential equation:

$$dx_t = \mu(t, x_t)\, dt, \quad x_0 = x,$$

for some function μ. Randomness can be introduced via an additional "noise term":

$$dx_t = \mu(t, x_t)\, dt + \sigma(t, x_t)\, dW_t, \quad x_0 = x.$$

We already know that a ω-W_t is not differentiable, so this is only a short-hand notation for:

$$x_t = x_0 + \int_0^t \mu(s, x_s)\, ds + \int_0^t \sigma(s, x_s)\, dW_s, \tag{5A.6}$$

where the first integral is Riemann and the second integral is an Itô's stochastic integral.

We have the following definitions. First, we say that an *Itô's process* is,

$$dx_t(\omega) = \mu_t(\omega)\, dt + \sigma_t(\omega)\, dW_t, \quad x_0 = x.$$

Moreover, we say that an *Itô's diffusion process* is,

$$dx_t = \mu(t, x_t)\, dt + \sigma(t, x_t)\, dW_t, \quad x_0 = x.$$

It is known that an Itô's diffusion process is a Markov process. The previous equation is also called a *stochastic differential equation* (SDE). In a SDE, μ and σ "depend" on ω only through x. Finally, we say that a *time-homogeneous diffusion process* is,

$$dx_t = \mu(x_t)\, dt + \sigma(x_t)\, dW_t, \quad x_0 = x.$$

There is a beautiful property that is used to price financial derivatives, using replication arguments, as explained in the main text, called the *unique decomposition property*. Suppose we were given two processes x and y with $x_0 = y_0$, and that:

$$dx_t = \mu_t^x dt + \sigma_t^x dW_t \quad \text{and} \quad dy_t = \mu_t^y dt + \sigma_t^y dW_t.$$

Then $x_t = y_t$ almost surely if and only if $\mu_t^x = \mu_t^y$ and $\sigma_t^x = \sigma_t^y$ almost everywhere, in the sense that $E(\int_0^T |a_t^x - a_t^y|\, dt = 0) = E(\int_0^T |b_t^x - b_t^y|\, dt) = 0$.

5.12.2.2 Basic definitions, properties and regularity conditions

How do we know whether the various integrals given before are well-defined. As an example, the Itô's integral representation $\int_0^t \sigma(s, x_s) dW_s$ works if σ is $\mathbb{F}_t$-adapted and $\int_0^t E\left[\sigma(s, x_s)^2\right] ds < \infty$. But how can be sure that these two basic conditions are satisfied if we don't know yet the *solution* of x? And, above all, what is a solution to a SDE? We have two concepts of such a solution, *strong* and *weak*.

Definition 5A.9. (Strong solution to a SDE) *A* strong *solution to Eq. (5A.6) is a stochastic process* $x = (x_t,\ t \in [0, T])$ *such that:*

(i) *x is $\mathbb{F}_t$-adapted.*

(ii) *The integrals in Eq. (5A.6) are well-defined in the Riemann's and Itô's sense and Eq. (5A.6) holds $P \otimes dt$-almost surely*

(iii) $E\left(\int_0^T |x_s|^2\, ds\right) < \infty$.

In other words, the definition of a *strong* solution requires that a Brownian motion be "given in advance," and that the solution x_t constructed from it be then $\mathbb{F}_t$-adapted.

Next, suppose, instead, that we were only given x_0 and two functions $\sigma(t,x)$ and $\mu(t,x)$, and that we were asked to find a pair of processes $(\tilde{x}, \tilde{W})$ on some probability space $(\Omega, \tilde{\mathcal{F}}, P)$ such that Eq. (5A.6) holds with $\tilde{x}$ being $\tilde{\mathbb{F}}_t$-adapted on some space, not necessarily the one in Eq. (5A.6). (Clearly such a $\tilde{x}$ needs not to be $\mathbb{F}_t$-adapted.) In this case $(\tilde{x}, \tilde{W})$ is called a *weak solution* on $(\Omega, \tilde{\mathcal{F}}, P)$. In the case of a weak solution, we are given x, μ, σ and then "we have to find" two things: a Brownian motion $\tilde{W}$ and a $\tilde{\mathbb{F}}_t$-adapted process $\tilde{x}$ such that $\tilde{x}_t = x_0 + \int_0^t \mu(s, \tilde{x}_s)\, ds + \int_0^t \sigma(s, \tilde{x}_s)\, d\tilde{W}_s$ holds $P \otimes dt$-almost surely. Clearly, a strong solution is also weak, but the converse is not true. Consider the following example.

Example 5A.10. (Tanaka equation) Let x_t satisfy:

$$dx_t = \text{sign}(x_t) dW_t,\ x_0 = 0. \tag{5A.7}$$

This equation has no strong solutions, for define

$$y_t = \int_0^t \text{sign}(\hat{W}_s) d\hat{W}_s, \tag{5A.8}$$

where $\hat{W}$ is a Brownian motion. It can be shown that y_t is $\mathbb{G}_t$-measurable, where $\mathbb{G}_t$ is the σ-algebra generated by $\mid \hat{W}_t \mid$. Clearly $\mathbb{G}_t \subset \hat{\mathbb{F}}_t$, where $\hat{\mathbb{F}}_t$ is the σ-algebra generated by $\hat{W}_t$. Therefore, the σ-algebra generated by y_t is also strictly contained in $\hat{\mathbb{F}}_t$. Armed with this result, one can show that there are no strong solutions to Eq. (5A.7). For assume the contrary. There is a theorem saying that x_t would then be a Brownian motion. On the other hand, Eq. (5A.7) can also be written

$$dW_t = \text{sign}(x_t) dx_t,\ x_0 = 0,$$

or

$$W_t = \int_0^t \text{sign}(x_s) dx_s.$$

By the same reasoning produced to show that the σ-algebra generated by y_t is strictly contained in $\hat{\mathbb{F}}_t$ in Eq. (5A.8), we conclude that the σ-algebra generated by W_t is strictly contained in the σ-algebra generated by x_t. But this contradicts that x_t is a strong solution to Eq. (5A.7).

Clearly, one needs to be able to impose conditions that allow to distinguish between weak and strong solutions. However, the only focus of the following discussion is about the regularity conditions ensuring existence and uniqueness of strong solutions—the case of interest in continuous-time finance. We need two types of restrictions on μ and σ. Consider the following definition. For a given function f, we say that it satisfies a Lipschitz condition in x if there exists a constant L, such that for all $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\|f(x,t) - f(y,t)\| \leq L \|x - y\| \text{ uniformly in } t.$$

where $\|A\| \equiv \sqrt{\text{Tr}(AA^\top)}$. In other words, f cannot change too widely. We also say f satisfies a growth condition in x, if there exists a constant G such that for all $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\|f(x,t)\|^2 \leq G\left(1 + \|x\|^2\right) \text{ uniformly in } t.$$

That is, f cannot grow too much.

Next, we turn to the concepts of existence and uniqueness of a solution to a stochastic differential equation. We say that if $x_t^{(1)}(\omega)$ and $x_t^{(2)}(\omega)$ are both strong solutions to Eq. (5A.6), then $x_t^{(1)}(\omega) = x_t^{(2)}(\omega)$ $P \otimes dt$-a.s. We have:

THEOREM 5A.11. *Suppose that μ, σ satisfy Lipschitz and growth conditions in x, then there exists a unique Itô's process x satisfying Eq. (5A.6) which is continuous adapted Markov.*

Consider the following stochastic differential equation:

$$dx_t = \mu(a - x_t)\,dt + \sigma\sqrt{x_t}dW_t,$$

for some constants μ, a, σ. This is the so-called square-root process utilized to model equity volatility (see Chapter 11), the short-term rate (see Chapter 13) or instantaneous probabilities of default of debt issuers (see Chapter 14). The point here, for now, is that the diffusion component does not satisfy the conditions in Theorem 5A.7. Yet it is possible to show that under suitable parameter restrictions there exists a strong solution. Incidentally, the solution to this simple equation is still unknown.

What about *uniqueness* of the solution? It is well-known that if μ, σ are locally Lipschitz continuous in x, then strong uniqueness holds. But even for ordinary differential equations, a local Lipschitz condition is not necessarily enough to guarantee global existence (i.e. for all t) of a solution. For example, consider the following equation:

$$\frac{dx_t}{dt} = \mu(x_t) \equiv x_t^2,\ x_0 = 1,$$

has as unique solution:

$$x_t = \frac{1}{1-t},\ 0 \leq t < 1.$$

Yet is impossible to find a global solution, i.e. one defined for all t. This is exactly the kind of pathology ruled out by linear-growth conditions. More generally, linear-growth conditions ensure that $|x_t(\omega)|$ is unique and doesn't explode in finite time. Naturally, Lipschitz and growth conditions are only *sufficient* conditions to guarantee the previous conclusions.

A final remark. The uniqueness concept used here refers to *strong* or *pathwise* uniqueness. There are also definitions of *weak* uniqueness to mean that any two solutions (weak or strong) have the same finite-dimensional distributions. For example, the Tanaka's equation introduced earlier has no strong solution, yet it can be shown that it has a (weakly) unique weak solution.

5.12.2.3 Itô's lemma

Itô's lemma is a fundamental tool of analysis in continuous-time finance. It helps build up "new" processes from "old" processes. Two examples might clarify.

(i) A share price is certainly a function of its dividend process. If the dividend process is solution to some SDE, then the asset price is a solution to another SDE. Which SDE? Itô's lemma will give us the answer.

(ii) As we know, derivatives are asset with a value depending on some underlying factors, whence, the terminology, "derivative." In other words, derivative prices are functions of these factors. If factors are solutions to SDE, derivative prices are also solutions to SDE. Once again, Itô's lemma will provide us with the right SDE.

Naturally, the functional form linking the dividend process (or the factors) to the asset prices is unknown. But in situations of interest, no-arbitrage restrictions will help to pin down such a functional form.

Let us proceed with a few preliminary heuristic considerations. A useful heuristic definition is that the increments of a Brownian motion, dW_t, can be thought of as being equal to $W_{T+\Delta t}-W_t$ as $\Delta t \to 0$. We may think of the "increments" dW_t as being normally distributed, $dW_t \sim N(0, dt)$. Heuristically, indeed, $\Delta W_t \equiv W_{t+\Delta t} - W_t \sim N(0, \Delta t)$. But then, by the previous normality property of ΔW_t,

$$E(\Delta W_t) = 0 \text{ and } E\left(\Delta W_t^2\right) = \Delta t, \text{ hence } var(\Delta W_t) = \Delta t, \quad \text{and} \quad var\left(\Delta W_t^2\right) = 2\Delta t^2,$$

where the second equality follows by the property χ^2 distributions.

The point of the previous computations is that for small Δt, the variance of ΔW_t^2, which is proportional to Δt^2, is negligible if compared to its expectation, which is Δt. Heuristically, $Q_n(t) \overset{q.m.}{\to} t$ and $dW_t^2 \equiv Q_n(dt) \overset{q.m.}{\to} dt$. These heuristic considerations lead to the following, celebrated table below.

Itô's multiplication table

$(dt)^n$	$=$	0	for $n > 1$
$dt \cdot dW$	$=$	0	
$(dW)^2$	$=$	dt	
$(dW)^n$	$=$	0	for $n > 2$
$dW_1 dW_2$	$=$	0	for two independent Brownian motions

We now use this table, and heuristically derive Itô's lemma. Let x_t be the solution to,

$$dx_t = \mu_t dt + \sigma_t dW_t,$$

and suppose we are given a function $f(x,t)$, which we assume to be as differentiable in (x,t) as many times as needed below. We expand f as follows:

$$df(x,t) = f_t(x,t)\,dt + f_x(x,t)\,dx + \frac{1}{2} f_{xx}(x,t)\,(dx)^2 + \text{Remainder},$$

where the remainder contains only terms of order higher than $(dx)^2$ and $(dt)^2$. So for reasons which will be clear in one moment we will discard it. We have,

$$\begin{aligned} df &= f_t dt + f_x dx + \frac{1}{2} f_{xx}(dx)^2 \\ &= f_t dt + f_x(\mu dt + \sigma dW) + \frac{1}{2} f_{xx}(\mu dt + \sigma dW_t)^2 \\ &= f_t dt + f_x \mu dt + f_x \sigma dW + \frac{1}{2} f_{xx}\left(\mu^2 (dt)^2 + \sigma^2 (dW)^2 + 2\mu\sigma(dt \cdot dW)\right). \end{aligned}$$

By the Itô's multiplication table,

$$df = f_t dt + f_x \mu dt + f_x \sigma dW + \frac{1}{2} f_{xx}(\mu^2 \underbrace{dt^2}_{\equiv 0} + \sigma^2 \underbrace{dW^2}_{\equiv dt} + 2\mu\sigma \underbrace{dt \cdot dW}_{\equiv 0}).$$

By rearranging terms,

$$df(x,t) = \left(f_t(x,t) + f_x(x,t)\,\mu + \frac{1}{2} f_{xx}(x,t)\,\sigma^2\right) dt + f_x(x,t)\,\sigma dW.$$

This is *Itô's lemma.*

To derive Itô's lemma in the case x is multidimensional, apply Itô's multiplication table to the following expansion:

$$df(x,t) = f_t dt + f_x dx + \frac{1}{2} \sum_{i,j} f_{x_i x_j} dx_i dx_j.$$

We have:

THEOREM 5A.12. (ITÔ'S LEMMA, MULTIDIMENSIONAL) *Assume that the the following process $x \in R^n$ is solution to*

$$dx_t = \mu(x_t, t)\, dt + \sigma(x_t, t)\, dW_t, \tag{5A.9}$$

where $\mu \in R^n$, σ is a $n \times d$ matrix and W is a d-dimensional standard Brownian motion. Consider the function $f(x,t)$, and assume it is twice continuously differentiable in x and differentiable in t. Then f is an Itô's process, solution to

$$df(x_t, t) = \mathcal{L}f(x_s, s)\, ds + f_x(x_s, s)\, \sigma(x_s, s)\, dW_s, \tag{5A.10}$$

where

$$\mathcal{L}f(x,t) = f_t(x,t) + f_x(x,t)\, \mu + \frac{1}{2}\mathrm{Tr}\left(\sigma\sigma^\top f_{xx}(x,t)\right)$$

and $f_x(x,t)$ and $f_{xx}(x,t)$ are the gradient and Hessian of f with respect to x.

Note that by Eq. (5A.10), and provided $f_x\sigma \in \mathcal{H}^2$, f is a martingale whenever $\mathcal{L}f(x,t) = 0$, for all x, t. As a matter of terminology, the operator $\mathcal{A}f(x,t) = f_x(x,t)\,\mu + \frac{1}{2}\mathrm{Tr}\left[\sigma\sigma^\top f_{xx}(x,t)\right]$ is usually referred to as the *infinitesimal generator* of the diffusion process in Eq. (5A.9).

5.13 Appendix 2: Self-financed strategies, from discrete to continuous time

5.13.1 With intermediate consumption

We have

$$c_t + S_t\theta_{1,t+1} + b_t\theta_{2,t+1} = (S_t + D_t)\,\theta_{1,t} + b_t\theta_{2,t} \equiv V_t + D_t\theta_{1,t},$$

where $V_t \equiv S_t\theta_{1,t} + b_t\theta_{2,t}$ is wealth net of dividends, and $b_t\theta_{2,t}$ is the overall value of the money market account.

We have,

$$\begin{aligned} V_t - V_{t-1} &= S_t\theta_{1,t} + b_t\theta_{2,t} - V_{t-1} \\ &= S_t\theta_{1,t} + b_t\theta_{2,t} - (c_{t-1} + S_{t-1}\theta_{1,t} + b_{t-1}\theta_{2,t} - D_{t-1}\theta_{1,t-1}) \\ &= (S_t - S_{t-1})\,\theta_{1,t} + (b_t - b_{t-1})\,\theta_{2,t} - c_{t-1} + D_{t-1}\theta_{1,t-1}, \end{aligned}$$

and more generally,

$$V_t - V_{t-\Delta} = (S_t - S_{t-\Delta})\,\theta_{1,t} + (b_t - b_{t-\Delta})\,\theta_{2,t} - (c_{t-\Delta}\cdot\Delta) + (D_{t-\Delta}\cdot\Delta)\,\theta_{1,t-\Delta}.$$

Now let $\Delta \downarrow 0$ and assume that θ_1 and θ_2 are constant between t and $t-\Delta$. We have:

$$dV_\tau = (dS_\tau + D_\tau d\tau)\,\theta_{1\tau} + db_\tau\theta_{2\tau} - c_\tau d\tau.$$

Assume that

$$\frac{db_\tau}{b_\tau} = r d\tau.$$

The budget constraint can then be written as:

$$\begin{aligned} dV_\tau &= (dS_\tau + D_\tau d\tau)\,\theta_{1\tau} + rb_\tau\theta_{2\tau}d\tau - c_\tau d\tau \\ &= (dS_\tau + D_\tau d\tau)\,\theta_{1\tau} + r\,(V - S_\tau\theta_{1\tau})\,d\tau - c_\tau d\tau \\ &= (dS_\tau + D_\tau d\tau - rS_\tau d\tau)\,\theta_{1\tau} + rVd\tau - c_\tau d\tau \\ &= \left(\frac{dS_\tau}{S_\tau} + \frac{D_\tau}{S_\tau}d\tau - rd\tau\right)\theta_{1\tau}S_\tau + rVd\tau - c_\tau d\tau \\ &= \left(\frac{dS_\tau}{S_\tau} + \frac{D_\tau}{S_\tau}d\tau - rd\tau\right)\pi_\tau + rVd\tau - c_\tau d\tau. \end{aligned}$$

5.13.2 With final consumption only

We might be interested in models where consumption takes place only at the end of the period. Let $\bar{S} = \left(S^{(0)}, S\right)^\intercal$ and $\bar{\theta} = (\theta^{(0)}, \theta)$, where θ and S are both m-dimensional. Define as usual wealth as of time t as $V_t \equiv \bar{S}_t\bar{\theta}_t$. There are no dividends. A *self-financing* strategy $\bar{\theta}$ satisfies,

$$\bar{S}_t\bar{\theta}_{t+1} = \bar{S}_t\bar{\theta}_t \equiv V_t, \quad t = 1, \cdots, T.$$

Therefore,

$$\begin{aligned} V_t &= \bar{S}_t\bar{\theta}_t + \bar{S}_{t-1}\bar{\theta}_{t-1} - \bar{S}_{t-1}\bar{\theta}_{t-1} \\ &= \bar{S}_t\bar{\theta}_t + \bar{S}_{t-1}\bar{\theta}_{t-1} - \bar{S}_{t-1}\bar{\theta}_t \qquad \text{(because } \bar{\theta} \text{ is self-financing)} \\ &= V_{t-1} + \Delta\bar{S}_t\bar{\theta}_t, \quad \Delta\bar{S}_t \equiv \bar{S}_t - \bar{S}_{t-1}, \quad t = 1, \cdots, T, \end{aligned}$$

or,

$$V_t = V_1 + \sum_{n=1}^{t} \Delta\bar{S}_n\bar{\theta}_n.$$

Next, suppose that

$$\Delta S_t^{(0)} = r_t S_{t-1}^{(0)}, \quad t = 1, \cdots, T,$$

with $\{r_t\}_{t=1}^T$ given and to be defined more precisely below. The term $\Delta \bar{S}_t \theta_t^+$ can then be rewritten as:

$$\begin{aligned}
\Delta \bar{S}_t \bar{\theta}_t &= \Delta S_t^{(0)} \theta_t^{(0)} + \Delta S_t \theta_t \\
&= r_t S_{t-1}^{(0)} \theta_t^{(0)} + \Delta S_t \theta_t \\
&= r_t S_{t-1}^{(0)} \theta_t^{(0)} + r_t S_{t-1} \theta_t - r_t S_{t-1} \theta_t + \Delta S_t \theta_t \\
&= r_t \bar{S}_{t-1} \bar{\theta}_t - r_t S_{t-1} \theta_t + \Delta q_t \theta_t \\
&= r_t \bar{S}_{t-1} \bar{\theta}_{t-1} - r_t S_{t-1} \theta_t + \Delta S_t \theta_t \quad \text{(because } \bar{\theta} \text{ is self-financing)} \\
&= r_t V_{t-1} - r_t S_{t-1} \theta_t + \Delta S_t \theta_t,
\end{aligned}$$

and we obtain

$$V_t = (1 + r_t) V_{t-1} - r_t S_{t-1} \theta_t + \Delta S_t \theta_t,$$

or,

$$V_t = V_1 + \sum_{n=1}^{t} \left(r_n V_{n-1} - r_n S_{n-1} \theta_n + \Delta S_n \theta_n \right).$$

Next, considering "small" time intervals. In the limit we obtain:

$$dV_t = r_t V_t dt - r_t S_t \theta_t dt + dS_t \theta_t.$$

Such an equation can also be arrived at by noticing that current wealth is nothing but initial wealth plus gains from trade accumulated up to now:

$$V_t = V_0 + \int_0^t d\bar{S}_u \bar{\theta}_u.$$

$\Leftrightarrow$

$$\begin{aligned}
dV_t &= d\bar{S}_t \bar{\theta}_t^+ \\
&= dS_{0t} \theta_{0t} + dS_t \theta_t \\
&= r_t S_{0t} \theta_{0t} dt + dS_t \theta_t \\
&= r_t \left(V_t - S_t \theta_t \right) dt + dS_t \theta_t \\
&= r_t V_t dt - r_t S_t \theta_t dt + dS_t \theta_t.
\end{aligned}$$

Now consider the sequence of problems of terminal wealth maximization:

$$\text{For } t = 1, \cdots, T, \quad \mathcal{P}_t : \left\{ \begin{array}{l} \max_{\theta_t} E\left[u(V_T) \middle| \mathbb{F}_{t-1} \right], \\ \text{s.t. } V_t = (1 + r_t) V_{t-1} - r_t S_{t-1} \theta_t + \Delta S_t \theta_t \end{array} \right.$$

Even if markets are incomplete, agents can solve the sequence of problems $\{\mathcal{P}_t\}_{t=1}^T$ as time unfolds. Each problem can be written as:

$$\max_{\theta_t} E\left[u\left(V_1 + \sum_{t=1}^{T} \left(r_t V_{t-1} - r_t S_{t-1} \theta_t + \Delta S_t \theta_t \right) \right) \middle| \mathbb{F}_{t-1} \right].$$

The FOC for any generic t leads to:

$$\hat{S}_t = \frac{E\left(u'(V_T) \cdot \hat{S}_{t+1} \middle| \mathbb{F}_t \right)}{E\left(u'(V_T) \middle| \mathbb{F}_t \right)}, \quad t = 0, \cdots, T-1,$$

where $\hat{S}_t = \frac{S_t}{M_t}$, i.e. the prices expressed in terms of the money market numéraire.

The previous relations suggest that we can define a *martingale measure* Q for the price process (expressed in terms of the money market numéraire), by defining

$$\left.\frac{dQ}{dP}\right|_{\mathbb{F}_t} = \frac{u'(V_T)}{E\left(u'(V_T)|\,\mathbb{F}_t\right)}.$$

5.14 Appendix 3: Proof of selected results

5.14.1 Proof of Theorem 5.3

We know that, by Girsanov's theorem, $\mathcal{Q}$ is non-empty if and only if Eq. (5.71) holds true (see the discussion in the main text) Therefore, the proofs of this theorem rely on Eq. (5.71).

If part. With $c \equiv 0$, Eq. (5.72) is

$$\frac{V_\tau}{S_{0\tau}} = V_t + \int_t^\tau \frac{\pi_u^\top \sigma_u}{S_{0u}} dW_{0u}, \quad \tau \in [t, T],$$

or $V_t = E_\tau^Q \left(S_{0T}^{-1} V_T^{x,\pi,0} \right)$. An arbitrage opportunity is $V_t \leq S_{0T}^{-1} V_T^{x,\pi,0}$ almost surely under Q, but the previous equality leaves $V_t = S_{0T}^{-1} V_T$ almost surely under Q (if a random variable $\tilde{y} \geq 0$ and $E_t(\tilde{y}) = 0$, then, $\tilde{y} = 0$ almost surely) and, then, almost surely under P. But this equality is in contradiction with $\Pr\left(S_{0T}^{-1} V_T - V_t > 0 \right) > 0$ under zero-consumption paths, as required by Definition 4.3.

Only if part. We combine portions of proofs in Karatzas (1997, thm. 0.2.4 pp. 6-7) and Øksendal (1998, thm. 12.1.8b, pp. 256-257). We let:

$$\begin{aligned} Z(\tau) &= \{\omega \in \Omega : \text{ Eq. (5.71) has no solutions}\} \\ &= \{\omega \in \Omega : a_\tau(\omega) - \mathbf{1}_m r_\tau(\omega) \notin \langle \sigma \rangle\} \\ &= \left\{ \omega \in \Omega : \exists \underline{\pi}_\tau(\omega) : \underline{\pi}_\tau(\omega)^\top \sigma_\tau(\omega) = 0 \text{ and } \underline{\pi}_\tau(\omega)^\top \left(a_\tau(\omega) - \mathbf{1}_m r_\tau(\omega) \right) \neq 0 \right\}, \end{aligned}$$

and consider the following portfolio,

$$\hat{\pi}(\tau; w) = \begin{cases} k \cdot sign\left[\underline{\pi}_\tau(\omega)^\top \left(a_\tau(\omega) - \mathbf{1}_m r_\tau(\omega) \right) \right] \cdot \underline{\pi}_\tau(\omega) & \text{for } \omega \in Z(\tau) \\ 0 & \text{for } \omega \notin Z(\tau) \end{cases}$$

Clearly $\hat{\pi}$ is $(\tau; \omega)$-measurable, and generates, by Eq. (5.68),

$$\begin{aligned} \frac{V_\tau}{S_{0\tau}} &= V_t + \int_t^\tau \frac{\hat{\pi}_u^\top \left(a_u - \mathbf{1}_m r_u \right)}{S_{0u}} \mathbb{I}_{Z_u} du + \int_t^\tau \frac{\hat{\pi}_u^\top \sigma_u}{S_{0u}} \mathbb{I}_{Z_u} dW_u \\ &= V_t + \int_t^\tau \left(\frac{\hat{\pi}_u^\top \left(a_u - \mathbf{1}_m r_u \right)}{S_{0u}} \right) \mathbb{I}_{Z_u} du \\ &\geq V_t. \end{aligned}$$

So the market has no arbitrage only if $\mathbb{I}_{Z_u} = 0$, i.e. only if Eq. (5.71) has at least one solution. ‖

5.14.2 Details regarding multibeta CAPM

We provide precise definitions utilized in Section 5.5.3. First, consider the covariance between asses and market returns

$$cov\left(\sigma dW, \tfrac{dV_M}{V_M} \right) = \sigma \sigma^\top \frac{\text{Cap}}{V_M} dt \equiv \text{cov}_{S,M} dt,$$

where V_M is solution to Eq. (5.99) with π as in Eq. (5.81) and equal to Cap in equilibrium. Furthermore, we have that

$$\begin{aligned} cov\left(\tfrac{dV_M}{V_M}, dy \right) &= \frac{\text{Cap}^\top}{V_M} \sigma \Sigma^\top dt \equiv \text{cov}_{M,y} dt \\ var\left(\tfrac{dV_M}{V_M} \right) &= \frac{\text{Cap}^\top}{V_M} \text{cov}_{S,M} dt \equiv \text{var}_M dt \\ cov\left(\tfrac{dV_{H_i}}{V_{H_i}}, \tfrac{dV_M}{V_M} \right) &= \frac{\pi_{H_i}^\top}{V_{H_i}} \sigma \sigma^\top \frac{\text{Cap}}{V_M} dt \equiv \text{cov}_{H_i,M} dt \end{aligned}$$

where V_H is now solution to Eq. (5.99) with π as in Eq. (5.89).

Based on these definitions, Eq. (5.88) follows because

$$\mu_M - r = \frac{\pi_M^\top}{V_M}\text{cov}_{S,M}\frac{-V\mathcal{J}_{VV}}{\mathcal{J}_V} + \frac{\pi_M^\top}{V_M}\sigma\Sigma^\top\frac{-\mathcal{J}_{Vy}}{\mathcal{J}_V} = \text{var}_M\frac{-V\mathcal{J}_{VV}}{\mathcal{J}_V} + \text{cov}_{M,y}\frac{-\mathcal{J}_{Vy}}{\mathcal{J}_V},$$

and Eq. (5.90) follows by similar arguments.

The solution to Eqs. (5.88) and (5.90) is

$$\begin{pmatrix} \frac{-V\mathcal{J}_{VV}}{\mathcal{J}_V} \\ \frac{-\mathcal{J}_{Vy}}{\mathcal{J}_V} \end{pmatrix} = \begin{pmatrix} \text{var}_M & \text{cov}_{M,y} \\ \text{cov}_{H,M} & \text{cov}_{H,y} \end{pmatrix}^{-1} \begin{pmatrix} \mu_M - r \\ \mu_H - \mathbf{1}_d r \end{pmatrix},$$

where μ_H and $\text{cov}_{H,M}$ are two d-dimensional (column) vectors with generic components μ_{H_i} and $\text{cov}_{H_i,M}$, respectively, and $\text{cov}_{H,y}$ is a $d \times d$ matrix with generic row $\text{cov}_{H_i,y}$. Replacing this result into Eq. (5.87) leaves, for each $i = 1, \cdots, m$,

$$\mu_i - r = \begin{pmatrix} \text{cov}_{S_i,M} & \text{cov}_{S_i,y} \end{pmatrix} \begin{pmatrix} \text{var}_M & \text{cov}_{M,y} \\ \text{cov}_{H,M} & \text{cov}_{H,y} \end{pmatrix}^{-1} \begin{pmatrix} \mu_M - r \\ \mu_H - \mathbf{1}_d r \end{pmatrix},$$

where $\text{cov}_{S_i,y} \equiv \left(\sigma\Sigma^\top\right)_i$ relates to the vector of the instantaneous covariance between the i-th asset returns and the state variables

$$cov\left(\frac{dS_i}{S_i}, dy\right) = \left(\sigma\Sigma^\top\right)_i dt,$$

and Eq. (5.91) follows by standard results on multiple regressions.

To illustrate, fix $d = 1$. Note that

$$\text{cov}_{S_i,y} = \kappa\text{cov}_{S_i,H}, \quad \text{cov}_{M,y} = \kappa\text{cov}_{H,M}, \quad \text{cov}_{H,y} = \kappa\text{var}_H,$$

where $\kappa \equiv \mathbf{1}_m^\top \left(\sigma\sigma^\top\right)^{-1}\sigma\Sigma_i^\top$, such that, for each $i = 1, \cdots, m$,

$$\mu_i - r = \begin{pmatrix} \text{cov}_{S_i,M} & \text{cov}_{S_i,H} \end{pmatrix} \begin{pmatrix} \text{var}_M & \text{cov}_{H,M} \\ \text{cov}_{H,M} & \text{var}_H \end{pmatrix}^{-1} \begin{pmatrix} \mu_M - r \\ \mu_H - r \end{pmatrix},$$

which can be regarded as the predictive part of a regression of the excess returns of security i onto the the excess returns of the market and the (single) hedging portfolio.

5.14.3 Proof of Eq. (5.109).

We have

$$\begin{aligned}
V_t &= \mathbb{E}_t\left(\frac{V_T}{S_{0T}} + \int_t^T \frac{c_u}{S_{0u}}du\right) \\
&= \eta_t^{-1}E_t\left(\frac{\eta_T V_T}{S_{0T}} + \int_t^T \frac{\eta_T c_u}{S_{0u}}du\right) \\
&= \eta_t^{-1}E_t\left(\frac{\eta_T V_T}{S_{0T}} + \int_t^T E_u\left(\frac{\eta_T c_u}{S_{0u}}\right)du\right) \\
&= \eta_t^{-1}E_t\left(\frac{\eta_T V_T}{S_{0T}} + \int_t^T \frac{E_u\left(\eta_T\right)c_u}{S_{0u}}du\right) \\
&= \eta_t^{-1}E_t\left(\frac{\eta_T V_T}{S_{0T}} + \int_t^T \frac{\eta_u c_u}{S_{0u}}du\right) \\
&= E_t\left(m_{t,T}V_T + \int_t^T m_{t,u}c_u du\right),
\end{aligned}$$

where we used the fact that c is adapted, the law of iterated expectations, the martingale property of η, and the definition of $m_{0,t}$.

5.14.4 Walras consistency tests

[This section is in progress]

First, we show that Eq. (**??**) ⇒ E*q*. (**??**). To develop intuition, consider the two-period economy of Chapter 2, where in the absence of arbitrage, there is a vector (interpreted as the vector of Arrow-Debreu securities) $\phi \in \mathbb{R}^d$ such that $\phi^\top(c^1 - w^1) = S\theta = -(c_0 - w_0)$, whence $c_s = w_s, s = 0, \cdots, d \Longleftrightarrow \theta = \mathbf{0}_m$. In the continuous time model, absence of arbitrage and market completeness imply that there exists a unique $Q \in \mathcal{Q}$ such that,

$$\frac{V_\tau}{S_{0\tau}} \equiv \frac{\theta_{0\tau} S_{0\tau} + \pi_\tau^\top \mathbf{1}_m}{S_{0\tau}} = \mathbf{1}_m^\top S_t + \int_t^\tau \frac{\pi^\top \sigma_u}{S_{0u}} dW_{0u} - \int_t^\tau \frac{c_u}{S_{0u}} du.$$

That is,

$$\begin{aligned} &\frac{\theta_{0\tau} S_{0\tau} + \left(\pi_\tau^\top - S_\tau^\top\right) \mathbf{1}_m}{S_{0\tau}} + \frac{S_\tau^\top \mathbf{1}_m}{S_{0\tau}} \\ &= \mathbf{1}_m^\top S_t + \int_t^\tau \frac{\left(\pi_u^\top - S_u^\top\right) \sigma_u}{S_{0u}} dW_{0u} - \int_t^\tau \frac{c_u}{S_{0u}} du + \int_t^\tau \frac{S_u^\top \sigma_u}{S_{0u}} dW_{0u}. \end{aligned}$$

Plugging the solution $\frac{S_{i\tau}}{S_{0\tau}} = S_{it} + \int_t^\tau \left(S_{0u}^{-1} S_{iu}\right) \sigma_{iu} dW_{0u} - \int_t^\tau \left(S_{0u}^{-1} D_{iu}\right) du$ in the previous relation,

$$\frac{\theta_{0T} S_{0T} + \left(\pi_T^\top - S_T^\top\right) \mathbf{1}_m}{S_{0T}} = \int_t^T \frac{\pi_u^\top - S_u^\top}{S_{0u}} \sigma_u dW_{0u} + \int_t^T \frac{D_u - c_u}{S_{0u}} du. \tag{5A.11}$$

When Eq. (**??**) holds, we have that $V_T^{x,\pi,c} = \theta_{0T} S_{0T} + \pi_T^\top \mathbf{1}_m = S_T^\top \mathbf{1}_m$, and $D = c$, and Eq. (5A.11) becomes:

$$0 = x_T \equiv \int_t^T \frac{\pi_u^\top - S_u^\top}{S_{0u}} \sigma_u dW_{0u},$$

a martingale starting at zero, satisfying:

$$dx_\tau = \frac{\pi_\tau^\top - S_\tau^\top}{S_{0\tau}} \sigma_\tau dW_{0\tau} = 0.$$

Since $\ker(\sigma) = \{\emptyset\}$ then, we have that $\pi_\tau = S_\tau$ a.s. for $\tau \in [t, T]$ and, hence, $\pi_\tau = S_\tau$ a.s. for $\tau \in [t, T]$. It is easily checked that this implies $\theta_{0T} = 0$ P-a.s. and that in fact, $\theta_{0\tau} = 0$ a.s.

Next, we show that Eq. (**??**) ⇒ Eq. (**??**). When Eq. (**??**) holds, Eq. (5A.11) becomes:

$$0 = y_T \equiv \int_t^T \frac{D_u - c_u}{S_{0u}} du,$$

a martingale starting at zero. We conclude by the same arguments used in the proof of the previous part. ‖

5.15 Appendix 4: The Green's function

5.15.1 Setup

In Section 5.5, it is shown that in frictionless markets, the value of a security as of time τ is:

$$V(x_\tau, \tau) = E\left(\frac{m_{t,T}}{m_{t,\tau}} V(x_T, T) + \int_\tau^T \frac{m_{t,s}}{m_{t,\tau}} h(x_s, s)\, ds\right), \tag{5A.12}$$

where $m_{t,\tau}$ is the stochastic discount factor,

$$m_{t,\tau} = \frac{\eta(t,\tau)}{S_{0\tau}} = \frac{1}{S_{0\tau}} \cdot \left.\frac{dQ}{dP}\right|_{\mathbb{F}_\tau}.$$

The Arrow-Debreu state price density is:

$$\phi_{t,T} = m_{t,T} dP = \frac{S_{0t}}{S_{0T}} dQ.$$

Our aim is to characterize this density in terms of partial differential equations. By the same reasoning produced in Section 5.5, Eq. (5A.12) can be rewritten as:

$$V(x_\tau, \tau) = \mathbb{E}\left(a(\tau, T) V(x_T, T) + \int_\tau^T a(\tau, s) h(x_s, s)\, ds\right), \quad a\left(t', t''\right) \equiv \frac{S_0(t')}{S_0(t'')}. \tag{5A.13}$$

Next, consider the state vector, $y_u \equiv (a(\tau, u),\ x_u)$, $\tau \leq u \leq T$, and let $q(y(t') | y_\tau)$ be the risk-neutral density of y. We have,

$$\begin{aligned} V(x_\tau, \tau) &= \mathbb{E}\left(a(\tau, T) V(x_T, T) + \int_\tau^T a(\tau, s) h(x_s, s)\, ds\right) \\ &= \int a(\tau, T) V(x_T, T) q(y_T | y_\tau)\, dy_T + \int_\tau^T \int a(\tau, s) h(x_s, s) q(y_s | y_\tau)\, dy_s ds. \end{aligned}$$

If $V(x_T, T)$ and $a(\tau, T)$ are independent,

$$\int a(\tau, T) V(x_T, T) q(y_T | y_\tau)\, dy_T = \int_X G(\tau, T) V(x_T, T)\, dx_T$$

where:

$$G(\tau, T) \equiv \int_A a(\tau, T) q(y_T | y_\tau)\, dy_T.$$

Assuming the same for h,

$$V(x_\tau, \tau) = \int_X G(\tau, T) V(x_T, T)\, dx_T + \int_\tau^T \int_X G(\tau, s) h(x_s, s)\, dx_s ds.$$

The function G is known as the *Green's function*:

$$G(t, \ell) \equiv G(x, t; \xi, \ell) = \int_A a(t, \ell) q(y(\ell) | y_t)\, da.$$

It is the value in state $x \in \mathbb{R}^d$ as of time t of a unit of numéraire at $\ell > t$, provided future states are in a neighborhood (in $\mathbb{R}^d$) of ξ. It is thus the Arrow-Debreu state-price density.

For example, a pure discount bond has $V(x, T) = 1\ \forall x$, and $h(x, s) = 1\ \forall x, s$, and

$$V(x_\tau, \tau) = \int_X G(x_\tau, \tau; \xi, T)\, d\xi, \quad \text{with } \lim_{\tau \uparrow T} G(x_\tau, \tau; \xi, T) = \delta(x_\tau - \xi),$$

where δ is the Dirac's delta.

5.15.2 The PDE connection

We show the Green's function satisfies the same partial differential equation (PDE) satisfied by the security price, but with a different boundary condition, and with the instantaneous dividend taken out. We have:

$$V\left(x_t,t\right)=\int_X G\left(x_t,t;\xi_T,T\right)V\left(\xi_T,T\right)d\xi_T+\int_t^T\int_X G(x_t,t;\xi_s,s)h\left(\xi_s,s\right)d\xi_s ds. \tag{5A.14}$$

Consider the scalar case. By Eq. (5A.13), and the Feynman-Kac connection between PDEs and conditional expectations reviewed in Section 5.2, we have that under regularity conditions, V is solution to:

$$0=V_t+\mu V_x+\frac{1}{2}\sigma^2 V_{xx}-rV+h, \tag{5A.15}$$

where μ is the risk-neutral drift of x. Next, take the following partial derivatives of $V\left(x,t\right)$ in Eq. (5A.14):

$$\begin{aligned}
V_t &= \int_X G_t V d\xi-\int_X \delta(x-\xi)h d\xi+\int_t^T\int_X G_t h d\xi ds=\int_X G_t V d\xi-h+\int_\tau^T\int_X G_t h d\xi ds\\
V_x &= \int_X G_x V d\xi+\int_t^T\int_X G_x h d\xi ds\\
V_{xx} &= \int_X G_{xx} V d\xi+\int_t^T\int_X G_{xx} h d\xi ds
\end{aligned}$$

and replace them into Eq. (5A.15) to obtain:

$$\begin{aligned}
0 &= \int_X\left(G_t+\mu G_x+\frac{1}{2}\sigma^2 G_{xx}-rG\right)V\left(\xi_T,T\right)d\xi_T\\
&\quad+\int_t^T\int_X\left(G_t+\mu G_x+\frac{1}{2}\sigma^2 G_{xx}-rG\right)h\left(\xi_s,s\right)d\xi_s ds.
\end{aligned}$$

This shows that G is solution to

$$0=G_t+\mu G_x+\frac{1}{2}\sigma^2 G_{xx}-rG,\quad \text{with } \lim_{t\uparrow T}G\left(x,t;\xi,T\right)=\delta\left(x-\xi\right).$$

5.16 Appendix 5: Portfolio constraints

We are looking for a portfolio-consumption policy $(p_{\hat{\nu}}, c_{\hat{\nu}})$ such that

$$\text{Val}\,(x;K) = E\left(\int_0^T u\,(t, c_{\hat{\nu}t})\,dt + U\left(V_T^{x,p_{\hat{\nu}},c_{\hat{\nu}}}\right)\right) \equiv \text{Val}_{\hat{\nu}}\,(x)\,, \tag{5A.16}$$

and such that $\hat{\nu}$ leads to $p_{\hat{\nu}t} \in K$ for all $t \in [0,T]$. Two small remarks on notation. We remind that we are defining as in Section 5.7, (i) $\text{Val}_\nu\,(x)$ as the value of the problem an investor faces in the *unconstrained* market in Eqs. (5.165), and (ii) p_t as the *normalized* portfolio process, $p_t \equiv \frac{\pi_t}{V_t}$.

It's really this. As Cvitanić and Karatzas (1992) put it, we simply search for a member of a family of unconstrained problems (those arising from the market in Eqs. (5.165)), one for which the optimal portfolio actually satisfies the constraint (i.e. without imposing it a priori), and thereby solves the original problem.

Note that because K contains the origin, the support function ζ in Eq. (5.163) satisfies $\zeta\,(\nu) \geq 0$ for each $\nu \in \tilde{K}$. Moreover, by construction,

$$p \in K \Longleftrightarrow \zeta\,(\nu) + p^\top \nu \geq 0, \quad \forall \nu \in \tilde{K}. \tag{5A.17}$$

Next, define the standard Brownian motion under the probability Q^ν, defined through the Radon-Nikodym in Eq. (5.167):

$$W_{\nu t} = W_t + \int_0^t \left(\lambda_u + \sigma_u^{-1}\nu_u\right) du \equiv W_{0t} + \int_0^t \left(\sigma_u^{-1}\nu_u\right) du,$$

where $\lambda = \sigma^{-1}\,(\hat{a} - \mathbf{1}_d r)$, and W_0 is the usual Brownian under the risk-neutral probability in a market without frictions. If the price system is as in the artificial market of Eqs. (5.165), then, for *any* unconstrained portfolio-consumption (p,c), the dynamics of wealth, $V_\nu^{x,p,c}$ say, are easily seen to be:

$$\begin{aligned} dV_\nu^{x,p,c} &= \left(\pi^\top\left(\hat{a}^\nu - \mathbf{1}_d r^\nu\right) + r^\nu V_\nu^{x,p,c} - c\right) dt + \pi^\top \sigma dW, \\ &= \left(\left(p^\top\nu + \zeta\,(\nu)\right) V_\nu^{x,p,c} + r V_\nu^{x,p,c} - c\right) dt + V_\nu^{x,p,c} p^\top \sigma dW_0, \end{aligned} \tag{5A.18}$$

where the second expression follows by a change in probability and the expressions for r^ν and $\hat{a}^\nu$ given in Eq. (5.166).

Therefore, for a given normalized portfolio-consumption (p,c), we have that the wealth difference, $\Delta\,(t) \equiv \frac{V_{\nu T}^{x,\pi,c} - V_T^{x,\pi,c}}{S_{0T}}$, satisfies:

$$d\Delta\,(t) = \underbrace{\frac{V_{\nu t}^{x,\pi,c}}{S_{0t}}\left(p_t^\top \nu_t + \zeta\,(\nu_t)\right)}_{\equiv m(t)} dt + \Delta_t p_t^\top \sigma_t dW_{0t}, \quad \Delta_0 = 0.$$

Next, consider, the simpler equation,

$$d\bar{\Delta}_t = \bar{\Delta}_t p_t^\top \sigma_t dW_{0t}, \quad \bar{\Delta}_0 = 0. \tag{5A.19}$$

Because $m_t \geq 0$ by Eq. (5A.17), then, by a comparison theorem (e.g., Karatzas and Shreve (1991, p. 291-295)), $\Delta_t \geq \bar{\Delta}_t = 0$, where the last equality follows because the solution to Eq. (5A.19) is $\bar{\Delta}_t = \bar{\Delta}_0 L_t$, for some positive process L_t. Therefore, we have,

$$V_{\nu t}^{x,p,c} \geq V_t^{x,p,c}, \quad \text{with an equality if} \quad \zeta\,(\nu_t) + p^\top \nu_t = 0 \text{ for all } t. \tag{5A.20}$$

Assume now that there exists a $\hat{\nu}$ and a portfolio-consumption pair $(p_{\hat{\nu}}, c_{\hat{\nu}})$, such that

$$p_{\hat{\nu}t} \in K \quad \text{and} \quad \zeta\,(\hat{\nu}_t) + p_{\hat{\nu}t}^\top \hat{\nu}_t = 0. \tag{5A.21}$$

(Meaning that it's a portfolio chosen without imposing the constraint, which then *happens* to satisfy the constraint.)

By the inequality in (5A.20), we have that $\text{Val}(x;K) \leq \text{Val}_\nu(x)$ for all ν and, hence,

$$\text{Val}(x;K) \leq \text{Val}_{\hat{\nu}}(x) \quad \text{and} \quad \text{Val}(x;K) \leq \inf_{\nu \in \tilde{K}} (\text{Val}_\nu(x)). \tag{5A.22}$$

Moreover, we have,

$$\begin{aligned}
\text{Val}(x;K) &= E\left(\int_0^T u(t,c_t)\,dt + U\left(V_T^{x,p,c}\right)\right), \quad p_t \in K \\
&\geq E\left(\int_0^T u(t,c_{\hat{\nu}t})\,dt + U\left(V_T^{x,p_{\hat{\nu}},c_{\hat{\nu}}}\right)\right) \\
&= E\left(\int_0^T u(t,c_{\hat{\nu}t})\,dt + U\left(V_{\hat{\nu}T}^{x,p_{\hat{\nu}},c_{\hat{\nu}}}\right)\right) \\
&= \text{Val}_{\hat{\nu}}(x),
\end{aligned} \tag{5A.23}$$

where the second line follows, because the value of the original problem is, of course, larger than that of any constrained and not-optimally chosen portfolio-consumption $(p_{\hat{\nu}}, c_{\hat{\nu}})$. In other words, note that as soon as the second equality in Eq. (5A.21) holds true, then, the dynamics of wealth in Eq. (5A.18) collapse to those in the original market with $\nu = 0$, where the agent is constrained in K ($p \in K$), such that the second line follows, because $(p_{\hat{\nu}}, c_{\hat{\nu}})$ are at this stage not optimally chosen.

The third line of (5A.23) follows by Eq. (5A.21) and (5A.20). The fourth line is the definition of $\text{Val}_\nu(x)$. Combining the first inequality in (5A.22) with Eq. (5A.23) leaves,

$$\text{Val}_{\hat{\nu}}(x) = \text{Val}(x;K) = \inf_{\nu \in \tilde{K}} (\text{Val}_\nu(x)),$$

where the last inequality follows by the second inequality of (5A.22).

5.17 Appendix 6: Topics on jumps

5.17.1 The Radon-Nikodym derivative

This appendix derives, heuristically, results about Radon-Nikodym derivatives for jump-diffusion processes. Precise mathematical details can be found in Brémaud (1981). Consider the jump times $0 < \tau_1 < \tau_2 < \cdots < \tau_n = \hat{T}$. The probability of a jump in a neighborhood of τ_i is $v(\tau_i)d\tau$. To define the same probability under the risk-neutral world, write $v^Q(\tau_i)d\tau$ under Q, and set $v^Q = v\lambda^J$, for some λ^J. The probability that no-jump would occur between any two adjacent random points τ_{i-1} and τ_i *and* a jump would at time τ_{i-1} is, for $i \geq 2$, proportional to:

$$v(\tau_{i-1})e^{-\int_{\tau_{i-1}}^{\tau_i} v_u du} \quad \text{under } P,$$

and to

$$v^Q(\tau_{i-1})e^{-\int_{\tau_{i-1}}^{\tau_i} v_u^Q du} = v(\tau_{i-1})\lambda^J(\tau_{i-1})e^{-\int_{\tau_{i-1}}^{\tau_i} v_u\lambda_u^J du} \quad \text{under } Q.$$

As explained in Section 5.7, these are in fact densities of time intervals elapsing from one arrival to the next one.

Next, let A be the event of marks at time $\tau_1, \tau_2, \cdots, \tau_n$. The Radon-Nikodym derivative is the likelihood ratio of the two probabilities Q and P of A:

$$\frac{Q(A)}{P(A)} = \frac{e^{-\int_t^{\tau_1} v_u\lambda_u^J du} \cdot v(\tau_1)\lambda^J(\tau_1)e^{-\int_{\tau_1}^{\tau_2} v_u\lambda_u^J du} \cdot v(\tau_2)\lambda^J(\tau_2)e^{-\int_{\tau_2}^{\tau_3} v_u\lambda_u^J du} \cdot \ \cdots}{e^{-\int_t^{\tau_1} v_u du} \cdot v(\tau_1)e^{-\int_{\tau_1}^{\tau_2} v_u du} \cdot v(\tau_2)e^{-\int_{\tau_2}^{\tau_3} v_u du} \cdot \ \cdots},$$

where we have used the fact that given that at $\tau_0 = t$, there are no-jumps, the probability that no-jumps would occur from t to τ_1 is $e^{-\int_t^{\tau_1} v_u du}$ under P, and $e^{-\int_t^{\tau_1} v_u\lambda_u^J du}$ under Q. Simple algebra yields,

$$\begin{aligned}
\frac{Q(A)}{P(A)} &= \lambda^J(\tau_1) \cdot \lambda^J(\tau_2) \cdot e^{-\int_t^{\tau_1} v_u\left(\lambda_u^J - 1\right)du} \cdot e^{-\int_{\tau_1}^{\tau_2} v_u\left(\lambda_u^J - 1\right)du} e^{-\int_{\tau_2}^{\tau_3} v_u\left(\lambda_u^J - 1\right)du} \cdot \ \cdots \\
&= \prod_{i=1}^{n} \lambda^J(\tau_i) \cdot e^{-\int_t^{\tau_n} v_u\left(\lambda_u^J - 1\right)du} \\
&= \exp\left[\ln\left(\prod_{i=1}^{n} \lambda^J(\tau_i) \cdot e^{-\int_t^{\tau_n} v_u\left(\lambda_u^J - 1\right)du}\right)\right] \\
&= \exp\left(\sum_{i=1}^{n} \ln \lambda^J(\tau_i) - \int_t^{\tau_n} v_u\left(\lambda_u^J - 1\right)du\right) \\
&= \exp\left(\int_t^{\hat{T}} \ln \lambda_u^J dZ_u - \int_t^{\hat{T}} v_u\left(\lambda_u^J - 1\right)du\right),
\end{aligned}$$

where the last equality follows from the definition of the Stieltjes integral.

Consider, finally, the following definition. Let M be a martingale. The unique solution to the equation:

$$L_\tau = 1 + \int_t^\tau L_u dM_u,$$

is named the *Doléans-Dade exponential semimartingale* and is denoted as $\mathcal{E}(M)$. We now turn to the arbitrage restrictions arising whilst dealing with asset prices driven by jump-diffusion processes.

5.17.2 Arbitrage restrictions

As in the main text, let now S be the price of a primitive asset, solution to:

$$\begin{aligned}\frac{dS}{S} &= bd\tau + \sigma dW + \ell\mathcal{S}dZ \\ &= bd\tau + \sigma dW + \ell\mathcal{S}\left(dZ - vd\tau\right) + \ell\mathcal{S}vd\tau \\ &= \left(b + \ell\mathcal{S}v\right)d\tau + \sigma dW + \ell\mathcal{S}\left(dZ - vd\tau\right).\end{aligned}$$

Next, define

$$d\tilde{Z} = dZ - v^Q d\tau \ \left(v^Q = v\lambda^J\right); \quad d\tilde{W} = dW + \lambda d\tau.$$

Both $\tilde{Z}$ and $\tilde{W}$ are Q-martingales. We have:

$$\frac{dS}{S} = \left(b + \ell\mathcal{S}v^Q - \sigma\lambda\right)d\tau + \sigma d\tilde{W} + \ell\mathcal{S}d\tilde{Z}.$$

The characterization of the equivalent martingale measure for the discounted price is given by the following Radon-Nikodym density of Q with respect to P:

$$\frac{dQ}{dP} = \mathcal{E}\left(-\int_t^T \lambda_\tau dW_\tau + \int_t^T \left(\lambda_\tau^J - 1\right)\left(dZ_\tau - v_\tau\right)d\tau\right),$$

where $\mathcal{E}\left(\cdot\right)$ is the Doléans-Dade exponential semimartingale, and so:

$$b = r + \sigma\lambda - \ell v^Q E_{\mathcal{S}}(\mathcal{S}) = r + \sigma\lambda - \ell v\lambda^J E_{\mathcal{S}}(\mathcal{S}).$$

Clearly, markets are incomplete here. It is possible to show that if $\mathcal{S}$ is deterministic, a representative agent with utility function $u(x) = \frac{x^{1-\gamma}-1}{1-\gamma}$ makes $\lambda^J(\mathcal{S}) = (1+\mathcal{S})^{-\gamma}$.

5.17.3 State price density: introduction

We have:

$$L_T = \exp\left(-\int_t^T v_\tau\left(\lambda_\tau^J - 1\right)d\tau + \int_t^T \ln\lambda_\tau^J dZ_\tau\right).$$

The objective here is to use Itô's lemma for jump processes to express L in differential form. Define the jump process y as:

$$y_\tau \equiv -\int_t^\tau v_u\left(\lambda_u^J - 1\right)du + \int_t^\tau \ln\lambda_u^J dZ_u.$$

In terms of y, L is $L_\tau = l(y_\tau)$ with $l(y) = e^y$. We have:

$$\begin{aligned}dL_\tau &= -e^{y_\tau}v_\tau\left(\lambda_\tau^J - 1\right)d\tau + \left(e^{y_\tau + \text{jump}} - e^{y_\tau}\right)dZ_\tau \\ &= -e^{y_\tau}v_\tau\left(\lambda_\tau^J - 1\right)d\tau + e^{y_\tau}\left(e^{\ln\lambda_\tau^J} - 1\right)dZ_\tau\end{aligned}$$

or,

$$\frac{dL_\tau}{L_\tau} = -v_\tau\left(\lambda_\tau^J - 1\right)d\tau + \left(\lambda_\tau^J - 1\right)dZ_\tau = \left(\lambda_\tau^J - 1\right)\left(dZ_\tau - v_\tau d\tau\right).$$

The general case (with stochastic distribution) is covered in the following subsection.

5.17.4 State price density: general case

Assume that the primitive is:

$$dx_t = \mu\left(x_{\tau-}\right)d\tau + \sigma\left(x_{\tau_-}\right)dW_\tau + dZ_\tau,$$

and let u denote the price of a derivative. Introduce the P-martingale,

$$dM_\tau = dZ_\tau - v(x_\tau)d\tau.$$

By Itô's lemma for jump-diffusion processes,

$$\begin{aligned}\frac{du(x_\tau,\tau)}{u(x_{\tau_-},\tau)} &= \mu^u(x_{\tau_-},\tau)d\tau + \sigma^u(x_{\tau_-},\tau)dW_\tau + J^u\left(\Delta x,\tau\right)dZ_\tau \\ &= \left(\mu^u(x_{\tau_-},\tau) + v(x_{\tau_-})J^u\left(\Delta x,\tau\right)\right)d\tau + \sigma^u(x_{\tau_-},\tau)dW_\tau + J^u\left(\Delta x,\tau\right)dM_\tau,\end{aligned}$$

where $\mu^u = \frac{1}{u}\left(\frac{\partial}{\partial t} + L\right)u$, $\sigma^u = \frac{1}{u}\left(\frac{\partial u}{\partial x}\sigma\right)$, $\frac{\partial}{\partial t} + L$ is the generator for pure diffusion processes and, finally:

$$J^u\left(\Delta x,\tau\right) \equiv \frac{u(x_\tau,\tau) - u(x_{\tau_-},\tau)}{u(x_{\tau_-},\tau)}.$$

To generalize the steps made to deal with the standard diffusion case, let

$$d\tilde{W} = dW + \lambda d\tau, \quad d\tilde{Z} = dZ - v^Q d\tau.$$

We wish to find restrictions on both λ and v^Q, such that both $\tilde{W}$ and $\tilde{Z}$ are Q-martingales. Let J^ξ be the jump component for the state price density ξ:

$$\frac{d\xi_\tau}{\xi_{\tau_-}} = -\lambda(x_{\tau_-})dW_\tau + J^\xi\left(\Delta x,\tau\right)dM_\tau, \quad \xi_t = 1.$$

We shall show that:

$$v^Q = v\left(1 + J^\xi\right).$$

Note that in this case,

$$\frac{d\xi_\tau}{\xi_{\tau_-}} = -\lambda(x_{\tau_-})dW_\tau + \left(\lambda^J - 1\right)dM_\tau, \quad \xi_t = 1,$$

a clear generalization of the pure diffusion case.

As for the derivative price:

$$\begin{aligned}\frac{du}{u} &= \left(\mu^u + vJ^u\right)d\tau + \sigma^u dW + J^u\left(dZ - vd\tau\right) \\ &= \left(\mu^u + v^Q J^u - \sigma^u\lambda\right)d\tau + \sigma^u d\tilde{W} + J^u d\tilde{Z} \\ &= \left(\mu^u + v(1+J^\xi)J^u - \sigma^u\lambda\right)d\tau + \sigma^u d\tilde{W} + J^u d\tilde{Z}.\end{aligned}$$

Finally, by the Q-martingale property of the discounted u,

$$\mu^u - r = \sigma^u\lambda - v^Q \cdot E_{\Delta x}\left(J^u\right) = \sigma^u\lambda - v \cdot E_{\Delta x}\left((1+J^\xi)J^u\right),$$

where $E_{\Delta x}$ is taken with respect to the jump-size distribution, which is the same under Q and P.

PROOF THAT $v^Q = v\left(1 + J^\xi\right)$. As usual, the state-price density ξ has to be a P-martingale in order to be able to price bonds (in addition to all other assets). In addition, ξ clearly "depends" on W and Z. Therefore, it satisfies:

$$\frac{d\xi_\tau}{\xi_{\tau_-}} = -\lambda(x_{\tau_-})dW_\tau + J_\xi\left(\Delta x, \tau\right) dM_\tau, \quad \xi_t = 1.$$

We wish to find v^Q in $d\tilde{Z} = dZ - v^Q d\tau$ such that $\tilde{Z}$ is a Q-martingale, viz

$$\tilde{Z}_\tau = \mathbb{E}_\tau(\tilde{Z}_T),$$

i.e.,

$$\mathbb{E}(\tilde{Z}_t) = \frac{E\left(\xi_T \cdot \tilde{Z}_T\right)}{\xi_t} = \tilde{Z}_t \;\Leftrightarrow\; \xi_t\tilde{Z}_t = E[\xi_T\tilde{Z}_T],$$

i.e.,

$$\xi_t\tilde{Z}_t \text{ is a } P\text{-martingale.}$$

By Itô's lemma,

$$\begin{aligned}
d(\xi\tilde{Z}) &= d\xi \cdot \tilde{Z} + \xi \cdot d\tilde{Z} + d\xi \cdot d\tilde{Z} \\
&= d\xi \cdot \tilde{Z} + \xi\left(dZ - v^Q d\tau\right) + d\xi \cdot d\tilde{Z} \\
&= d\xi \cdot \tilde{Z} + \xi[\underbrace{dZ - vd\tau}_{dM} + \left(v - v^Q\right) d\tau] + d\xi \cdot d\tilde{Z} \\
&= d\xi \cdot \tilde{Z} + \xi \cdot dM + \xi\left(v - v^Q\right) d\tau + d\xi \cdot d\tilde{Z}.
\end{aligned}$$

Because ξ, M and $\xi\tilde{Z}$ are P-martingales,

$$\forall T,\ \ 0 = E\left(\int_t^T \xi_\tau \cdot \left(v_\tau - v_\tau^Q\right) d\tau + \int_t^T d\xi_\tau \cdot d\tilde{Z}_\tau\right).$$

But

$$d\xi \cdot d\tilde{Z} = \xi\left(-\lambda dW + J^\xi dM\right)\left(dZ - v^Q d\tau\right) = \xi\left[-\lambda dW + J^\xi\left(dZ - vd\tau\right)\right]\left(dZ - v^Q d\tau\right),$$

and since $(dZ)^2 = dZ$,

$$E(d\xi \cdot d\tilde{Z}) = \xi \cdot J^\xi v \cdot d\tau,$$

and the previous condition collapses to:

$$\forall T,\ \ 0 = E\left(\int_t^T \xi_\tau \cdot \left(v_\tau - v_\tau^Q + J^\xi(\Delta x)v_\tau\right) d\tau\right),$$

implying that

$$v_\tau^Q = v_\tau\left(1 + J^\xi(\Delta x)\right),$$

$\|$

References

Arnold, L. (1974): *Stochastic Differential Equations: Theory and Applications*, New York: Wiley.

Bajeux-Besnainou, I. and J.-C. Rochet (1996): "Dynamic Spanning: Are Options an Appropriate Instrument?" *Mathematical Finance* 6, 1-16.

Black, F. and M. Scholes (1973): "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81, 637-659.

Breeden, D.T. (1979): "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities." *Journal of Financial Economics* 7, 265-296.

Brémaud, P. (1981): *Point Processes and Queues*: *Martingale Dynamics*. Berlin: Springer Verlag.

Campbell, J.Y. and L.M. Viceira (2002): *Strategic Asset Allocation*. Oxford: Oxford University Press.

Chacko, G. and L.M. Viceira (2005): "Dynamic Consumption and Portfolio Choice with Stochastic Volatility in Incomplete Markets." *Review of Financial Studies* 18, 1369-1402.

Cox, J.C., J.E. Ingersoll and S.A. Ross (1985a): "An Intertemporal General Equilibrium Model of Asset Prices." *Econometrica* 53, 363-384.

Cox, J.C., J.E. Ingersoll and S.A. Ross (1985b): "A Theory of the Term Structure of Interest Rates." Econometrica 53, 385-407.

Cvitanić, J. and I. Karatzas (1992): "Convex Duality in Constrained Portfolio Optimization." *Annals of Applied Probability* 2, 767-818.

Föllmer, H. and M. Schweizer (1991): "Hedging of Contingent Claims under Incomplete Information." In: Davis, M. and R. Elliott (Editors): *Applied Stochastic Analysis*. New York: Gordon & Breach, 389-414.

Friedman, A. (1975): *Stochastic Differential Equations and Applications* (Vol. I). New York: Academic Press.

Lucas, R. E. (1978): "Asset Prices in an Exchange Economy." *Econometrica* 46, 1429-1445.

Harrison, J.M. and S. Pliska (1983): "A Stochastic Calculus Model of Continuous Trading: Complete Markets." *Stochastic Processes and Their Applications* 15, 313-316.

Harrison, J.M, R. Pitbladdo and S.M. Schaefer (1984): Continuous Price Processes in Frictionless Markets Have Infinite Variation." *Journal of Business* 57, 353-365.

He, H. and N. Pearson (1991): "Consumption and Portfolio Policies with Incomplete Markets and Short-Sales Constraints: The Infinite Dimensional Case." *Journal of Economic Theory* 54, 259-304.

Huang, C-F. (1987): "An Intertemporal General Equilibrium Asset Pricing Model: the Case of Diffusion Information." *Econometrica* 55, 117-142.

Jeanblanc-Picqué, M. and A.N. Shirayev (1995): "Optimization of the Flow of Dividends." *Russian Mathematical Surveys* 50, 257-277.

Karatzas, I. and S.E. Shreve (1991): *Brownian Motion and Stochastic Calculus*. New York: Springer Verlag.

Krugman, P. (1991): "Target Zones and Exchange Rate Dynamics." *Quarterly Journal of Economics* 106, 669-682.

Lewis, A.L. (2000): *Option Valuation under Stochastic Volatility*. Finance Press: Newport Beach, CA.

Longstaff, F.A. and E.S. Schwartz (1992): "Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model." *Journal of Finance* 47, 1259-1282.

McDonald, R.L. and D.R. Siegel (1986): "The Value of Waiting to Invest." *Quarterly Journal of Economics* 101: 707-727.

Merton, R.C. (1969): "Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case." *Review of Economics and Statistics* 51, 247-257.

Merton, R.C. (1971): "Optimum Consumption and Portfolio Rules in a Continuous-Time Model." *Journal of Economic Theory* 3, 373-413.

Merton, R.C. (1973): "An Intertemporal Capital Asset Pricing Model." *Econometrica* 41, 867-887.

Mikosch, T. (1998): *Elementary Stochastic Calculus with Finance in View*. Singapore: World Scientific.

Pham, H. and N. Touzi (1996): "Equilibrium State Prices in a Stochastic Volatility Model." *Mathematical Finance* 6, 215-236.

Revuz, D. and M. Yor (1999): *Continuous Martingales and Brownian Motion*. New York: Springer Verlag.

Romano, M. and N. Touzi (1997): "Contingent Claims and Market Completeness in a Stochastic Volatility Model." *Mathematical Finance* 7, 399-412.

Ross, S.A. (1976): "Options and Efficiency." *Quarterly Journal of Economics* 90, 75-89.

Shreve, S. (1991): "A Control Theorist's View of Asset Pricing." In: Davis, M. and R. Elliot (Editors): *Applied Stochastic Analysis*. New York: Gordon & Breach, 415-445.

Steele, J.M. (2001): *Stochastic Calculus and Financial Applications*. New York: Springer-Verlag.

6
Taking models to data

6.1 Introduction

This chapter surveys methods to estimate and test dynamic models of asset prices. It begins with foundational issues on identification, specification and testing. Then, it surveys classical estimation and testing methodologies such as the Method of Moments, where the number of moment conditions equals the dimension of the parameter vector (Pearson, 1894); Maximum Likelihood (ML) (Gauss, 1816; Fisher, 1912); the Generalized Method of Moments (GMM), where the number of moment conditions exceeds the dimension of the parameter vector, leading to the minimum chi-squared (Neyman and Pearson, 1928; Hansen, 1982); and, finally, the relatively more recent developments relying on simulations, which aim to implement ML and GMM estimation for models that are analytically quite complex, but that can be simulated. The chapter concludes with an illustration of how joint estimation of fundamentals and asset prices in arbitrage-free models can asymptotically lead to statistical efficiency.

6.2 Data generating processes

6.2.1 Basics

Given is a multidimensional stochastic process y_t, a *data generating process* (DGP). While we do not know the probability distribution underlying y_t, we use the available data to get insights into its nature. A few definitions. A DGP is a conditional law, say the law of y_t given the set of past values $\underline{y_{t-1}} = \{y_{t-1}, y_{t-2}, \cdots\}$, and some exogenous [define] variable z, with $\underline{z_t} = \{z_t, z_{t-1}, z_{t-2}, \cdots\}$,

$$\text{DGP} : \ell_0(\, y_t |\, x_t),$$

where $x_t = (\underline{y_{t-1}},\ \underline{z_t})$, and ℓ_0 denotes the conditional density of the data, the true law. Then, we have three basic definitions. First, we define a *parametric model* as a set of conditional laws for y_t, indexed by a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^p$,

$$(M) = \{\ell\,(\, y_t |\, x_t; \theta)\,,\ \theta \in \Theta \subseteq \mathbb{R}^p\}\,.$$

Second, we say that the model (M) is *well-specified* if,

$$\exists \theta_0 \in \Theta : \ell\left(y_t|\, x_t; \theta_0\right) = \ell_0\left(y_t|\, x_t\right).$$

Third, we say that the model (M) is *identifiable* if θ_0 is unique. The main goal of this chapter is to review tools aimed at drawing inference about the true parameter θ_0, given the observations.

6.2.2 Restrictions on the DGP

The previous definition of DGP is too rich to be of practical relevance. This chapter deals with estimation methods applying to DGPs satisfying a few restrictions. Two fundamental restrictions are usually imposed on the DGP:

- Restrictions on the *heterogeneity* of the stochastic process, which lead to *stationary* random processes.
- Restrictions on the *memory* of the stochastic process, which pave the way to *ergodic* processes.

6.2.2.1 Stationarity

Stationary processes describe phenomena leading to long run equilibria, in some statistical sense: as time unfolds, the probability generating the observations settles down to some "long-run" probability density, a time invariant probability. As explained in Chapter 3, in the early 1980s, theorists begun to define a long-run equilibrium as a well-defined stationary, probability distribution generating economic outcomes. We have two notions of stationarity: (i) Strong, or strict, stationarity. Definition: Homogeneity in law; (ii) Weak stationarity, or stationarity of order p. Definition: Homogeneity in moments.

[In progress]

Even with stationary DGP, there might be situations where the number of parameters to be estimated increases with the sample size. As an example, consider two stochastic processes: one, for which $cov(y_t, y_{t+\tau}) = \tau^2$; and another, for which $cov(y_t, y_{t+\tau}) = \exp\left(-|\tau|\right)$. In both cases, the DGP is stationary. Yet for the first process, the dependence increases with τ, and for the second, the dependence decreases with τ. As this simple example reveals, a stationary stochastic process may have "long memory." "Ergodicity" further restricts DGP, so as to make this memory play a more limited role.

6.2.2.2 Ergodicity

We shall deal with DGPs where the dependence between y_{t_1} and y_{t_2} decreases with $|t_2 - t_1|$. To introduce some concepts and notation, say two events A and B are independent, when $P(A \cap B) = P(A)P(B)$. A stochastic process is *asymptotically independent* if, for some function β_τ,

$$\beta_\tau \geq \left|F(y_{t_1}, \cdots, y_{t_n}, y_{t_1+\tau}, \cdots, y_{t_n+\tau}) - F\left(y_{t_1}, \cdots, y_{t_n}\right) F\left(y_{t_1+\tau}, \cdots, y_{t_n+\tau}\right)\right|,$$

we also have that $\lim_{\tau\to\infty} \beta_\tau \to 0$. A stochastic process is *p-dependent* if $\forall \tau \leq p$, $\beta_\tau \neq 0$. A stochastic process is *asymptotically uncorrelated* if there exists ρ_τ such that for all t, $\rho_\tau \geq cov(y_t, y_{t+\tau})/\sqrt{var(y_t) \cdot var(y_{t+\tau})}$, and that $0 \leq \rho_\tau \leq 1$ with $\sum_{\tau=0}^{\infty} \rho_\tau < \infty$. For example, $\rho_\tau = \tau^{-(1+\delta)}$, $\delta > 0$, in which case $\rho_\tau \downarrow 0$ as $\tau \uparrow \infty$.

Let $\mathbb{B}_1^t$ denote the σ-algebra generated by $\{y_1, \cdots, y_t\}$ and $A \in \mathbb{B}_{-\infty}^t$, $B \in \mathbb{B}_{t+\tau}^{\infty}$, and define:

$$\alpha_\tau = \sup_\tau \left|P(A \cap B) - P(A)P(B)\right|, \quad \varphi(\tau) = \sup_\tau \left|P(B \mid A) - P(B)\right|,\ P(A) > 0.$$

We say that (i) y is *strongly mixing*, or α-*mixing* if $\lim_{\tau\to\infty}\alpha_\tau \to 0$; (ii) y is *uniformly mixing* if $\lim_{\tau\to\infty}\varphi_\tau \to 0$. Clearly, a uniformly mixing process is also strongly mixing. A second order stationary process is *ergodic* if $\lim_{T\to\infty}\sum_{\tau=1}^{T} cov\left(y_t, y_{t+\tau}\right) < \infty$. If a second order stationary process is strongly mixing, it is also ergodic.

6.2.3 Parameter estimators

Consider an estimator of the parameter vector θ of the model,

$$(M) = \left\{\ell(y_t|\, x_t;\theta),\ \theta\in\Theta\subset\mathbb{R}^p\right\}.$$

Naturally, any estimator does necessarily depend on the sample size, which we write as $\hat{\theta}_T \equiv t_T(y)$. Of a given estimator $\hat{\theta}_T$, we say that it is:

- *Correct*, or *unbiased*, if $E(\hat{\theta}_T) = \theta_0$. The difference $E(\hat{\theta}_T) - \theta_0$ is called distortion, or bias.
- *Weakly consistent* if $\text{plim}\hat{\theta}_T = \theta_0$. And *strongly consistent* if $\hat{\theta}_T \xrightarrow{a.s.} \theta_0$.

Finally, an estimator $\hat{\theta}_T^{(1)}$ is more *efficient* than another estimator $\hat{\theta}_T^{(2)}$ if, for any vector of constants c, we have that $c^\top \cdot var(\hat{\theta}_T^{(1)}) \cdot c < c^\top \cdot var(\hat{\theta}_T^{(2)}) \cdot c$.

6.2.4 Basic properties of density functions

We have T observations $y_T^1 = \{y_1, \cdots, y_T\}$. Suppose these observations are the realization of a T-dimensional random variable with joint density, $f\left(\tilde{y}_1, \cdots, \tilde{y}_T; \theta\right) = f\left(\tilde{y}_1^T; \theta\right)$. We have momentarily put tildes on y_i, to emphasize that we view each $\tilde{y}_i$ as a random variable.[1] However, to ease notation, from now on, we write y_i instead of $\tilde{y}_i$. By construction, $\int f\left(y|\,\theta\right) dy \equiv \int\cdots\int f\left(y_1^T\big|\,\theta\right) dy_1^T = 1$ or,

$$\forall\theta\in\Theta,\ \int f\left(y;\theta\right) dy = 1.$$

Now suppose that the support of y doesn't depend on θ. Under regularity conditions,

$$\nabla_\theta \int f\left(y;\theta\right) dy = \int \nabla_\theta f\left(y;\theta\right) dy = \mathbf{0}_p,$$

where $\mathbf{0}_p$ is a column vector of zeros in $\mathbb{R}^p$. Moreover, for all $\theta\in\Theta$,

$$\mathbf{0}_p = \int \nabla_\theta f\left(y;\theta\right) dy = E_\theta\left[\nabla_\theta \ln f\left(y;\theta\right)\right]. \tag{6.1}$$

Finally, we have,

$$\begin{aligned}\mathbf{0}_{p\times p} &= \nabla_\theta \int \left[\nabla_\theta \ln f\left(y;\theta\right)\right] f\left(y;\theta\right) dy \\ &= \int \left[\nabla_{\theta\theta} \ln f\left(y;\theta\right)\right] f\left(y;\theta\right) dy + \int \left|\nabla_\theta \ln f\left(y;\theta\right)\right|_2 f\left(y;\theta\right) dy,\end{aligned}$$

where $|x|_2$ denotes the outer product, i.e. $|x|_2 = x\cdot x^\top$. Hence, by Eq. (6.1),

$$E_\theta\left[\nabla_{\theta\theta} \ln f\left(y;\theta\right)\right] = -E_\theta\left|\nabla_\theta \ln f\left(y;\theta\right)\right|_2 = -var_\theta\left[\nabla_\theta \ln f\left(y;\theta\right)\right] \equiv -\mathcal{J}(\theta),\ \forall\theta\in\Theta.$$

The matrix $\mathcal{J}$ is known as the Fisher's information matrix.

[1] Therefore, we follow a classical perspective. A Bayesian statistician would view the sample as *given*. We do not review Bayesian methods in this chapter.

6.2.5 The Cramer-Rao lower bound

Let $t(y)$ some unbiased estimator of θ, and set the dimension of the parameter space to $p = 1$. We have,

$$E\left[t(y)\right] = \int t(y) f\left(y;\theta\right) dy.$$

Under regularity conditions,

$$\nabla_\theta E\left[t(y)\right] = \int t(y)\left[\nabla_\theta \ln f\left(y;\theta\right)\right] f\left(y;\theta\right) dy = cov\left(t(y), \nabla_\theta \ln f\left(y;\theta\right)\right).$$

By Cauchy-Schwartz inequality, $\left[cov\left(t(y), \nabla_\theta \ln f\left(y;\theta\right)\right)\right]^2 \leq var\left[t(y)\right] \cdot var\left[\nabla_\theta \ln f\left(y;\theta\right)\right]$. Therefore,

$$\left[\nabla_\theta E\left(t(y)\right)\right]^2 \leq var\left[t(y)\right] \cdot var\left[\nabla_\theta \ln f\left(y;\theta\right)\right] = -var\left[t(y)\right] \cdot E\left[\nabla_{\theta\theta} \ln f\left(y;\theta\right)\right].$$

But if $t(y)$ is unbiased, or $E\left[t(y)\right] = \theta$,

$$var\left[t(y)\right] \geq \left[-E\left(\nabla_\theta \ln f\left(y;\theta\right)\right)\right]^{-1} \equiv \mathcal{J}\left(\theta\right)^{-1}.$$

This is the celebrated Cramer-Rao bound. The same results holds in the multidimensional case, through a mere change in notation (see, e.g., Amemiya, 1985, p. 14-17).

6.3 Maximum likelihood estimation

6.3.1 Basics

The density of the data, $f(y_1^T \big| \theta)$, maps every possible sample and parameter values of θ on to positive numbers, the "likelihood" of occurence of any given sample, given the parameter θ: $\mathbb{R}^{nT} \times \Theta \mapsto \mathbb{R}_+$. We trace the joint density of the entire sample through a thought experiment, in which we change the sample y_1^T. So the sample is viewed as the realization of a random variable, a view opposite to the Bayesian perspective. We ask: Which value of θ makes the sample we observed the most likely to have occurred? We introduce the "likelihood function," $L(\theta|\, y_1^T) \equiv f(y_1^T;\theta)$. It is the function $\theta \mapsto f(y;\theta)$ for y_1^T given and equal to $\bar{y}$, say:

$$L(\theta|\, \bar{y}) \equiv f(\bar{y};\theta).$$

Then, we maximize $L(\theta|\, y_1^T)$ with respect to θ. That is, we look for the value of θ, which maximizes the probability to observe the sample we have effectively observed. The resulting estimator is called *maximum likelihood estimator* (MLE). As we shall see, the MLE attains the Cramer-Rao lower bound, provided the model is not misspecified.

6.3.2 Factorizations

Consider a series of events $\{A_i\}$. In the Appendix, we show that,

$$\Pr\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \Pr\left(A_i \left| \bigcap_{j=1}^{i-1} A_j \right.\right). \tag{6.2}$$

By Eq. (6.2), then, the MLE satisfies:

$$\hat{\theta}_T = \arg\max_{\theta\in\Theta} L_T(\theta) = \arg\max_{\theta\in\Theta}\left(\frac{1}{T}\ln L_T(\theta)\right),$$

where, assuming IID data,

$$\ln L_T(\theta) \equiv \ln\prod_{t=1}^{T} f\left(y_t\left|y_1^{t-1};\theta\right.\right) = \sum_{t=1}^{T}\ln f\left(y_t\left|y_1^{t-1};\theta\right.\right) \equiv \sum_{t=1}^{T}\ln f\left(y_t;\theta\right) \equiv \sum_{t=1}^{T}\ell_t(\theta), \qquad (6.3)$$

and $\ell_t(\theta)$ is the "log-likelihood" of a single observation.

6.3.3 Asymptotic properties

We consider the i.i.d. case only, as in Eq. (6.3). Moreover, we provide heuristic arguments, leaving more rigorous proofs and general results in the Appendix.

6.3.3.1 The limiting problem

The MLE satisfies the following first order conditions,

$$\mathbf{0}_p = \left.\nabla_\theta \ln L_T\left(\theta\right)\right|_{\theta=\hat{\theta}_T} \equiv \nabla_\theta \ln L_T(\hat{\theta}_T).$$

Consider a Taylor expansion of the first order conditions around θ_0,

$$\mathbf{0}_p = \nabla_\theta \ln L_T(\hat{\theta}_T) \overset{d}{=} \nabla_\theta \ln L_T(\theta_0) + \nabla_{\theta\theta} \ln L_T(\theta_0)(\hat{\theta}_T - \theta_0), \qquad (6.4)$$

where the notation $x_T \overset{d}{=} y_T$ means that the difference $x_T - y_T = o_p\left(1\right)$, and θ_0 is defined as the solution to the limiting problem,

$$\theta_0 = \arg\max_{\theta\in\Theta}\left[\lim_{T\to\infty}\left(\frac{1}{T}\ln L_T(\theta)\right)\right] = \arg\max_{\theta\in\Theta}\left[E\left(\ell\left(\theta\right)\right)\right],$$

and, finally, ℓ satisfies regularity conditions needed to ensure that,

$$\theta_0 : E\left[\nabla_\theta \ell\left(\theta_0\right)\right] = \mathbf{0}_p.$$

To show that this is indeed the solution, suppose θ_0 is identified; that is, $\theta \neq \theta_0$ and $\theta, \theta_0 \in \Theta \Leftarrow f(y|\,\theta) \neq f(y|\,\theta_0)$. Suppose, further, that for each $\theta \in \Theta$, $E_\theta\left[\ln f(y|\,\theta)\right] < \infty$. Then, we have that $\theta_0 = \arg\max_{\theta\in\Theta} E_\theta\left[\ln f(y|\,\theta)\right]$, and this value of θ is unique. The proof is, indeed, very simple. We have,

$$\begin{aligned} E_{\theta_0}\left[-\ln\left(\frac{f(y|\,\theta)}{f(y|\,\theta_0)}\right)\right] &> -\ln E_{\theta_0}\left(\frac{f(y|\,\theta)}{f(y|\,\theta_0)}\right) \\ &= -\ln\int\frac{f(y|\,\theta)}{f(y|\,\theta_0)}f(y|\,\theta_0)dy \\ &= -\ln\int f(y|\,\theta)dy = 0. \end{aligned}$$

6.3.3.2 Consistency and asymptotic normality

Provided the model is well-specified, we have that $\hat{\theta}_T \xrightarrow{p} \theta_0$ and even $\hat{\theta}_T \xrightarrow{a.s.} \theta_0$, under regularity conditions. One example of conditions required to obtain weak consistency is that the following uniform weak law of large numbers holds,

$$\lim_{T\to\infty} \Pr\left[\sup_{\theta\in\Theta} |\ell_T(\theta) - E(\ell(\theta))|\right] \to 0.$$

Next, consider again the asymptotic expansion in Eq. (6.4), which can be elaborated, so as to have,

$$\begin{aligned}\sqrt{T}(\hat{\theta}_T - \theta_0) &\overset{d}{=} -\left[\frac{1}{T}\nabla_{\theta\theta} \ln L_T(\theta_0)\right]^{-1} \frac{1}{\sqrt{T}}\nabla_\theta \ln L_T(\theta_0) \\ &= -\left[\frac{1}{T}\sum_{t=1}^{T} \nabla_{\theta\theta}\ell_t(\theta_0)\right]^{-1} \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \nabla_\theta \ell_t(\theta_0).\end{aligned}$$

By the law of large numbers reviewed in the Appendix (weak law no. 1),

$$\frac{1}{T}\sum_{t=1}^{T} \nabla_{\theta\theta}\ell_t(\theta_0) \xrightarrow{p} E_{\theta_0}\left[\nabla_{\theta\theta}\ell_t(\theta_0)\right] = -\mathcal{J}(\theta_0).$$

Therefore, asymptotically,

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \overset{d}{=} \mathcal{J}(\theta_0)^{-1} \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \nabla_\theta \ell_t(\theta_0).$$

We also have,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \nabla_\theta \ell_t(\theta_0) \xrightarrow{d} N(0, \mathcal{J}(\theta_0)).$$

Indeed, let $\overline{\nabla_\theta \ell(\theta_0)}_T = \frac{1}{T}\sum_{t=1}^{T} \nabla_\theta \ell_t(\theta_0)$, and note that $E(\nabla_\theta \ell_t(\theta_0)) = 0$. Then, by the central limit theorem reviewed in the Appendix:

$$\frac{1}{\sqrt{T}} \frac{\sum_{t=1}^{T} \nabla_\theta \ell_t(\theta_0)}{\sqrt{var\left[\nabla_\theta \ell_t(\theta_0)\right]}} = \frac{\sqrt{T}\left(\overline{\nabla_\theta \ell(\theta_0)}_T - E(\nabla_\theta \ell_t(\theta_0))\right)}{\sqrt{var\left[\nabla_\theta \ell_t(\theta_0)\right]}},$$

where, for each t, $var\left[\nabla_\theta \ell_t(\theta_0)\right] = \mathcal{J}(\theta_0)$.

Finally, by the Slutzky's theorem reviewed in the Appendix,

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N\left(0, \mathcal{J}(\theta_0)^{-1}\right).$$

Therefore, the ML estimator attains the Cramer-Rao lower bound.

6.4 M-estimators

Consider a function g of the unknown parameters θ. Given a function Ψ, a *M-estimator* of the function $g(\theta)$ is the solution to,

$$\max_{g\in G}\sum_{t=1}^{T}\Psi\left(x_t,y_t;g\right),$$

where y and x are as in Section 6.2.1. We assume that a solution to this problem exists, that it is interior and that it is unique. Let us denote the M-estimator with $\hat{g}_T(x_1^T,y_1^T)$. Naturally, the M-estimator satisfies the following first order conditions,

$$0=\frac{1}{T}\sum_{t=1}^{T}\nabla_g\Psi\left(y_t,x_t;\hat{g}_T(x_1^T,y_1^T)\right).$$

To simplify the presentation, we assume that (x,y) are independent in time, and that they have the same law. By the law of large numbers,

$$\frac{1}{T}\sum_{t=1}^{T}\Psi\left(y_t,x_t;g\right)\xrightarrow{p}\iint\Psi\left(y,x;g\right)dF(x,y)=\iint\Psi\left(y,x;g\right)dF\left(y|\,x\right)dZ\left(x\right)\equiv E_xE_0\left[\Psi\left(y,x;g\right)\right],$$

where E_0 is the expectation operator taken with respect to the true conditional law of y given x and E_x is the expectation operator taken with respect to the true marginal law of x. The limit problem is,

$$g_\infty=g_\infty\left(\theta_0\right)=\arg\max_{g\in G}E_xE_0\left[\Psi\left(y,x;g\right)\right].$$

Under standard regularity conditions,[2] there exists a sequence of M-estimators $\hat{g}_T(x,y)$ converging a.s. to $g_\infty=g_\infty\left(\theta_0\right)$. Under additional regularity conditions, the M-estimator is also asymptotic normal:

THEOREM 6.1: *Let* $\mathcal{I}\equiv E_xE_0\left(\nabla_g\Psi\left(y,x;g_\infty\left(\theta_0\right)\right)\left[\nabla_g\Psi\left(y,x;g_\infty\left(\theta_0\right)\right)\right]^\top\right)$ *and assume that the matrix* $\mathcal{J}\equiv E_xE_0\left[-\nabla_{gg}\Psi\left(y,x;g\right)\right]$ *exists and has an inverse. We have,*

$$\sqrt{T}\left(\hat{g}_T-g_\infty\left(\theta_0\right)\right)\xrightarrow{d}N\left(0,\mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}\right).$$

PROOF (Sketch). The M-estimator satisfies the following first order conditions,

$$\begin{aligned}0&=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla_g\Psi\left(y_t,x_t;\hat{g}_T\right)\\&\overset{d}{=}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla_g\Psi\left(y_t,x_t;g_\infty\right)+\sqrt{T}\left[\frac{1}{T}\sum_{t=1}^{T}\nabla_{gg}\Psi\left(y_t,x_t;g_\infty\right)\right]\cdot\left(\hat{g}_T-g_\infty\right).\end{aligned}$$

[2] G is compact; Ψ is continuous with respect to g and integrable with respect to the true law, for each g; $\frac{1}{T}\sum_{t=1}^{T}\Psi\left(y_t,x_t;g\right)\overset{a.s.}{\to}E_xE_0\left[\Psi\left(y,x;g\right)\right]$ uniformly on G; the limit problem has a unique solution $g_\infty=g_\infty\left(\theta_0\right)$.

By rearranging terms,

$$\begin{aligned}\sqrt{T}\left(\hat{g}_T - g_\infty\right) &\overset{d}{=} \left[-\frac{1}{T}\sum_{t=1}^{T}\nabla_{gg}\Psi\left(y_t, x_t; g_\infty\right)\right]^{-1} \cdot \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla_g\Psi\left(y_t, x_t; g_\infty\right) \\ &\overset{d}{=} \left[E_x E_0\left(-\nabla_{gg}\Psi\left(y, x; g\right)\right)\right]^{-1} \cdot \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla_g\Psi\left(y_t, x_t; g_\infty\right) \\ &\overset{d}{=} \mathcal{J}^{-1} \cdot \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla_g\Psi\left(y_t, x_t; g_\infty\right).\end{aligned}$$

By the limiting problem, $E_x E_0\left[\nabla_g\Psi\left(y, x; g_\infty\right)\right] = 0$. Then, $var\left(\nabla_g\Psi\right) = E\left(\nabla_g\Psi \cdot \left[\nabla_g\Psi\right]^\top\right) = \mathcal{I}$, and, then,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla_g\Psi\left(y_t, x_t; g_\infty\right) \xrightarrow{d} N\left(0, \mathcal{I}\right).$$

The result follows by the Slutzky's theorem and the symmetry of $\mathcal{J}$. ‖

One simple example of M-estimator is the Nonlinear Least Squares estimator,

$$\hat{\theta}_T = \arg\min_{\theta\in\Theta}\sum_{t=1}^{T}\left[y_t - m\left(x_t; \theta\right)\right]^2,$$

for some function m. In this case, $\Psi\left(x, y; \theta\right) = \left[y - m\left(x; \theta\right)\right]^2$.

6.5 Pseudo, or quasi, maximum likelihood

The maximum likelihood estimator is an M-estimator: set $\Psi = \ln L$, the log-likelihood function. Indeed, assume the model is well-specified, in which case $\mathcal{J} = \mathcal{I}$, which confirms we are back to the MLE.

Next, suppose that we implement the MLE to estimate a model, when in fact the model is *misspecified* in that the true DGP $\ell_0\left(y_t \middle| x_t\right)$ does not belong to the family of laws spanned by our model,

$$\ell_0\left(y_t \middle| x_t\right) \notin \left(M\right) = \left\{f\left(y_t \middle| x_t; \theta\right), \quad \theta \in \Theta\right\}.$$

Suppose we insist in maximizing $\Psi = \ln L$, where $L = \sum_t f\left(y_t \middle| x_t; \theta\right)$. In this case,

$$\sqrt{T}(\hat{\theta}_T - \theta_0^*) \xrightarrow{d} N\left(0, \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}\right),$$

where θ_0^* is the "pseudo-true" value,[3] and

$$\mathcal{J} = -E_x E_0\left[\nabla_{\theta\theta}\ln f(y_t|\underline{y_{t-1}}; \theta_0^*)\right], \quad \mathcal{I} = E_x E_0\left(\nabla_\theta \ln f(y_t|\underline{y_{t-1}}; \theta_0^*) \cdot \left[\nabla_\theta \ln f(y_t|\underline{y_{t-1}}; \theta_0^*)\right]^\top\right).$$

In the presence of specification errors, $\mathcal{J} \neq \mathcal{I}$. By comparing the two estimated matrices leads to detect specification errors. Finally, note that in this general case, the variance-covariance

[3] That is, θ_0^* is, clearly, the solution to some misspecified limiting problem. This θ_0^* has an appealing interpretation in terms of some entropy distance minimizer.

matrix $\mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}$ depends on the unknown law of (y_t, x_t). To assess the precision of the estimates of $\hat{g}_T$, one needs to estimate such a variance-covariance matrix. A common practice is to use the following a.s. consistent estimators,

$$\hat{J} = -\frac{1}{T}\sum_{t=1}^{T} \nabla_{gg}\Psi(y_t, x_t; \hat{g}_T), \quad \text{and} \quad \hat{\mathcal{I}} = -\frac{1}{T}\sum_{t=1}^{T} \left(\nabla_g \Psi(y_t, x_t; \hat{g}_T)\left[\nabla_g \Psi(y_t, x_t; \hat{g}_T)^\top\right]\right).$$

6.6 GMM

Economic theory often places restrictions on models that have the following format,

$$E\left[h\left(y_t; \theta_0\right)\right] = \mathbf{0}_q, \tag{6.5}$$

where $h : \mathbb{R}^n \times \Theta \mapsto \mathbb{R}^q$, θ_0 is the true parameter vector, y_t is the n-dimensional vector of the observable variables and $\Theta \subseteq \mathbb{R}^p$. Typically, then, the MLE cannot be used to estimate θ_0. Moreover, MLE requires specifying a density function. Hansen (1982) proposed the following Generalized Method of Moments (GMM) estimation procedure. Consider the sample counterpart to the population in Eq. (6.5),

$$\bar{h}\left(y_1^T; \theta\right) = \frac{1}{T}\sum_{t=1}^{T} h\left(y_t; \theta\right), \tag{6.6}$$

where we have rewritten h as a function of the parameter vector $\theta \in \Theta$. The basic idea of GMM is to find a θ which makes $\bar{h}\left(y_1^\top; \theta\right)$ as close as possible to zero. Precisely, we have,

Definition (GMM estimator): The GMM estimator *is the sequence* $\hat{\theta}_T$ *satisfying*,

$$\hat{\theta}_T = \arg\min_{\theta \in \Theta \subseteq \mathbb{R}^p} \underset{1\times q}{\bar{h}\left(y_1^T; \theta\right)^\top} \cdot \underset{q\times q}{W_T} \cdot \underset{q\times 1}{\bar{h}\left(y_1^T; \theta\right)},$$

where $\{W_T\}$ *is a sequence of weighting matrices, with elements that may depend on the observations.*

When $p = q$, we say the GMM is *just-identified*, and is, simply, the MM, satisfying:

$$\hat{\theta}_T : \bar{h}(y_1^T; \hat{\theta}_T) = \mathbf{0}_q.$$

When $p < q$, we say the GMM estimator imposes *overidentifying* restrictions.

We analyze the i.i.d. case only. Under regularity conditions, there exists a matrix W_T that minimizes the asymptotic variance of the GMM estimator, which satisfies asymptotically,

$$W = \left[\lim_{T\to\infty} T \cdot E\left(\bar{h}(y_1^T; \hat{\theta}_T) \cdot \bar{h}(y_1^T; \hat{\theta}_T)^\top\right)\right]^{-1} \equiv \Sigma_0^{-1}. \tag{6.7}$$

An estimator of Σ_0 can be:

$$\Sigma_T = \frac{1}{T}\sum_{t=1}^{T}\left[h(y_t; \hat{\theta}_T) \cdot h(y_t; \hat{\theta}_T)^\top\right].$$

Note that $\hat{\theta}_T$ depends on the weighting matrix Σ_T, and the weighting matrix Σ_T depends on $\hat{\theta}_T$. Therefore, we need to implement an iterative procedure. The more one iterates, the less likely the final outcome depends on the initial weighting matrix $\Sigma_T^{(0)}$. For example, one can start with $\Sigma_T^{(0)} = \mathbf{I}_q$.

We have:

THEOREM 6.2: *Suppose to be given a sequence of GMM estimators $\hat{\theta}_T$ with weigthing matrix as in Eq. (6.7), and such that: $\hat{\theta}_T \xrightarrow{p} \theta_0$. We have,*

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N\left(\mathbf{0}_p, \left[E\left(h_\theta\right)\Sigma_0^{-1}E\left(h_\theta\right)^\top\right]^{-1}\right), \quad \textit{where } h_\theta \equiv \nabla_\theta h(y;\theta_0).$$

PROOF (Sketch). The assumption that $\hat{\theta}_T \xrightarrow{p} \theta_0$ is easy to check under mild regularity conditions. Moreover, the GMM satisfies,

$$\underset{p\times 1}{\mathbf{0}_p} = \underset{p\times q}{\nabla_\theta \bar{h}(y_1^T;\hat{\theta}_T)}\underset{q\times q}{\Sigma_T^{-1}}\underset{q\times 1}{\bar{h}(y_1^T;\hat{\theta}_T)}. \tag{6.8}$$

Eq. (6.8) confirms that if $p = q$, the GMM satisfies $\hat{\theta}_T : \bar{h}(y_1^T;\hat{\theta}_T) = 0$. Indeed, $\nabla_\theta h \Sigma_T^{-1}$ is full-rank with $p = q$, and Eq. (6.8) can only be satisfied with $\bar{h} = 0$. In the general case, $q > p$, we have,

$$\underset{q\times 1}{\sqrt{T}\bar{h}(y_1^T;\hat{\theta}_T)} = \sqrt{T}\underset{q\times 1}{\bar{h}\left(y_1^T;\theta_0\right)} + \underset{q\times p}{\left[\nabla_\theta \bar{h}\left(y_1^T;\theta_0\right)\right]^\top}\sqrt{T}(\hat{\theta}_T - \theta_0) + o_p(1).$$

By premultiplying both sides of the previous equality by $\nabla_\theta \bar{h}(y_1^T;\hat{\theta}_T)\Sigma_T^{-1}$,

$$\begin{aligned} & \sqrt{T}\nabla_\theta \bar{h}(y_1^T;\hat{\theta}_T)\Sigma_T^{-1}\cdot \bar{h}(y_1^T;\hat{\theta}_T) \\ = \; & \sqrt{T}\nabla_\theta \bar{h}(y_1^T;\hat{\theta}_T)\Sigma_T^{-1}\cdot \bar{h}\left(y_1^T;\theta_0\right) + \nabla_\theta \bar{h}(y_1^T;\hat{\theta}_T)\Sigma_T^{-1}\cdot\left[\nabla_\theta \bar{h}\left(y_1^T;\theta_0\right)\right]^\top\sqrt{T}(\hat{\theta}_T - \theta_0) + o_p(1). \end{aligned}$$

The l.h.s. of this equality is zero by the first order conditions in Eq. (6.8). By rearranging terms,

$$\begin{aligned} \sqrt{T}(\hat{\theta}_T - \theta_0) &\overset{d}{=} -\left(\nabla_\theta \bar{h}\left(y_1^T;\theta_0\right)\Sigma_T^{-1}\left[\nabla_\theta \bar{h}\left(y_1^T;\theta_0\right)\right]^\top\right)^{-1}\nabla_\theta \bar{h}(y_1^T;\hat{\theta}_T)\Sigma_T^{-1}\cdot\sqrt{T}\bar{h}\left(y_1^T;\theta_0\right) \\ &= -\left(\frac{1}{T}\sum_{t=1}^T \nabla_\theta h(y_t;\hat{\theta}_T)\Sigma_T^{-1}\frac{1}{T}\sum_{t=1}^T[\nabla_\theta h(y_t;\hat{\theta}_T)]^\top\right)^{-1}\frac{1}{T}\sum_{t=1}^T \nabla_\theta h(y_t;\hat{\theta}_T)\Sigma_T^{-1}\sqrt{T}\bar{h}\left(y_1^T;\theta_0\right) \\ &\overset{d}{=} -\left(E\left(h_\theta\right)\Sigma_0^{-1}E\left(h_\theta\right)^\top\right)^{-1}E\left(h_\theta\right)\Sigma_0^{-1}\cdot\frac{1}{\sqrt{T}}\sum_{t=1}^T h\left(y_t;\theta_0\right). \end{aligned}$$

We have: $\frac{1}{\sqrt{T}}\sum_{t=1}^T h\left(y_t;\theta_0\right) \xrightarrow{d} N\left(E(h), var(h)\right)$, where, by Eq. (6.5), $E(h) = 0$, and $var(h) = E\left(h\cdot h^\top\right) = \Sigma_0$. Hence:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^T h\left(y_t;\theta_0\right) \xrightarrow{d} N\left(0,\Sigma_0\right).$$

Therefore, $\sqrt{T}(\hat{\theta}_T - \theta_0)$ is asymptotically normal with expectation $\mathbf{0}_p$, and variance,

$$\left(E\left(h_\theta\right)\Sigma_0^{-1}E\left(h_\theta\right)^\top\right)^{-1}E\left(h_\theta\right)\Sigma_0^{-1}\Sigma_0\Sigma_0^{-1}E\left(h_\theta\right)^\top\left(E\left(h_\theta\right)\Sigma_0^{-1}E\left(h_\theta\right)^\top\right)^{\top-1} = \left(E\left(h_\theta\right)\Sigma_0^{-1}E\left(h_\theta\right)^\top\right)^{-1}.$$

‖

A widely used global specification test is that of the celebrated "overidentifying restrictions." Consider the following intuitive result:

$$\sqrt{T}\bar{h}\left(y_1^T;\theta_0\right)^\top \Sigma_0^{-1}\sqrt{T}\bar{h}\left(y_1^T;\theta_0\right)^\top \xrightarrow{d} \chi^2(q).$$

Would we be expecting the same, if we were to replace the true parameter θ_0 with the GMM estimator $\hat{\theta}_T$, which is, anyway, a consistent estimator for θ_0? The anwer is no. Define:

$$\mathcal{C}_T = \sqrt{T}\bar{h}(y_1^T;\hat{\theta}_T)^\top \Sigma_T^{-1}\cdot\sqrt{T}\bar{h}(y_1^T;\hat{\theta}_T).$$

We have,

$$\begin{aligned}
&\sqrt{T}\bar{h}(y_1^T;\hat{\theta}_T) \overset{d}{=} \sqrt{T}\bar{h}\left(y_1^T;\theta_0\right) + \nabla_\theta \bar{h}\left(y_1^T;\theta_0\right)\sqrt{T}(\hat{\theta}_T-\theta_0)\\
&\overset{d}{=} \sqrt{T}\bar{h}\left(y_1^T;\theta_0\right) - \left[\nabla_\theta \bar{h}\left(y_1^T;\theta_0\right)\right]^\top \left[E\left(h_\theta\right)\Sigma_0^{-1}E\left(h_\theta\right)^\top\right]^{-1} E\left(h_\theta\right)\Sigma_0^{-1}\cdot\sqrt{T}\bar{h}\left(y_1^T;\theta_0\right)\\
&\overset{d}{=} \sqrt{T}\bar{h}\left(y_1^T;\theta_0\right) - E\left(h_\theta\right)^\top \left[E\left(h_\theta\right)\Sigma_0^{-1}E\left(h_\theta\right)^\top\right]^{-1} E\left(h_\theta\right)\Sigma_0^{-1}\cdot\sqrt{T}\bar{h}\left(y_1^T;\theta_0\right)\\
&= \underset{q\times q}{\left(\mathbf{I}_q-\mathbf{P}_q\right)}\underset{q\times 1}{\sqrt{T}\bar{h}\left(y_1^T;\theta_0\right)},
\end{aligned}$$

and

$$\mathbf{P}_q \equiv E\left(h_\theta\right)^\top \left[E\left(h_\theta\right)\Sigma_0^{-1}E\left(h_\theta\right)^\top\right]^{-1} E\left(h_\theta\right)\Sigma_0^{-1}$$

is the orthogonal projector in the space generated by the columns of $E\left(h_\theta\right)$ by the inner product Σ_0^{-1}. Thus, we have shown that,

$$\mathcal{C}_T \overset{d}{=} \sqrt{T}\bar{h}\left(y_1^T;\theta_0\right)^\top \left(\mathbf{I}_q-\mathbf{P}_q\right)^\top \Sigma_T^{-1}\left(\mathbf{I}_q-\mathbf{P}_q\right)\sqrt{T}\bar{h}\left(y_1^T;\theta_0\right).$$

But,

$$\sqrt{T}\bar{h}\left(y_1^T;\theta_0\right) \xrightarrow{d} N\left(0,\Sigma_0\right),$$

and, by a classical result,

$$\mathcal{C}_T \xrightarrow{d} \chi^2\left(q-p\right).$$

Hansen and Singleton (1982, 1983) started the literature on the estimation and testing of dynamic asset pricing models within a fully articulated rational expectations framework. Consider the classical system of Euler equations arising in the Lucas tree,

$$E\left[\beta\frac{u'(c_{t+1})}{u'(c_t)}\left(1+r_{i,t+1}\right)-1\middle|\mathcal{F}_t\right]=0,\quad i=1,\cdots,m,$$

where u is the utility function of the representative agent, r_i is the return on asset i, β is the time-discount factor, $\mathcal{F}_t$ is the information set as of time t, and m is the number of assets. Consider the CRRA utility function, $u(x)=x^{1-\eta}/\left(1-\eta\right)$. If the model is well-specified, then, there exist some β_0 and η_0 such that:

$$E\left[\beta_0\left(\frac{c_{t+1}}{c_t}\right)^{-\eta_0}\left(1+r_{i,t+1}\right)-1\middle|\mathcal{F}_t\right]=0,\quad i=1,\cdots,m.$$

To sumup, the dimension of the parameter vector is $p = 2$. To estimate the true parameter vector $\theta_0 \equiv (\beta_0, \eta_0)$, we may build up a system of orthogonality conditions. This system can be based on projecting observable variables predicted by the model onto other variables, some "instruments" included in the information set $\mathcal{F}_t$:

$$E\left[h\left(y_t;\theta_0\right)\right] = 0,$$

where, for some vector of z instruments, say, $\text{In}_t = [i_{1,t}, \cdots, i_{z,t}]^\top$,

$$\underset{q\times 1}{h\left(y_t;\theta\right)} = \begin{pmatrix} \left[\beta\left(\frac{c_{t+1}}{c_t}\right)^{-\eta}\left(1+r_{1,t+1}\right)-1\right]\cdot \text{In}_t \\ \vdots \\ \left[\beta\left(\frac{c_{t+1}}{c_t}\right)^{-\eta}\left(1+r_{m,t+1}\right)-1\right]\cdot \text{In}_t \end{pmatrix}, \quad q = m\cdot z. \tag{6.9}$$

The instruments used to produce the orthogonality restrictions, may include constants, past values of consumption growth, $\frac{c_{t+1}}{c_t}$, or even past returns.

6.7 Simulation-based estimators

Ideally, MLE should be the preferred estimation method of parametric Markov models, as it leads to first-order efficiency. Yet economic theory places restrictions that make these models problematic to estimate through maximum ML. In these cases, GMM is a natural estimation method. But GMM can be unfeasible as well, in situations of interest. Assume, for example, that the data generating process is not i.i.d. Instead, data are generated by the transition function,

$$y_{t+1} = H\left(y_t, \epsilon_{t+1};\theta_0\right), \tag{6.10}$$

where $H : \mathbb{R}^n \times \mathbb{R}^d \times \Theta \mapsto \mathbb{R}^n$, and ϵ_t is a vector of i.i.d. disturbances in $\mathbb{R}^d$. Assume the econometrician knows the function H. Let $z_t = (y_t, y_{t-1}, \cdots, y_{t-l+1})$, $l < \infty$. In many cases of interest, the function $\bar{h}$ in Eq. (6.6) can be written as,

$$\bar{h}\left(y_1^T;\theta\right) = \frac{1}{T}\sum_{t=1}^{T}\underbrace{\left[f_t^* - E\left(f\left(z_t,\theta\right)\right)\right]}_{\equiv h(z_t,\theta)}, \tag{6.11}$$

where,

$$f_t^* = f\left(z_t,\theta_0\right),$$

is a vector-valued moment function, or "observation function," a function that summarizes satisfactorily the data, so to speak. Consider, for example, Eq. (6.9) without the instruments In_t, where $f_t^* = \left(1+r_{i,t+1}\right)^{-1}$ and $E\left(f\left(z_t,\theta\right)\right) = \beta E((\frac{c_{t+1}}{c_t})^{-\eta})$. Once we identify consumption growth with y_{t+1}, $y_{t+1} = \ln\frac{c_{t+1}}{c_t}$, and take the transition law in Eq. (6.10) to be log-normally distributed, as in some basic models we shall see in Part II, we can compute $E\left(f\left(z_t,\theta\right)\right)$ in closed form. Needless to say, the GMM estimator is unfeasible, if we are not able to compute the expectation $E\left(f\left(z_t,\theta\right)\right)$ in closed form, for each θ. Simulation-based methods can make the method of moments feasible in this case.

6.7.1 Three simulation-based estimators

The basic idea underlying simulation-based methods is quite simple. While the moment conditions are too complex to be evaluated analytically, the model in Eq. (6.10) can be simulated. Accordingly, draw ϵ_t from its distribution, and save the simulated values $\hat{\epsilon}_t$. Compute recursively,

$$y_{t+1}^{\theta} = H\left(y_t^{\theta}, \hat{\epsilon}_{t+1}, \theta\right),$$

and create simulated moment functions as follows,

$$f_t^{\theta} \equiv f\left(z_t^{\theta}, \theta\right).$$

Consider the following parameter estimator,

$$\theta_T = \arg\min_{\theta\in\Theta} G_T(\theta)^{\top} W_T G_T(\theta), \tag{6.12}$$

where W_T is some weigthing matrix, $G_T(\theta)$ is the simulated counterpart to $\bar{h}$ in Eq. (6.11),

$$G_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left(f_t^* - \frac{1}{S(T)}\sum_{s=1}^{S(T)} f_s^{\theta}\right),$$

and $S(T)$ is the simulated sample size, which we write as a function of the sample size T, for the purpose of the asymptotic theory.

The estimator θ_T, also known as the *Simulated Method of Moments* (SMM) estimator, aims to match the sample properties of the actual and simulated processes f_t^* and f_t^{θ}. It was introduced in a series of works, by McFadden (1989), Pakes and Pollard (1989), Lee and Ingram (1991) and Duffie and Singleton (1993). The simulated pseudo-maximum likelihood method of Laroque and Salanié (1989, 1993, 1994) can also be interpreted as a SMM estimator.

A second simulation-based estimator relies on the *indirect inference* principle (IIP), and was proposed by Gouriéroux, Monfort and Renault (1993) and Smith (1993). Instead of minimizing the distance of some moment conditions, the IIP relies on minimizing the parameters of an auxiliary, possibly misspecified model. For example, consider the following auxiliary parameter estimator,

$$\beta_T = \arg\max_{\beta} \ln L\left(y_1^T; \beta\right), \tag{6.13}$$

where L is the likelihood of some possibly misspecified model. Consider simulating S times the process y_t in Eq. (6.10), and computing,

$$\beta_T^s(\theta) = \arg\max_{\beta} \ln L(y_s(\theta)_1^T; \beta), \quad s = 1, \cdots, S,$$

where $y_s(\theta)_1^T = (y_t^{\theta,s})_{t=1}^T$ are the simulated variables (for $s = 1, \cdots, S$) when the parameter vector is θ. The IIP-based estimator is defined similarly as θ_T in Eq. (6.12), but with the function G_T given by,

$$G_T(\theta) = \beta_T - \frac{1}{S}\sum_{s=1}^{S}\beta_T^s(\theta). \tag{6.14}$$

The diagram in Figure 6.1 illustrates the main ideas underlying the IIP.

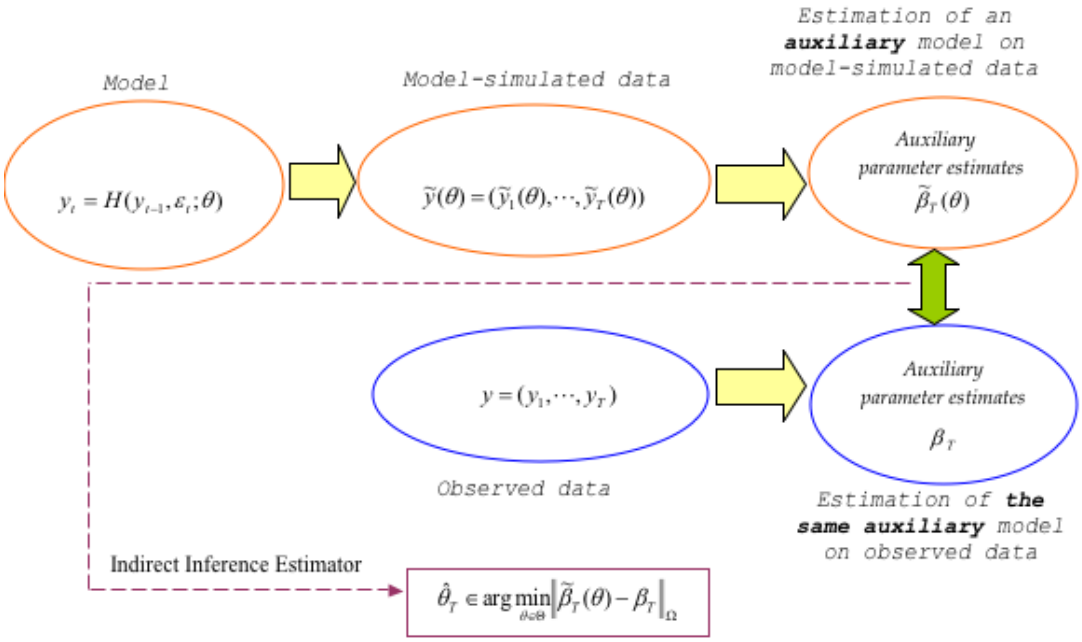


FIGURE 6.1. *The Indirect Inference principle.* Given the *true* model $y_t = H(y_{t-1}, \epsilon_t; \theta)$, an estimator of θ based on the indirect inference principle ($\hat{\theta}_T$ say) makes the parameters of some *auxiliary* model $\tilde{\beta}_T(\hat{\theta}_T)$ as close as possible to the parameters β_T of *the same auxiliary* model estimated on the observations. That is, $\hat{\theta}_T = \arg\min_{\theta \in \Theta} \left\| \tilde{\beta}_T(\theta) - \beta_T \right\|_\Omega$, for some norm Ω.

Finally, Gallant and Tauchen (1996) propose a simulation-based estimation method they label *efficient method of moments* (EMM). Their estimator sets,

$$G_T(\theta, \beta_T) = \frac{1}{N} \sum_{n=1}^{N} \frac{\partial}{\partial \beta} \ln f\left(y_n^\theta \middle| z_{n-1}^\theta; \beta_T\right),$$

where $\frac{\partial}{\partial \beta} \ln f(y| z; \beta)$ is the score of some auxiliary model f, also known as the score generator, β_T is the Pseudo ML estimator of the auxiliary model, and $(y_n^\theta)_{n=1}^N$ is a long simulation (i.e. N is very large) of Eq. (6.10), with parameter vector set equal to θ. Finally, the weighting matrix W_T in Eq. (6.12) is taken to be any matrix $\mathcal{I}_T^{-1}$ converging in probability to:

$$\mathcal{I} = E\left[\left| \frac{\partial}{\partial \beta} \ln f(y_2| z_1; \beta) \right|_2 \right]. \tag{6.15}$$

To motivate this choice of $G_T(\theta)$, note that the auxiliary score, $\frac{\partial}{\partial \beta} \ln f(y_t| z_{t-1}; \beta_T)$, satisfies the following first order conditions:

$$\frac{1}{T} \sum_{n=1}^{T} \frac{\partial}{\partial \beta} \ln f(y_t| z_{t-1}; \beta_T) = 0,$$

which is the sample equivalent of

$$E\left[\frac{\partial}{\partial\beta}\ln f\left(y_2|\,z_1;\beta^*\right)\right]=0,$$

for some β^*. Likewise, we must have that with $\theta=\theta_0$, $G_T(\theta_0,\beta_T)=0$, *for large* N. All in all, we want to find a stochastic process $H\left(y_t,\epsilon_{t+1};\cdot\right)$ in Eq. (6.10), or a parameter vector θ such that the expectation of the score of the auxiliary model is zero, a very property of the score, arising even when the model is misspecified.

6.7.2 Asymptotic normality

We show, heuristically, how asymptotic normality obtains for the three estimators of Section 6.7.1, and then, define conditions under which asymptotic efficiency might obtain for the EMM.

6.7.2.1 SMM

Let,

$$\Sigma_0=\sum_{j=-\infty}^{\infty}E\left[\left(f_t^*-E\left(f_t^*\right)\right)\left(f_{t-j}^*-E\left(f_{t-j}^*\right)\right)^\top\right],$$

and suppose that

$$W_T\xrightarrow{p}W_0=\Sigma_0^{-1}.$$

We now demonstrate that under this condition, as $T\to\infty$ and $S\left(T\right)\to\infty$,

$$\sqrt{T}(\theta_T-\theta_0)\xrightarrow{d}N\left(\mathbf{0}_p,\left(1+\tau\right)\left(D_0^\top\Sigma_0^{-1}D_0\right)^{-1}\right),\tag{6.16}$$

where $\tau=\lim_{T\to\infty}\frac{T}{S(T)}$, $D_0=E\left(\nabla_\theta G_\infty\left(\theta_0\right)\right)=E\left(\nabla_\theta f_\infty^{\theta_0}\right)$, and the notation G_∞ means that $G_{\cdot}$ is drawn from its stationary distribution.

Indeed, the first order conditions satisfied by the SMM in Eq. (6.12) are,

$$\mathbf{0}_p=\left[\nabla_\theta G_T\left(\theta_T\right)\right]^\top W_T G_T\left(\theta_T\right)=\left[\nabla_\theta G_T\left(\theta_T\right)\right]^\top W_T\cdot\left[G_T\left(\theta_0\right)+\nabla_\theta G_T\left(\theta_0\right)\left(\theta_T-\theta_0\right)\right]+o_p\left(1\right).$$

That is,

$$\begin{aligned}\sqrt{T}\left(\theta_T-\theta_0\right)&\overset{d}{=}-\left(\left[\nabla_\theta G_T\left(\theta_T\right)\right]^\top W_T\nabla_\theta G_T\left(\theta_0\right)\right)^{-1}\left[\nabla_\theta G_T\left(\theta_T\right)\right]^\top W_T\cdot\sqrt{T}G_T\left(\theta_0\right)\\&\overset{d}{=}-\left(D_0^\top W_0 D_0\right)^{-1}D_0^\top W_0\cdot\sqrt{T}G_T\left(\theta_T\right)\\&=-\left(D_0^\top\Sigma_0^{-1}D_0\right)^{-1}D_0^\top\Sigma_0^{-1}\cdot\sqrt{T}G_T\left(\theta_0\right).\end{aligned}\tag{6.17}$$

We have,

$$\begin{aligned}\sqrt{T}G_T\left(\theta_0\right)&=\sqrt{T}\cdot\frac{1}{T}\sum_{t=1}^{T}\left(f_t^*-\frac{1}{S\left(T\right)}\sum_{s=1}^{S(T)}f_s^{\theta_0}\right)\\&=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(f_t^*-E\left(f_\infty^*\right)\right)-\frac{\sqrt{T}}{\sqrt{S\left(T\right)}}\cdot\frac{1}{\sqrt{S\left(T\right)}}\sum_{s=1}^{S(T)}\left(f_s^{\theta_0}-E\left(f_\infty^{\theta_0}\right)\right)\\&\xrightarrow{d}N\left(0,\left(1+\tau\right)\Sigma_0\right),\end{aligned}$$

where we have used the fact that $E\left(f_{\infty}^{*}\right)=E\left(f_{\infty}^{\theta_0}\right)$. By using this result into Eq. (6.17) produces the convergence in Eq. (6.16). If $\tau=\lim_{T\to\infty}\frac{T}{S(T)}=0$ (i.e. if the number of simulations grows faster than the sample size), the SMM estimator is as efficient as the GMM estimator. Finally, and obviously, we need that $\tau=\lim_{T\to\infty}\frac{T}{S(T)}<\infty$: the number of simulations $S\left(T\right)$ cannot grow more slowly than the sample size.

6.7.2.2 Indirect inference

The IIP-based estimator works slightly differently. For this estimator, even if the number of simulations S is fixed, asymptotic normality obtains without requiring S to go to infinity faster than the sample size. Basically, what really matters here is that ST goes to infinity.

By Eq. (6.17), and the discussion in Section 6.7.1, we know that asymptotically, the first order conditions satisfied by the IIP-based estimator are,

$$\sqrt{T}\left(\theta_T-\theta_0\right)\overset{d}{=}-\left(D_0^{\top}W_0D_0\right)^{-1}D_0^{\top}W_0\cdot\sqrt{T}G_T\left(\theta_0\right),$$

where G_T is as in Eq. (6.14), $D_0=\nabla_{\theta}b\left(\theta\right)$, and $b\left(\theta\right)$ is solution to the limiting problem corresponding to the estimator in Eq. (6.13), viz

$$\beta\left(\theta\right)=\arg\max_{\beta}\left(\lim_{T\to\infty}\frac{1}{T}\ln L\left(y_1^T;\beta\right)\right).$$

We need to find the distribution of G_T in Eq. (6.14). We have,

$$\begin{aligned}\sqrt{T}G_T\left(\theta_0\right)&=\frac{1}{S}\sum_{s=1}^{S}\sqrt{T}\left(\beta_T-\beta_T^s\left(\theta_0\right)\right)\\&=\frac{1}{S}\sum_{s=1}^{S}\sqrt{T}\left[\left(\beta_T-\beta_0\right)-\left(\beta_T^s\left(\theta_0\right)-\beta_0\right)\right]\\&=\sqrt{T}\left(\beta_T-\beta_0\right)-\frac{1}{S}\sum_{s=1}^{S}\sqrt{T}\left(\beta_T^s\left(\theta_0\right)-\beta_0\right),\end{aligned}$$

where $\beta_0=\beta\left(\theta_0\right)$. Hence, given the independence of the sample and the simulations,

$$\sqrt{T}G_T\left(\theta_0\right)\xrightarrow{d}N\left(0,\left(1+\frac{1}{S}\right)\cdot\text{Asy.Var}\left(\sqrt{T}\beta_T\right)\right).$$

That is, asymptotically S can be fixed with respect to T.

6.7.2.3 Efficient method of moments

We have,

$$\theta_T=\arg\min_{\theta}G_T\left(\theta,\beta_T\right)^{\top}W_TG_T\left(\theta,\beta_T\right),\quad G_T\left(\theta,\beta_T\right)=\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\beta}\ln f\left(y_n^{\theta}\middle|\, z_{n-1}^{\theta};\beta_T\right).$$

The first order conditions are:

$$\begin{aligned}0&=\nabla_{\theta}G_T\left(\theta_T,\beta_T\right)^{\top}W_TG_T\left(\theta_T,\beta_T\right)\\&\overset{d}{=}\nabla_{\theta}G_T\left(\theta_0,\beta_T\right)^{\top}W_T\left(G_T\left(\theta_0,\beta_T\right)+\nabla_{\theta}G_T\left(\theta_0,\beta_T\right)\left(\theta_T-\theta_0\right)\right),\end{aligned}$$

or

$$\sqrt{T}\left(\theta_T-\theta_0\right)\overset{d}{=}-\left(\nabla_\theta G_T\left(\theta_0,\beta_T\right)^\top W_T\nabla_\theta G_T\left(\theta_0,\beta_T\right)\right)^{-1}\nabla_\theta G_T\left(\theta_0,\beta_T\right)^\top W_T\sqrt{T}G_T\left(\theta_0,\beta_T\right).$$

We have, for some β^*,

$$\sqrt{T}G_T\left(\theta_0,\beta_T\right)\overset{d}{=}\mathcal{J}\sqrt{T}\left(\beta_T-\beta^*\right)\overset{d}{\to}N\left(0,\mathcal{I}\right),$$

where $\mathcal{J}=E\left(\frac{\partial}{\partial\beta\partial\beta^\intercal}\ln f\left(y_2|\,y_1;\beta\right)\right)$ and $\mathcal{I}$ is as in Eq. (6.15). Hence,

$$\sqrt{T}\left(\theta_T-\theta_0\right)\overset{d}{\to}N\left(0,V\right),$$

where,

$$V=\left(\nabla_\theta G^\top W\nabla_\theta G\right)^{-1}\nabla_\theta G^\top W\mathcal{I}W^\top\nabla_\theta G\left(\nabla_\theta G^\top W\nabla_\theta G\right)^{-1}.$$

With $W=\mathcal{I}^{-1}$, this variance collapses to,

$$V=\left(\nabla_\theta G^\top\mathcal{I}^{-1}\nabla_\theta G\right)^{-1}. \tag{6.18}$$

6.7.2.4 Spanning scores

This section provides a heuristic discussion of the conditions under which the EMM achieves the Cramer-Rao lower bound. Consider the following definition, which is similar to that in Tauchen (1997). Of a given span of moment conditions s_f, say that of the EMM, we say that it also spans the true score if,

$$var\left(s|\,s_f\right)=0, \tag{6.19}$$

where s denotes the true score. From Eq. (6.18), we know that the asymptotic variance of the EMM, say var_{EMM}, satisfies:

$$var_{\text{EMM}}^{-1}\equiv V^{-1}=\nabla_\theta G^\top var\left(s_f\right)^{-1}\nabla_\theta G.$$

By the linear projection,

$$s=Bs_f+\epsilon,\quad B=cov\left(s,s_f\right)var\left(s_f\right)^{-1},$$

we have,

$$var_{\text{MLE}}^{-1}=var\left(s\right)=Bvar\left(s_f\right)B^\top+var\left(s|\,s_f\right)=cov\left(s,s_f\right)var\left(s_f\right)^{-1}cov\left(s,s_f\right)^\top+var\left(s|\,s_f\right), \tag{6.20}$$

where var_{MLE} denotes the asymptotic variance of the MLE. We claim that:

$$cov\left(s,s_f\right)^\top=\nabla_\theta G. \tag{6.21}$$

Indeed, under regularity conditions,

$$\begin{aligned}\nabla_\theta G\left(\theta_0,\beta^*\right)&=\left[\frac{\partial}{\partial\theta}\left(\int\frac{\partial}{\partial\beta}\ln f\left(y;\beta^*\right)p\left(y,\theta\right)dy\right)\right]_{\theta=\theta_0}\\&=\int\frac{\partial}{\partial\beta}\ln f\left(y;\beta^*\right)\frac{\partial}{\partial\theta}p\left(y,\theta_0\right)dy\\&=\int\left(\frac{\partial}{\partial\beta}\ln f\left(y;\beta^*\right)\frac{\partial}{\partial\theta}\ln p\left(y,\theta_0\right)\right)p\left(y,\theta_0\right)dy\\&=cov\left(s,s_f\right)^\top,\end{aligned}$$

where $p(y, \theta_0)$ is the true density. Next, replace Eq. (6.21) into Eq. (6.20),

$$var_{\text{MLE}}^{-1} = \nabla_\theta G^\top var(s_f)^{-1} \nabla_\theta G + var(s|\, s_f) = var_{\text{EMM}}^{-1} + var(s|\, s_f).$$

Therefore, the EMM estimator achieves the Cramer-Rao lower bound under the spanning condition in Eq. (6.19).

6.7.3 A fourth simulation-based estimator: Simulated maximum likelihood

Estimating the parameters of stochastic differential equations is a recurrent theme in empirical finance. Consider a continuous time model,

$$dy(\tau) = b(y(\tau); \theta)\, d\tau + \Sigma(y(\tau); \theta)\, dW(\tau), \tag{6.22}$$

where $W(\tau)$ is a Brownian motion and b and Σ are two functions guaranteeing a strong solution to Eq. (6.22). Except in special cases (e.g., the affine models reviewed in Chapters 11 and 13), the likelihood function of the data generated by this process is unknown. We can then use one of the three estimators we have presented in section 6.7.1. Alternatively, we might use simulated maximum likelihood, a method introduced in finance by Santa-Clara (1995) (see, also, Brandt and Santa-Clara, 2002). We only provide the idea of the method, not the asymptotic theory.

Suppose, then, that we observe discretely sample data generated by Eq. (6.22): y_0, y_1,$\cdots$, y_t, $\cdots$, y_T, where T is the sample size. We need to know the transition density, say $p(y_{t+1}|\, y_t; \theta)$, to implement maximum likelihood, which we assume we do not know. Consider, then, the Euler approximation to Eq. (6.22),

$$y_{(k+1)/n} = y_{k/n} + b\left(y_{k/n}; \theta\right) \frac{1}{n} + \Sigma\left(y_{k/n}; \theta\right) \sqrt{\frac{1}{n}} \epsilon_{k+1}, \tag{6.23}$$

where ϵ_k is a sequence of i.i.d. random variables with expectation zero and unit variance. This stochastic process is defined at the dates $\frac{k}{n}$, for k integer. Let $[Tn]$ denote the integer part of Tn, and for $k = 1, \cdots, [Tn]$, set

$$\hat{y}_\tau^{(n)} = y_{k/n}, \quad \text{if } \frac{k}{n} \leq \tau \leq \frac{k+1}{n}.$$

In other words, we are chopping the time interval between two observations, $[t, t+1]$, in n pieces, and then take n to be large. We know that as $n \to \infty$, $\hat{y}_t^{(n)} \Rightarrow y(t)$ as $n \to \infty$, where $\Rightarrow$ denotes "weak convergence," or "convergence in distribution," meaning that all finite dimensional distributions of $\hat{y}_t^{(n)}$ converge to those of $y(t)$ as $n \to \infty$. The idea underlying simulated maximum likelihood, then, is to estimate the transition density, $p(y_{t+1}|\, y_t; \theta)$, through simulations of Eq. (6.23), performed using a large value of n. Note, we cannot guarantee the transition density is recovered by simulating Eq. (6.23), not even for a large value of n. We can only perform an imperfect simulation of Eq. (6.23).

The likelihood function is,

$$L = p(y_0; \theta) \prod_{t=0}^{T-1} p(y_{t+1}|\, y_t; \theta),$$

where $p(y_0; \theta)$ denotes the marginal density of the first observation, y_0.

Let $p^n\left(y'|\,y;\theta\right)$ the transition density of the data generated by Eq. (6.23). Then, if ϵ is normally distributed,

$$p^n\left(y_{(k+1)/n}\big|\,y_{k/n};\theta\right)=\varphi\left(y_{(k+1)/n};y_{k/n}+b\left(y_{k/n};\theta\right)\frac{1}{n};\Sigma^2\left(y_{k/n};\theta\right)\frac{1}{n}\right),$$

where $\varphi\left(u;\mu;\sigma^2\right)$ denotes the Gaussian density with mean μ and variance σ^2. Moreover, we have, approximately,

$$\begin{aligned}p^n\left(y_{t+1}|\,y_t;\theta\right)&=\int p^n\left(y_{t+1}|\,x;\theta\right)p^n\left(x|\,y_t;\theta\right)dx\\&=\int\varphi\left(y_{t+1};x+b\left(x;\theta\right)\frac{1}{n};\Sigma^2\left(x;\theta\right)\frac{1}{n}\right)p^n\left(x|\,y_t;\theta\right)dx,\end{aligned}$$

where we have set $x=y_{t+1-\frac{1}{n}}$. We may, now, draw values of x from $p^n\left(x|\,y_t;\theta\right)$, as explained in a moment, and estimate $p^n\left(y_{t+1}|\,y_t;\theta\right)$ through:

$$p^{n,S}\left(y_{t+1}|\,y_t;\theta\right)\equiv\frac{1}{S}\sum_{j=1}^{S}\varphi\left(y_{(k+1)/n};\tilde{x}^j+b\left(\tilde{x}^j;\theta\right)\frac{1}{n};\Sigma^2\left(\tilde{x}^j;\theta\right)\frac{1}{n}\right),$$

where $\tilde{x}^j$ is obtained by iterating Eq. (6.23) from time t to time $t+1-\frac{1}{n}$. Under regularity conditions, we have that for all $\theta\in\Theta$, $\sup_{y',y}\left|p^{n,S}\left(y'|\,y;\theta\right)-p\left(y'|\,y;\theta\right)\right|\to 0$ as n and S get large, with $\frac{\sqrt{S}}{n}\to 0$.

6.7.4 Advances

The three estimators examined in Sections 6.7.1 and 6.7.2 are general-purpose, but in general, they do not lead to to asymptotic efficiency, unless the true score belongs to the span of the moment conditions, as explained in Section 6.7.2.4. There exist other simulation-based methods, which aim to approximate the likelihood function through simulations (e.g., Lee, 1995; Hajivassiliou and McFadden, 1998): for example, the simulated maximum likelihood estimator in Section 6.7.2.3 can be used to estimate the parameters of stochastic differential equations. While methods based on simulated likelihood lead to asymptotically efficient estimators, they address specific estimation problems, just as the example of Section 6.7.2.3 illustrates.

There exist estimators that are both general purpose and that can lead to asymptotic efficiency. Fermanian and Salanié (2004) consider an estimator that relies on approximating the likelihood function through kernel estimates obtained simulating the model of interest. Carrasco, Chernov, Florens and Ghysels (2007) rely on a "continuum of moment conditions" matching model-based (simulated) characteristic functions to data-based characteristic functions. Altissimo and Mele (2009) propose an estimator based on a continuum of moment conditions, which minimizes a certain distance between conditional densities estimated with the true data and conditional densities estimated with data simulated from the model, where both conditional densities are estimated through kernel methods.

6.7.5 In practice? Latent factors and identification

The estimation theory of this section does not rule out the situation where some of the variables in Eq. (6.10) are unobservable. The principle to follow is very simple, one applies any of the

methods we have discussed to those variables simulated out of Eq. (6.10), which correspond to the observed ones. For example, we may want to estimate the following model of the short-term rate $r(\tau)$, discussed at length in Chapter 13:

$$\begin{array}{l} dr(\tau) = \kappa_r(\bar{r} - r(\tau))\, d\tau + \sqrt{v(\tau)} dW_1(\tau) \\ dv(\tau) = \kappa_v(\bar{v} - v(\tau))\, d\tau + \xi\sqrt{v(\tau)} dW_2(\tau) \end{array} \tag{6.24}$$

where $v(\tau)$ is the short-term rate instantaneous, stochastic variance, W_1 and W_2 are two standard Brownian motions, and the parameter vector of interest is $\theta = [\kappa_r\ \bar{r}\ \kappa_v\ \bar{v}\ \xi]$. Let us consider one of the methods discussed so far, say indirect inference. The logical steps to follow, then, are (i) to simulate Eqs. (6.24), and (ii) to calibrate an auxiliary model to the short term rate data simulated out of Eqs. (6.24) which is as close as possible to the very same auxiliary model fitted on true data. Note, in doing so, we just have to neglect the volatility data simulated out of Eqs. (6.24), as these data are obviously unobservable.

The question arises, therefore, as to whether the auxiliary model one chooses is rich enough to allow identifying the model's parameter vector θ. There might be many combinations of unobserved random processes $v(\tau)$ that are consistent with the likelihood of any given auxiliary model. So which auxiliary model to fit, in practice? Gallant and Tauchen (1996) asked this question long time ago. Needless to mention, there are no general answers to this question. Very simply, one requires the model to be identifiable, which is likely to happen once the auxiliary model is "rich enough." In an impressive series of applied work, Gallant and Tauchen and their co-authors have proposed semi-nonparametric score generators, as a way to get as close as possible to a "rich" model. Intuitively, by increasing the order of Hermite expansions, semi-nonparametric scores might converge to the true ones. Alternatively, one might use a continuum of moment conditions, as explained in Section 6.7.4. For example, the nonparametric density estimators of Altissimo and Mele (2009) converge to the true parameter once the bandwidth parameters used to smooth out these kernel estimates gets smaller and smaller. In the next section, we provide a discussion of how asset prices might help convey information about unobserved processes and lead to statistical efficiency.

6.8 Asset pricing, prediction functions, and statistical inference

We develop conditions, which ensure the feasibility of estimation methods in a context where an unobservable multidimensional process is estimated in conjunction with prediction functions suggested by asset pricing models.[4] We assume that the data generating process is a multidimensional partially observed diffusion process solution to,

$$dy(\tau) = b(y(\tau);\theta)\, d\tau + \Sigma(y(\tau);\theta)\, dW(\tau), \tag{6.25}$$

where W is a multidimensional process and (b, Σ) satisfy some regularity conditions we single out below. We analyze situations where the original partially observed system in Eq. (6.25) can be estimated by augmenting it with a number of observable deterministic functions of the state. In many situations, such deterministic functions are suggested by asset pricing theories in a natural way. Typical examples include the price of derivatives or in general, any functional of asset prices (such as asset returns, bond yields, implied volatilities).

[4]This section is based on an unpublished appendix of Altissimo and Mele (2009).

The idea to use asset pricing predictions to improve the fit of models with unobservable factors has been explored at least by, e.g., Christensen (1992), Pastorello, Renault and Touzi (2000), Chernov and Ghysels (2000), Singleton (2001), and Pastorello, Patilea and Renault (2003).

We consider a standard Markov pricing setting. For fixed $t \geq 0$, we let M be the expiration date of a contingent claim with rational price process $c = \{c(y(\tau), M-\tau)\}_{\tau\in[t,M)}$, and let $\{z(y(\tau))\}_{\tau\in[t,M]}$ and $\Pi(y)$ be the associated intermediate payoff process and final payoff function, respectively. Let $\partial/\partial\tau + L$ be the usual infinitesimal generator of the system in Eq. (6.25), taken under the risk-neutral probability. Then, as we know from Chapter 5, in a frictionless market, the arbitrage-free price c is the solution to the following partial differential equation

$$\begin{cases} 0 = \left(\dfrac{\partial}{\partial\tau} + L - R\right) c(y, M-\tau) + z(y), \ \forall (y,\tau) \in Y \times [t, M) \\ c(y, 0) = \Pi(y), \ \forall y \in Y \end{cases} \tag{6.26}$$

where $R \equiv R(y)$ is the short-term rate. We call *prediction function* any continuous and twice differentiable function $c\,(y; M-\tau)$ solution to the partial differential equation and boundary condition in (6.26). Examples of contingent claims with prices satisfying (6.26) are derivatives, typically.

Next, we augment the system in Eq. (6.25) with $d-q$ prediction functions, where q denotes the number of the observable variables in Eq. (6.25). Precisely, we let:

$$C(\tau) \equiv \left(c\left(y(\tau), M_1 - \tau\right), \cdots, c\left(y(\tau), M_{d-q} - \tau\right)\right), \quad \tau \in [t, M_1]$$

where $\{M_i\}_{i=1}^{d-q}$ is an increasing sequence of fixed maturity dates. Furthermore, we define the measurable vector valued function:

$$\phi\left(y(\tau); \theta, \gamma\right) \equiv \left(y^o(\tau), C\left(y(\tau)\right)\right), \quad \tau \in [t, M_1], \quad (\theta, \gamma) \in \Theta \times \Gamma, \tag{6.27}$$

where $y^o(\tau)$ denotes the vector of observable variables in Eq. (6.25), and $\Gamma \subset \mathbb{R}^{p_\gamma}$ is a compact parameter set containing additional parameters. These new parameters arise from the change of measure leading to the pricing model in Eq. (6.27), and are now part of our estimation problem.

We assume that the pricing model in Eq. (6.27) is correctly specified. That is, all contingent claim prices in the economy are taken to be generated by the prediction function $c(y, M-\tau)$ for some $(\theta_0, \gamma_0) \in \Theta \times \Gamma$. For simplicity, we also consider a stylized situation in which all contingent claims have the same contractual characteristics specified by $\mathcal{C} \equiv (z, \Pi)$. More generally, one may define a series of classes of contingent claims $\{\mathcal{C}_j\}_{j=1}^J$, where the class of contingent claims j has characteristics specified by $\mathcal{C}_j \equiv (z_j, \Pi_j)$. As an example, assets belonging to the class $\mathcal{C}_1$ can be European options, assets belonging to the class $\mathcal{C}_1$ can be bonds. The number of prediction functions that we would introduce in this case would be equal to $d - q = \sum_{j=1}^J M^j$, where M^j is the number of prediction functions within class of assets j. To keep the presentation simple, we do not consider such a more general situation.

The objective is to define estimators of the parameter vector (θ_0, γ_0), under which observations were generated. We want to use any of the simulation methods reviewed in Section 6.7 to produce an estimator of (θ_0, γ_0). The idea, as usual, is to make the finite dimensional distributions of ϕ implied by the pricing model in Eq. (6.27) and the fundametals in Eq. (6.26) as close as possible to the sample counterparts of ϕ. Let $\Phi \subseteq \mathbb{R}^d$ be the domain on which ϕ takes values. As illustrated by Figure 6.2, we want to move from the unfeasible domain Y of

the original state variables in Eq. (6.25) (observables and not) to the domain Φ on which only observable variables take value. Ideally, we would like to implement such a change in domain in order to recover as much information as possible about the original unobserved process in (6.25). Clearly, ϕ is fully revealing whenever it is globally invertible. However, we will show that estimation is feasible even when ϕ is only locally one-to-one.

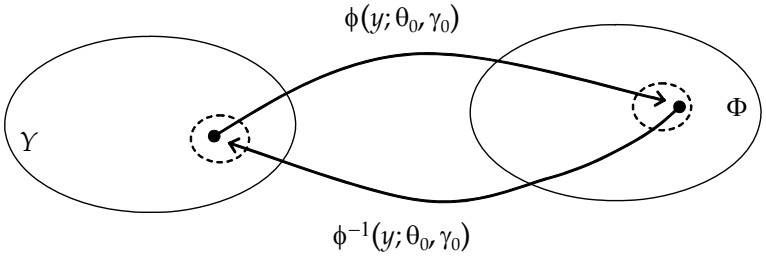


FIGURE 6.2. *Asset pricing, the Markov property, and statistical efficiency.* Y is the domain on which the partially observed primitive state process $y \equiv (y^o \quad y^u)^\top$ takes values, Φ is the domain on which the observed system $\phi \equiv (y^o \; C(y))^\top$ takes values in Markovian economies, and $C(y)$ is a contingent claim price process in $\mathbb{R}^{d-q^*}$. Let $\phi^c = (y^o, c(y,\ell_1), \cdots, c(y,\ell_{d-q^*}))$, where $\{c(y,\ell_j)\}_{j=1}^{d-q^*}$ forms an intertemporal cohort of contingent claim prices, as in Definition 5.3. If the *local restrictions* of ϕ are one-to-one and onto, statistical inference about θ and γ can be made, using information about the price of derivative contracts, ϕ^c. If ϕ is also *globally* invertible, statistical inference can lead to first-order asymptotic efficiency, once conditioned upon ϕ^c.

An important feature of the theory in this section is that it does not hinge upon the availability of contingent prices data covering the same sample period covered by the observables in Eq. (6.25). First, the price of a given contingent claim is typically not available for a long sample period. As an example, available option data often include option prices with a life span smaller than the usual sample span of the underlying asset prices. By contrast, it is common to observe long time series of option prices having the same maturity. Second, the price of a single contingent claim depends on the time-to-maturity of the claim; therefore, it does not satisfy the stationarity assumptions maintained in this paper. To address these issues, we deal with data on assets having the same characteristics at each point in time. Precisely, consider the data generated by the following random processes:

DEFINITION 6.3. (Intertertemporal (ℓ, N)-cohort of contingent claim prices) *Given a prediction function $c(y; M-\tau)$ and a N-dimensional vector $\ell \equiv (\ell_1, \cdots, \ell_N)$ of fixed time-to-maturity, an* intertemporal (ℓ, N)-cohort of contingent claim prices *is any collection of contingent claim price processes $c(\tau, \ell) \equiv (c(y(\tau), \ell_1), \cdots, c(y(\tau), \ell_N))$ $(\tau \geq 0)$ generated by the pricing model (6.27).*

Consider for example a sample realization of three-months at-the-money option prices, or a sample realization of six-months zero-coupon bond prices. Long sequences such as the ones in these examples are common to observe. If these sequences were generated by the pricing model in Eq. (6.27), as in Definition 6.3, they would be deterministic functions of y, and hence

stationary. We now develop conditions ensuring both feasibility and first-order efficiency of the class of simulation-based estimators, as applied to this kind of data. Let $\bar{a}$ denote the matrix having the first q rows of Σ, the diffusion matrix in Eq. (6.25). Let ∇C denote the Jacobian of C with respect to y. We have:

THEOREM 6.4. (Asset pricing and Cramer-Rao lower bound) *Suppose to observe an intertemporal $(\ell, d-q)$-cohort of contingent claim prices $c(\tau, \ell)$, and that there exist prediction functions C in $\mathbb{R}^{d-q}$ with the property that for $\theta = \theta_0$ and $\gamma = \gamma_0$,*

$$\begin{pmatrix} \bar{a}(\tau) \cdot \Sigma(\tau)^{-1} \\ \nabla C(\tau) \end{pmatrix} \neq 0, \quad P \otimes d\tau\text{-a.s.} \quad \text{all } \tau \in [t, t+1], \tag{6.28}$$

where C satisfies the initial condition $C(t) = c(t, \ell) \equiv (c(y(t), \ell_1), \cdots, c(y(t), \ell_{d-q}))$. Let $\phi_t^c = (y^o(t), c(y(t), \ell_1), \cdots, c(y(t), \ell_{d-q}))$. Then, any simulation-based estimator applied to ϕ_t^c is feasible. Moreover, asssume ϕ_t^c is also Markov. Then, any estimator with a span of moment conditions for ϕ_t^c that also spans the true score, attains the Cramer-Rao lower bound, with respect to the fields generated by ϕ_t^c.

According to Theorem 6.4, any estimator is feasible, whenever ϕ is locally invertible for a time span equal to the sampling interval. Figure 6.2 illustrates that condition (6.28) is satisfied whenever ϕ is locally one-to-one and onto.[5] If ϕ is also globally invertible for the same time span, ϕ^c is Markov. The last part of this theorem says that in this case, any estimator is asymptotically efficient. We emphasize that this conclusion is about first-order efficiency in the *joint* estimation of θ *and* γ given the observations on ϕ^c.

Naturally, condition (6.28) does not ensure that ϕ is globally one-to-one and onto: ϕ might have many locally invertible restrictions.[6] In practice, ϕ might fail being globally invertible because monotonicity properties of ϕ may break down in multidimensional diffusion models. For example, in models with stochastic volatility, option prices can be decreasing in the underlying asset price (see Bergman, Grundy and Wiener, 1996). In models of the yield curve with stochastic volatility, to cite a second example, medium-long term bond prices can be increasing in the short-term rate (see Mele, 2003). These cases might arise as there is no guarantee that the solution to a stochastic differential system is nondecreasing in the initial condition of one if its components, which is, instead, always true in the scalar case.

When all components of vector y^o represent the prices of assets actively traded in frictionless markets, (6.28) corresponds to a condition ensuring market completeness in the sense of Harrison and Pliska (1983). As an example, condition (6.28) for Heston's (1993) model is $\partial c / \partial \sigma \neq 0$ $P \otimes d\tau$-a.s, where σ denotes instantaneous volatility of the price process. This condition is satisfied by the Heston's model. In fact, Chapter 11 reviews results by which, in fairly general models with stochastic volatility, option prices are *always* strictly increasing in σ whenever they are convex in Q. Theorem 6.4 can be used to implement efficient estimators in other complex multidimensional models. Consider for example a three-factor model of the yield curve. Consider a state-vector (r, σ, ℓ), where r is the short-term rate and σ, ℓ are additional factors (such as, say, instantaneous short-term rate volatility and a central tendency factor). Let $u^{(i)} = u(r(\tau), \sigma(\tau), \ell(\tau); M_i - \tau)$

[5] Local invertibility of ϕ means that for every $y \in Y$, there exists an open set Y_* containing y such that the restriction of ϕ to Y_* is invertible. Let $J\phi$ denote the Jacobian of ϕ. Then, we have that ϕ is locally invertible on Y_* if $\det J\phi \neq 0$ on Y_*, which is condition (6.28).

[6] As an example, consider the mapping $\mathbb{R}^2 \mapsto \mathbb{R}^2$ defined as $\phi(y_1, y_2) = (e^{y_1} \cos y_2, e^{y_1} \sin y_2)$. The Jacobian satisfies $\det J\phi(y_1, y_2) = e^{2y_1}$, yet ϕ is 2π-periodic with respect to y_2. For example, $\phi(0, 2\pi) = \phi(0, 0)$.

be the time τ rational price of a pure discount bond expiring at $M_i \geq \tau$, $i = 1, 2$, and take $M_1 < M_2$. Let $\phi \equiv (r, u^{(1)}, u^{(2)})$. Condition (6.28) for this model is then,

$$u_\sigma^{(1)} u_\ell^{(2)} - u_\ell^{(1)} u_\sigma^{(2)} \neq 0, \;\; P \otimes dt\text{-a.s.} \;\; \tau \in [t, t+1], \tag{6.29}$$

where subscripts denote partial derivatives. It is easily checked that this same condition must be satisfied by models with correlated Brownian motions and by yet more general models. Classes of models of the short-term rate for which condition (6.29) holds are more intricate to identify than in the European option pricing case seen above (see Mele, 2003).

6.9 Appendix 1: Proof of selected results

Proof of Eq. (6.2). We have: $P\left(A_1 \bigcap A_2\right) = P\left(A_1\right) \cdot P\left(A_2 \left| A_1\right.\right)$. Consider the event $E \equiv A_1 \bigcap A_2$. We still have,

$$\Pr\left(A_3 \left| A_1 \bigcap A_2\right.\right) = \Pr\left(A_3 \left| E\right.\right) = \frac{\Pr\left(A_3 \bigcap E\right)}{\Pr\left(E\right)} = \frac{\Pr\left(A_3 \bigcap A_1 \bigcap A_2\right)}{\Pr\left(A_1 \bigcap A_2\right)}.$$

That is,

$$\Pr\left(\bigcap_{i=1}^{3} A_i\right) = \Pr\left(A_1 \bigcap A_2\right) \cdot \Pr\left(A_3 \left| A_1 \bigcap A_2\right.\right) = \Pr\left(A_1\right) \cdot \Pr\left(A_2 \left| A_1\right.\right) \cdot \Pr\left(A_3 \left| A_1 \bigcap A_2\right.\right).$$

Continuing, we obtain Eq. (6.2). ‖

6.10 Appendix 2: Collected notions and results

CONVERGENCE IN PROBABILITY. *A sequence of random vectors* $\{x_T\}$ converges in probability *to the random vector* $\tilde{x}$ *if for each* $\epsilon > 0$, $\delta > 0$ *and each* $i = 1, 2, \cdots, N$, *there exists a* $T_{\epsilon,\delta}$ *such that for every* $T \geq T_{\epsilon,\delta}$,

$$\Pr\left(|x_{Ti} - \tilde{x}_i| > \delta\right) < \epsilon.$$

This is succinctly written as $x_T \xrightarrow{p} \tilde{x}$, *or* $\operatorname{plim} x_T = \bar{x}$, *if* $\tilde{x} \equiv \bar{x}$, *a constant.*

Convergence in probability generalizes the standard notion of a limit of a deterministic sequence. Of a deterministic sequence x_T, we say it converges to some limit $\bar{x}$ if, for $\kappa > 0$, there exists a T_κ : for each $T \geq T_\kappa$ we have that $|x_T - \bar{x}| < \kappa$. Convergence in probability can also be restated as saying that:

$$\lim_{T\to\infty} \Pr\left(|x_{Ti} - \tilde{x}_i| > \delta\right) = 0.$$

The following is a stronger notion of convergence:

ALMOST SURE CONVERGENCE. *A sequence of random vectors* $\{x_T\}$ converges almost surely *to the random vector* $\tilde{x}$ *if, for each* $i = 1, 2, \cdots, N$, *we have*:

$$\Pr\left(\omega : x_{Ti}(\omega) \to \tilde{x}_i\right) = 1,$$

where ω *denotes the entire random sequence* x_{Ti}. *This is succinctly written as* $x_T \xrightarrow{a.s.} \tilde{x}$.

Almost sure convergence implies convergence in probability. Convergence in probability means that for each $\epsilon > 0$, $\lim_{T\to\infty} \Pr\left(\omega : |x_{Ti}(\omega) - \tilde{x}_i| < \epsilon\right) = 1$. Almost sure convergence requires that $\Pr\left(\lim_{T\to\infty} x_{Ti} \to \tilde{x}_i\right) = 1$ or that

$$\lim_{T'\to\infty} \Pr\left(\sup_{T\geq T'} |x_{Ti} - \tilde{x}_i| > \delta\right) = \lim_{T'\to\infty} \Pr\left(\bigcup_{T\geq T'} |x_{Ti} - \tilde{x}_i| > \delta\right) = 0.$$

Next, assume that the second order moments of all x_i are finite. We have:

CONVERGENCE IN QUADRATIC MEAN. *A sequence of random vectors* $\{x_T\}$ converges in quadratic mean *to the random vector* $\tilde{x}$ *if for each* $i = 1, 2, \cdots, N$, *we have*:

$$\lim_{T'\to\infty} E\left[(x_{Ti} - \tilde{x}_i)^2\right] \to 0.$$

This is succinctly written as $x_T \xrightarrow{q.m.} \tilde{x}$.

REMARK. By Chebyshev's inequality,

$$\Pr\left(|x_{Ti} - \tilde{x}_i| > \delta\right) \leq \frac{E[(x_{Ti} - \tilde{x}_i)^2]}{\delta^2},$$

which shows that convergence in quadratic mean implies convergence in probability.

We now turn to a weaker notion of convergence:

CONVERGENCE IN DISTRIBUTION. *Let* $\{f_T(\cdot)\}_T$ *be the sequence of probability distributions* (*that is,* $f_T(x) = pr\left(x_T \leq x\right)$) *of the sequence of the random vectors* $\{x_T\}$. *Let* $\tilde{x}$ *be a random vector with probability distribution* $f(x)$. *A sequence* $\{x_T\}$ *converges in distribution to* $\tilde{x}$ *if, for each* $i = 1, 2, \cdots, N$, *we have:*

$$\lim_{T\to\infty} f_T(x) = f(x).$$

This is succinctly written as $x_T \xrightarrow{d} \tilde{x}$.

The following two results are useful to the purpose of this chapter:

SLUTZKY'S THEOREM. *If* $y_T \xrightarrow{p} \bar{y}$ *and* $x_T \xrightarrow{d} \tilde{x}$, *then:*

$$y_T \cdot x_T \xrightarrow{d} \bar{y} \cdot \tilde{x}.$$

CRAMER-WOLD DEVICE. *Let* λ *be a* N*-dimensional vector of constants. We have*:

$$x_T \xrightarrow{d} \tilde{x} \Leftrightarrow \lambda^\top \cdot x_T \xrightarrow{d} \lambda^\top \cdot \tilde{x}.$$

The following example illustrates the Cramer-Wold device. If $\lambda^\top \cdot x_T \xrightarrow{d} N\left(0; \lambda^\top \Sigma \lambda\right)$, then $x_T \xrightarrow{d} N\left(0; \Sigma\right)$.

We now state two laws about convergence in probability.

WEAK LAW (NO. 1) (Khinchine). *Let* $\{x_T\}$ *be a i.i.d. sequence satistfying* $E(x_T) = \mu < \infty \ \forall T$. *We have*:

$$\bar{x}_T \equiv \frac{1}{T}\sum_{t=1}^{T} x_t \xrightarrow{p} \mu.$$

WEAK LAW (NO. 2) (Chebyshev). *Let* $\{x_T\}$ *be a sequence independent but not identically distributed, satisfying* $E(x_T) = \mu_T < \infty$ *and* $E\left[(x_T - \mu_T)^2\right] = \sigma_T^2 < \infty$. *If* $\lim_{T\to\infty} \frac{1}{T^2}\sum_{t=1}^{T} \sigma_t^2 \to 0$, *then:*

$$\bar{x}_T \equiv \frac{1}{T}\sum_{t=1}^{T} x_t \xrightarrow{p} \bar{\mu}_T \equiv \frac{1}{T}\sum_{t=1}^{T} \mu_t.$$

We now state and provide a proof of the central limit theorem in a simple setting.

CENTRAL LIMIT THEOREM. *Let* $\{x_T\}$ *be a i.i.d. sequence, satisfying* $E(x_T) = \mu < \infty$ *and* $E\left[(x_T - \mu)^2\right] = \sigma^2 < \infty \ \forall T$. *Let* $\bar{x}_T \equiv \frac{1}{T}\sum_{t=1}^{T} x_t$. *We have,*

$$\frac{\sqrt{T}\left(\bar{x}_T - \mu\right)}{\sigma} \xrightarrow{d} N(0,1).$$

The multidimensional version of this theorem requires a mere change in notation. For the proof, the classic method relies on the characteristic functions. Let:

$$\varphi(t) \equiv E\left(e^{\mathtt{i}tx}\right) = \int e^{\mathtt{i}tx} f(x)dx, \ \mathtt{i} \equiv \sqrt{-1}.$$

We have $\frac{\partial^r}{\partial t^r}\varphi(t)\big|_{t=0} = \mathtt{i}^r m^{(r)}$, where $m^{(r)}$ is the r-th order moment. By a Taylor's expansion,

$$\varphi(t) = \varphi(0) + \frac{\partial}{\partial t}\varphi(t)\bigg|_{t=0} t + \frac{1}{2}\frac{\partial^2}{\partial t^2}\varphi(t)\bigg|_{t=0} t^2 + \cdots = 1 + \mathtt{i}m^{(1)}t - m^{(2)}\frac{1}{2}t^2 + \cdots.$$

Next, let $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$, and consider the random variable,

$$Y_T \equiv \frac{\sqrt{T}\left(\bar{x}_T - \mu\right)}{\sigma} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \frac{x_t - \mu}{\sigma}.$$

The characteristic function of Y_T is the product of the characteristic functions of $a_t \equiv \frac{x_t-\mu}{\sqrt{T}\sigma}$, which are all the same: $\varphi_{Y_T}(t) = (\varphi_a(t))^T$, where $\varphi_a(t) = 1 - \frac{t^2}{2T} + \cdots$. Therefore,

$$\varphi_{Y_T}(t) = \varphi\left(\frac{t}{\sqrt{T}}\right)^T = \left(1 - \frac{1}{2}\frac{t^2}{T} + o\left(T^{-1}\right)\right)^T.$$

Clearly, $\lim_{T\to\infty} \varphi_{Y_T}(t) = e^{-\frac{1}{2}t^2}$, which is the characteristic function of a standard Gaussian variable.

6.11 Appendix 3: Theory for maximum likelihood estimation

Assume that $\hat{\theta}_T \overset{a.s.}{\to} \theta_0$, and that $H(y,\theta) \equiv \nabla_{\theta\theta} \ln L(\theta|\, y)$ exists, it is continuous in θ uniformly in y and that we can differentiate twice inside the integral $\int L(\theta|\, y)dy = 1$. We have:

$$s_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} \nabla_\theta \ln L(\theta|\, y).$$

Consider the c-parametrized curves $\theta(c) = c\square(\theta_0 - \hat{\theta}_T) + \hat{\theta}_T$ where, for all $c \in (0,1)^p$ and $\theta \in \Theta$, $c\square\theta$ denotes a vector in Θ where the ith element is $c^{(i)}\theta^{(i)}$. By the intermediate value theorem, there exists then a c^* in $(0,1)^p$ such that we have almost surely:

$$s_T(\hat{\theta}_T) = s_T(\theta_0) + H_T(\theta^*) \cdot (\hat{\theta}_T - \theta_0),$$

where $\theta^* \equiv \theta(c^*)$ and:

$$H_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} H(\theta|\, y_t).$$

The first order conditions tell us that $s_T(\hat{\theta}_T) = 0$. Hence,

$$0 = s_T(\theta_0) + H_T(\theta_T^*) \cdot (\hat{\theta}_T - \theta_0).$$

We also have that:

$$|H_T(\theta_T^*) - H_T(\theta_0)| \leq \frac{1}{T}\sum_{t=1}^{T} |H(\theta_T^*) - H_T(\theta_0)| \leq \sup |H(\theta_T^*) - H_T(\theta_0)|, \tag{6A.1}$$

where the supremum is taken over the set of all the observations. Since $\hat{\theta}_T \overset{a.s.}{\to} \theta_0$, we also have that $\theta_T^* \overset{a.s.}{\to} \theta_0$. Moreover, by the law of large numbers,

$$H_T(\theta_0) = \frac{1}{T}\sum_{t=1}^{T} H(\theta_0|\, y_t) \overset{p}{\to} E[H(\theta_0|\, y_t)] = -\mathcal{J}(\theta_0). \tag{6A.2}$$

Since H is continuous in θ uniformly in y, the inequality in (6A.1), and (6A.2) both imply that:

$$H_T(\theta_T^*) \overset{a.s.}{\to} -\mathcal{J}(\theta_0).$$

Therefore, as $T \to \infty$,

$$\sqrt{T}\left(\hat{\theta}_T - \theta_0\right) = -H_T^{-1}(\theta_0) \cdot s_T(\theta_0)\sqrt{T} = \mathcal{J}^{-1} \cdot \sqrt{T} s_T(\theta_0).$$

By the central limit theorem, and $E(s_T) = 0$, the score, $s_T(\theta_0) = \frac{1}{T}\sum_{t=1}^{T} s(\theta_0, y_t)$, is such that

$$\sqrt{T} \cdot s_T(\theta_0) \overset{d}{\to} N(0, var(s(\theta_0, y_t))),$$

where

$$var(s(\theta_0, y_t)) = \mathcal{J}.$$

The result follows by the Slutzky's theorem and the symmetry of $\mathcal{J}$.

Finally, one should show the existence of a sequence $\hat{\theta}_T$ converging a.s. to θ_0. Proofs on this type of convergence can be found in Amemiya (1985), or in Newey and McFadden (1994).

6.12 Appendix 4: Dependent processes

6.12.1 Weak dependence

Let $\sigma_T^2 = var(\sum_{t=1}^T x_t)$, and assume that that $\sigma_T^2 = O(T)$, and that $\sigma_T^2 = O\left(T^{-1}\right)$. If

$$\sigma_T^{-1} \sum_{t=1}^{T} (x_t - E(x_t)) \xrightarrow{d} N(0,1),$$

we say that $\{x_t\}$ is *weakly dependent*. Of a process, we say it is "nonergodic," when it exhibits such a strong dependence that it does not even satisfy the law of large numbers.

- Stationarity
- Weak dependence
- Ergodicity

6.12.2 The central limit theorem for martingale differences

Let x_t be a martingale difference sequence with $E\left(x_t^2\right) = \sigma_t^2 < \infty$ for all t, and define $\bar{x}_T \equiv \frac{1}{T}\sum_{t=1}^T x_t$, and $\bar{\sigma}_T^2 \equiv \frac{1}{T}\sum_{t=1}^T \sigma_t^2$. Let,

$$\forall \epsilon > 0, \ \lim_{T\to\infty} \frac{1}{T\bar{\sigma}_T^2} \sum_{t=1}^{T} x_t^2 \mathbb{I}_{|x_t| \geq \epsilon \cdot T \cdot \bar{\sigma}_T^2} = 0, \quad \text{and} \quad \frac{1}{T}\sum_{t=1}^{T} x_t^2 - \bar{\sigma}_T^2 \xrightarrow{p} 0.$$

Under the previous condition,

$$\frac{\sqrt{T}\cdot \bar{x}_T}{\bar{\sigma}_T} \xrightarrow{d} N(0,1).$$

6.12.3 Applications to maximum likelihood

We use the central limit theorem for martingale differences to prove asymptotic normality of the MLE, in the case of weakly dependent processes. We have,

$$\ln L_T(\theta) = \sum_{t=1}^{T} \ell_t(\theta), \quad \ell_t(\theta) \equiv \ell(\theta; y_t | x_t).$$

The MLE satisfies the following first order conditions,

$$\mathbf{0}_p = \nabla_\theta \ln L_T(\theta)|_{\theta=\hat{\theta}_T} \overset{d}{=} \sum_{t=1}^{T} \nabla_\theta \ell_t(\theta)|_{\theta=\theta_0} + \sum_{t=1}^{T} \nabla_{\theta\theta} \ell_t(\theta)|_{\theta=\theta_0} (\hat{\theta}_T - \theta_0),$$

whence

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \overset{d}{=} -\left[\frac{1}{T}\sum_{t=1}^{T} \nabla_{\theta\theta}\ell_t(\theta_0)\right]^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla_\theta \ell_t(\theta_0). \tag{6A.3}$$

We have:

$$E_{\theta_0}\left[\nabla_\theta \ell_{t+1}(\theta_0) | F_t\right] = \mathbf{0}_p,$$

which shows that $\frac{\partial \ell_t(\theta_0)}{\partial \theta}$ is a martingale difference. Naturally, here we also have that:

$$E_{\theta_0}\left(|\nabla_\theta \ell_{t+1}(\theta_0)|_2 | F_t\right) = -E_{\theta_0}\left(\nabla_{\theta\theta}\ell_{t+1}(\theta_0) | F_t\right) \equiv \mathcal{J}_t(\theta_0).$$

Next, for a given constant $c \in \mathbb{R}^p$, let:

$$x_t \equiv c^\top \nabla_\theta \ell_t(\theta_0).$$

Clearly, x_t is also a martingale difference. Furthermore,

$$E_{\theta_0}\left(x_{t+1}^2 \middle| F_t\right) = -c^\top \mathcal{J}_t(\theta_0)\, c,$$

and because x_t is a martingale difference, $E(x_t x_{t-i}) = E[E(x_t \cdot x_{t-i}| F_{t-i})] = E[E(x_t| F_{t-i}) \cdot x_{t-i}] = 0$, for all i. That is, x_t and x_{t-i} are mutually uncorrelated. It follows that,

$$\begin{aligned} var\left(\sum_{t=1}^T x_t\right) &= \sum_{t=1}^T E\left(x_t^2\right) \\ &= \sum_{t=1}^T c^\top E_{\theta_0}\left(|\nabla_\theta \ell_t(\theta_0)|_2\right) c \\ &= \sum_{t=1}^T c^\top E_{\theta_0}\left[E_{\theta_0}\left(|\nabla_\theta \ell_t(\theta_0)|_2| F_{t-1}\right)\right] c \\ &= -\sum_{t=1}^T c^\top E_{\theta_0}\left[\mathcal{J}_{t-1}(\theta_0)\right] c \\ &= -c^\top \left[\sum_{t=1}^T E_{\theta_0}\left(\mathcal{J}_{t-1}(\theta_0)\right)\right] c. \end{aligned}$$

Next, define:

$$\bar{x}_T \equiv \frac{1}{T}\sum_{t=1}^T x_t \quad \text{and} \quad \bar{\sigma}_T^2 \equiv \frac{1}{T}\sum_{t=1}^T E\left(x_t^2\right) = -c^\top \left[\frac{1}{T}\sum_{t=1}^T E_{\theta_0}\left(\mathcal{J}_{t-1}(\theta_0)\right)\right] c.$$

Under the conditions underlying the central limit theorem for weakly dependent processes provided earlier, to be spelled out below,

$$\frac{\sqrt{T}\bar{x}_T}{\bar{\sigma}_T} \xrightarrow{d} N(0,1).$$

By the Cramer-Wold device,

$$\left[\frac{1}{T}\sum_{t=1}^T E_{\theta_0}\left(\mathcal{J}_{t-1}(\theta_0)\right)\right]^{-1/2} \frac{1}{\sqrt{T}}\sum_{t=1}^T \nabla_\theta \ell_t(\theta_0) \xrightarrow{d} N(0, \mathbf{I}_p).$$

The conditions that need to be satisfied are,

$$\frac{1}{T}\sum_{t=1}^T \nabla_{\theta\theta}\ell_t(\theta_0) - \frac{1}{T}\sum_{t=1}^T E_{\theta_0}\left[\mathcal{J}_{t-1}(\theta_0)\right] \xrightarrow{p} 0, \quad \text{and} \quad \text{plim}\frac{1}{T}\sum_{t=1}^T E_{\theta_0}\left[\mathcal{J}_{t-1}(\theta_0)\right] \equiv \mathcal{J}_\infty(\theta_0).$$

Under the previous conditions, it follows from Eq. (6A.3) that,

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N\left(\mathbf{0}_p, \mathcal{J}_\infty(\theta_0)^{-1}\right).$$

6.13 Appendix 5: Proof of Theorem 6.4

Let $\pi_t \equiv \pi_t\left(\phi\left(y(t+1), \mathbf{M}-(t+1)\mathbf{1}_{d-q}\right)\middle|\, \phi\left(y(t), \mathbf{M}-t\mathbf{1}_{d-q}\right)\right)$ denote the transition density of

$$\phi\left(y(t), \mathbf{M}-t\mathbf{1}_{d-q}\right) \equiv \phi\left(y(t)\right) \equiv \left(y^o(t), c(y(t), M_1-t), \cdots, c(y(t), M_{d-q}-t)\right),$$

where we have emphasized the dependence of ϕ on the time-to-maturity vector:

$$\mathbf{M}-t\mathbf{1}_{d-q} \equiv (M_1-t, \cdots, M_{d-q}-t).$$

By $\Sigma(\tau)$ full rank $P \otimes d\tau$-a.s., and Itô's lemma, ϕ satisfies, for $\tau \in [t, t+1]$,

$$\left\{ \begin{array}{lcl} dy^o(\tau) & = & b^o(\tau)d\tau + F(\tau)\Sigma(\tau)dW(\tau) \\ dc(\tau) & = & b^c(\tau)d\tau + \nabla c(\tau)\Sigma(\tau)dW(\tau) \end{array} \right.$$

where b^o and b^c are, respectively, q-dimensional and $(d-q)$-dimensional measurable functions, and $F(\tau) \equiv \bar{a}(\tau)\cdot\Sigma(\tau)^{-1}$ $P \otimes d\tau$-a.s. Under condition (6.28), π_t is not degenerate. Furthermore, $C\left(y(t); \ell\right) \equiv C(t)$ is deterministic in $\ell \equiv (\ell_1, \cdots, \ell_{d-q})$. That is, for all $(\bar{c}, \bar{c}^+) \in \mathbb{R}^d \times \mathbb{R}^d$, there exists a function μ such that for any neighbourhood $N(\bar{c}^+)$ of $\bar{c}^+$, there exists another neighborhood $N(\mu(\bar{c}^+))$ of $\mu(\bar{c}^+)$ such that,

$$\begin{array}{rl} & \left\{\omega \in \Omega : \phi\left(y(t+1), \mathbf{M}-(t+1)\mathbf{1}_{d-q}\right) \in N(\bar{c}^+)\middle|\ \ \phi\left(y(t), \mathbf{M}-t\mathbf{1}_{d-q}\right) = \bar{c}\right\} \\ = & \left\{\omega \in \Omega : (y^o(t+1), c(y(t+1), M_1-t)), \cdots, c(y(t+1), M_{d-q}-t)) \in N(\mu(\bar{c}^+))\right. \\ & \qquad\qquad \left.\middle|\phi\left(y(t), \mathbf{M}-t\mathbf{1}_{d-q}\right) = \bar{c}\right\} \\ = & \left\{\omega \in \Omega : (y^o(t+1), c(y(t+1), M_1-t)), \cdots, c(y(t+1), M_{d-q}-t)) \in N(\mu(\bar{c}^+))\right. \\ & \qquad\qquad \left.\middle|(y^o(t), c(y(t), M_1-t), \cdots, c(y(t), M_{d-q}-t)) = \bar{c}\right\} \end{array}$$

where the last equality follows by the definition of ϕ. In particular, the transition laws of ϕ_t^c given ϕ_{t-1}^c are not degenerate; and ϕ_t^c is stationary. The feasibility of simulation based method of moments estimation is proved. The efficiency claim follows by the Markov property of ϕ, and the usual score martingale difference argument. ■

References

Altissimo, F. and A. Mele (2009): "Simulated Nonparametric Estimation of Dynamic Models." *Review of Economic Studies* 76, 413-450.

Amemiya, T. (1985): *Advanced Econometrics*. Cambridge, Mass.: Harvard University Press.

Bergman, Y. Z., B. D. Grundy, and Z. Wiener (1996): "General Properties of Option Prices." *Journal of Finance* 51, 1573-1610.

Brandt, M. and P. Santa-Clara (2002): "Simulated Likelihood Estimation of Diffusions with an Applications to Exchange Rate Dynamics in Incomplete Markets." *Journal of Financial Economics* 63, 161-210.

Carrasco, M., M. Chernov, J.-P. Florens and E. Ghysels (2007): "Efficient Estimation of General Dynamic Models with a Continuum of Moment Conditions." *Journal of Econometrics* 140, 529-573.

Chernov, M. and E. Ghysels (2000): "A Study towards a Unified Approach to the Joint Estimation of Objective and Risk-Neutral Measures for the Purpose of Options Valuation." *Journal of Financial Economics* 56, 407-458.

Christensen, B. J. (1992): "Asset Prices and the Empirical Martingale Model." Working paper, New York University.

Duffie, D. and K. J. Singleton (1993): "Simulated Moments Estimation of Markov Models of Asset Prices." *Econometrica* 61, 929-952.

Fermanian, J.-D. and B. Salanié (2004): "A Nonparametric Simulated Maximum Likelihood Estimation Method." *Econometric Theory* 20, 701-734.

Fisher, R. A. (1912): "On an Absolute Criterion for Fitting Frequency Curves." *Messages of Mathematics* 41, 155-157.

Gallant, A. R. and G. Tauchen (1996): "Which Moments to Match?" *Econometric Theory* 12, 657-681.

Gauss, C. F. (1816): "Bestimmung der Genanigkeit der Beobachtungen." *Zeitschrift für Astronomie und Verwandte Wissenschaften* 1, 185-196.

Gouriéroux, C., A. Monfort and E. Renault (1993): "Indirect Inference." *Journal of Applied Econometrics* 8, S85-S118.

Hajivassiliou, V. and D. McFadden (1998): "The Method of Simulated Scores for the Estimation of Limited-Dependent Variable Models." *Econometrica* 66, 863-896.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50, 1029-1054.

Hansen, L. P. and K. J. Singleton (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models." *Econometrica* 50, 1269-1286.

Hansen, L. P. and K. J. Singleton (1983): "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns." *Journal of Political Economy* 91, 249-265.

Harrison, J. M. and S. R. Pliska (1983): "A Stochastic Calculus Model of Continuous Trading: Complete Markets." *Stochastic Processes and their Applications* 15, 313-316.

Heston, S. (1993): "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options." *Review of Financial Studies* 6, 327-343.

Laroque, G. and B. Salanié (1989): "Estimation of Multimarket Fix-Price Models: An Application of Pseudo-Maximum Likelihood Methods." *Econometrica* 57, 831-860.

Laroque, G. and B. Salanié (1993): "Simulation-Based Estimation of Models with Lagged Latent Variables." *Journal of Applied Econometrics* 8, S119-S133.

Laroque, G. and B. Salanié (1994): "Estimating the Canonical Disequilibrium Model: Asymptotic Theory and Finite Sample Properties." *Journal of Econometrics* 62, 165-210.

Lee, B-S. and B. F. Ingram (1991): "Simulation Estimation of Time-Series Models." *Journal of Econometrics* 47, 197-207.

Lee, L. F. (1995): "Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models." *Econometric Theory* 11, 437-483.

McFadden, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration." *Econometrica* 57, 995-1026.

Mele, A. (2003): "Fundamental Properties of Bond Prices in Models of the Short-Term Rate." *Review of Financial Studies* 16, 679-716.

Newey, W. K. and D. L. McFadden (1994): "Large Sample Estimation and Hypothesis Testing." In: Engle, R. F. and D. L. McFadden (Editors): *Handbook of Econometrics*, Vol. 4, Chapter 36, 2111-2245. Amsterdam: Elsevier.

Neyman, J. and E. S. Pearson (1928): "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference." *Biometrika* 20A, 175-240, 263-294.

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators." *Econometrica* 57, 1027-1057.

Pastorello, S., E. Renault and N. Touzi (2000): "Statistical Inference for Random-Variance Option Pricing." *Journal of Business and Economic Statistics* 18, 358-367.

Pastorello, S., V. Patilea, and E. Renault (2003): "Iterative and Recursive Estimation in Structural Non Adaptive Models." *Journal of Business and Economic Statistics* 21, 449-509.

Pearson, K. (1894): "Contributions to the Mathematical Theory of Evolution." *Philosophical Transactions of the Royal Society of London*, Series A 185, 71-78.

Santa-Clara, P. (1995): "Simulated Likelihood Estimation of Diffusions With an Application to the Short Term Interest Rate." Ph.D. dissertation, INSEAD.

Singleton, K. J. (2001): "Estimation of Affine Asset Pricing Models Using the Empirical Characteristic Function." *Journal of Econometrics* 102, 111-141.

Smith, A. (1993): "Estimating Nonlinear Time Series Models Using Simulated Vector Autoregressions." *Journal of Applied Econometrics* 8, S63-S84.

Tauchen, G. (1997): "New Minimum Chi-Square Methods in Empirical Finance." In D. Kreps and K. Wallis (Editors): *Advances in Econometrics*, 7th World Congress, Econometrics Society Monographs, Vol. III. Cambridge UK: Cambridge University Press, 279-317.

Part II

Empirical lessons and market inefficiencies

7
Neo-classical kernels and puzzles

7.1 Introduction

Asset pricing models impose a number of restrictions on security returns, which can conveniently be summarized by a few but key properties of the pricing kernel, consistent with a data-reduction principle. This chapter discusses methods of statistical inference based on this data-reduction, by relying on restrictions on the moments of the pricing kernel based on the celebrated Hansen and Jagannathan (1991) bounds—evidence against any model mounts when the volatility of the pricing kernel is below a certain threshold. We illustrate how these bounds are useful, by revisiting a simplified version of the Lucas' tree model introduced in Chapter 3. We shall examine issues within this methodology arising in finite samples, and review the efforts needed to tackle them.

The next section contains a simpified version of the Lucas model, which we take as a useful benchmark in this chapter. Section 7.3 develops the central tools of analysis, a non-parametric bound on the volatility of the pricing kernels, i.e. the risk-premium, consistent with a given level of the short-term rate. Section 7.4 considers multifactor extensions, with closed-form solutions, arising under a number of analytically convenient assumptions on the stochastic discounting factor. One of the striking points of Section 7.4 is that in spite of these added dimensionalities, the resulting models might spectacularly fail explain the dynamics of asset prices—pumping up volatility is not enough, if this added volatility is not accompanied by time-varying countercyclical statistics. Section 7.5 develops a link between stochastic discount factors and Sharpe ratios, and Section 7.6 develops dynamic versions of the core bounds at the heart of this chapter.

7.2 The equity premium puzzle

7.2.1 A single-factor model

7.2.1.1 Assumptions

We consider an economy with a single agent with CRRA equal to γ, and a constant discount factor β. We assume cum-dividends gross returns, $(S_t + D_t)/\, S_{t-1}$, are generated by:

$$\begin{cases} \ln(S_t + D_t) &= \ln S_{t-1} + \mu_S - \dfrac{1}{2}\sigma_S^2 + \epsilon_{S,t} \\ \ln D_t &= \ln D_{t-1} + \mu_D - \dfrac{1}{2}\sigma_D^2 + \epsilon_{D,t} \end{cases} \tag{7.1}$$

where

$$\begin{bmatrix} \epsilon_{S,t} \\ \epsilon_{D,t} \end{bmatrix} \sim NID\left(\mathbf{0}_2; \begin{bmatrix} \sigma_S^2 & \sigma_{SD} \\ \sigma_{SD} & \sigma_D^2 \end{bmatrix}\right).$$

The second equation in (7.1) is obviously given. The first, is endogenous, so to speak. We want to find parameter values such that Eqs. (7.1) hold in equilibrium.

7.2.1.2 Euler's restrictions

The coefficients μ_S, μ_D, σ_S^2, σ_D^2 and σ_{SD} need to satisfy joint restrictions consistent with Euler equations, which is indeed the equilibrium outcome, as explained in Part I of the Lectures. By results in Part I of the Lectures, these Euler equations are:

$$1 = E\left(m_{t+1} \cdot e^{Q_{t+1}} \middle|\, F_t\right), \quad Q_{t+1} = \ln\left(\frac{S_{t+1} + D_{t+1}}{S_t}\right), \quad m_{t+1} \equiv e^{Z_{t+1}} \equiv \beta\left(\frac{D_{t+1}}{D_t}\right)^{-\gamma}, \tag{7.2}$$

where F_t is the information set as of time t, and m_{t+1} is the stochastic discounting factor. Naturally, Eq. (7.2) holds for any asset. In particular, it holds for a one-period bond with price $S_t^b \equiv b_t$, $S_{t+1}^b \equiv 1$ and $D_{t+1}^b \equiv 0$. Define, then, $Q_{t+1}^b \equiv \ln\left(b_t^{-1}\right) \equiv \ln R_t$. By replacing R_t into Eq. (7.2), one gets $R_t^{-1} = E\left(e^{Z_{t+1}} \middle|\, F_t\right)$, such that,

$$\frac{1}{R_t} = E\left(m_{t+1} \middle|\, F_t\right), \quad 1 = E\left(m_{t+1} \cdot e^{Q_{t+1}} \middle|\, F_t\right). \tag{7.3}$$

The following result helps determine the expectations in Eqs. (7.3) in closed-form:

LEMMA 7.1: *Let Z_{t+1} be normally distributed, conditionally upon the information set at time t. Then, for any $\alpha \in \mathbb{R}$,*

$$\begin{aligned} E\left(e^{-\alpha Z_{t+1}} \middle|\, F_t\right) &= e^{-\alpha E(Z_{t+1}|F_t) + \frac{1}{2}\alpha^2 var(Z_{t+1}|F_t)} \\ \sqrt{var\left(e^{-\alpha Z_{t+1}} \middle|\, F_t\right)} &= e^{-\alpha E(Z_{t+1}|F_t) + \alpha^2 var(Z_{t+1}|F_t)} \sqrt{1 - e^{-\alpha^2 var(Z_{t+1}|F_t)}} \end{aligned}$$

By the definition of Z, Eq. (7.1), and Lemma 7.1,

$$\frac{1}{R_t} = E\left(e^{Z_{t+1}} \middle|\, F_t\right) = e^{E(Z_{t+1}|F_t) + \frac{1}{2}var(Z_{t+1}|F_t)} = e^{\ln\beta - \gamma\left(\mu_D - \frac{1}{2}\sigma_D^2\right) + \frac{1}{2}\gamma^2\sigma_D^2}. \tag{7.4}$$

Therefore, the equilibrium interest rate is constant—its expression is given in the second of Eqs. (7.5) below.

The second of equations (7.3) can be written as,

$$1 = E\left(\exp\left(Z_{t+1}+Q_{t+1}\right)\middle|\, F_t\right) = e^{\ln\beta-\gamma\left(\mu_D-\frac{1}{2}\sigma_D^2\right)+\mu_S-\frac{1}{2}\sigma_S^2}\cdot E\left(e^{\tilde{n}_{t+1}}\middle|\, F_t\right),$$

where $\tilde{n}_{t+1}\equiv\epsilon_{S,t+1}-\gamma\epsilon_{D,t+1}\sim N(0,\sigma_S^2+\gamma^2\sigma_D^2-2\gamma\sigma_{SD})$. The expectation in the above equation can be determined through Lemma 7.1, resulting in,

$$\underbrace{\mu_S-r}_{\text{risk premium}} = \gamma\sigma_{SD}, \quad \text{and} \quad r_t\equiv\ln R_t = -\ln\beta+\gamma\mu_D-\frac{1}{2}\gamma\left(\gamma+1\right)\sigma_D^2, \tag{7.5}$$

where we have used Eq. (7.4) to calculate the interest rate, r. Note that the expected gross return on the risky asset is,

$$E\left[\left.\frac{S_{t+1}+D_{t+1}}{S_t}\right|\, F_t\right] = e^{\mu_S-\frac{1}{2}\sigma_S^2}\cdot E\left[e^{\epsilon_{S,t+1}}\middle|\, F_t\right] = e^{\mu_S} = e^{r+\gamma\sigma_{SD}}.$$

Therefore, if $\sigma_{SD}>0$, then $E\left(\left(S_{t+1}+D_{t+1}\right)/S_t\middle|\, F_t\right) > E\left(b_t^{-1}\middle|\, F_t\right)$, as expected.

The expressions for the equity premium and the short-term rate are the discrete-time counterpart to those derived in Part I (see Chapter 5). Consider, for example, the interest rate. The second term, $\gamma\mu_D$, reflects "intertemporal substitution" effects: consumption endowment increases, on average, as μ_D increases, which reduces the demand for bonds, thereby increasing the interest rate. The last term, instead, relates to "precautionary" motives: an increase in the uncertainty related to consumption endowment, σ_D, raises concerns with our representative agent, who then increases his demand for bonds, thereby leading to a drop in the interest rate.

7.2.1.3 Solution for the P/D ratio and absence of sunspots

We check the internal consistency of the model. The coefficients of the model satisfy some restrictions. In particular, the asset price volatility must be determined endogenously. Let us conjecture, and later verify, that the following "no-sunspots" condition holds, for each period t:

$$\epsilon_{S,t} = \epsilon_{D,t}. \tag{7.6}$$

By Eqs. (7.5) and (7.6),

$$\mu_S - r = \lambda\sigma_D, \quad \lambda\equiv\gamma\sigma_D, \tag{7.7}$$

and by the definition of Z_{t+1} in Eq. (7.2),

$$Z_{t+1} = -\left(r+\frac{1}{2}\lambda^2\right)-\lambda u_{D,t+1}, \quad u_{D,t+1}\equiv\frac{\epsilon_{D,t+1}}{\sigma_D},$$

such that we can define the pricing kernel ξ, from the stochastic discounting factor, recursively, as follows:

$$\frac{\xi_{t+1}}{\xi_t} = m_{t+1}\equiv e^{Z_{t+1}}, \quad \xi_0 = 1.$$

It is the discrete-time counterpart to the continuous-time representation of the Arrow-Debreu state price density given in Chapter 5.

Let us iterate the asset price equation (7.2),

$$\begin{aligned} S_t &= E\left[\left(\prod_{j=1}^{n} e^{Z_{t+i}}\right)\cdot S_{t+n}\middle| F_t\right] + \sum_{i=1}^{n} E\left[\left(\prod_{j=1}^{i} e^{Z_{t+j}}\right)\cdot D_{t+i}\middle| F_t\right] \\ &= E\left(\frac{\xi_{t+n}}{\xi_t}\cdot S_{t+n}\middle| F_t\right) + \sum_{i=1}^{n} E\left(\frac{\xi_{t+i}}{\xi_t}\cdot D_{t+i}\middle| F_t\right). \end{aligned}$$

By letting $n\to\infty$ and assuming the first term in the previous equation goes to zero,

$$S_t = \sum_{i=1}^{\infty} E\left(\frac{\xi_{t+i}}{\xi_t}\cdot D_{t+i}\middle| F_t\right). \tag{7.8}$$

Eq. (7.8) holds, as just mentioned, under a transversality condition, similar to that analyzed in Chapter 5 (see Section 5.3.3), which always holds, under the inequalities given in (7.9) below.

The expectations in Eq. (7.8) are, by Lemma 7.1,

$$E\left(\frac{\xi_{t+i}}{\xi_t}\cdot D_{t+i}\middle| F_t\right) = E\left(e^{\sum_{j=1}^{i} Z_{t+j}}\cdot D_{t+i}\middle| F_t\right) = D_t e^{(\mu_D - r - \sigma_D\lambda)i}.$$

Suppose the "risk-adjusted" discount rate $r+\sigma_D\lambda$ is higher than the growth rate of the economy, viz

$$r+\sigma_D\lambda > \mu_D \Leftrightarrow k \equiv e^{\mu_D - r - \sigma_D\lambda} < 1. \tag{7.9}$$

Under this condition, the summation in Eq. (7.8) converges, leaving:

$$\frac{S_t}{D_t} = \frac{k}{1-k}. \tag{7.10}$$

Eq. (7.10) links to the celebrated Gordon's formula (Gordon, 1962). The price-dividend ratio increases with the expected dividend growth, μ_D, and decreases with the (risk-adjusted) discount rate, $r+\sigma_D\lambda$. The model predicts that price-dividend ratios are constant, a counterfactual prediction addressed in the next two chapters. Finally, the solution for the price-dividend ratio in Eq. (7.10) is, of course, consistent with that of the Lucas model in Part I (see Section 3.2.4 in Chapter 3), as shown below.

We now check that the no-sunspots condition in Eq. (7.6) holds, and derive the asset price variance. By Eq. (7.10), $\ln(S_t+D_t) = -\ln(1-k)+\ln D_t$ & $\ln S_{t-1} = \ln k - \ln(1-k) + \ln D_{t-1}$, or

$$\ln(S_t+D_t) - \ln S_{t-1} = -\ln k + \ln D_t - \ln D_{t-1} = -\ln k + \mu_D - \frac{1}{2}\sigma_D^2 + \epsilon_{D,t},$$

where the last equaility follows by the second of Eqs. (7.1). By identifying terms in the previous equation and the first of Eqs. (7.1) leaves:

$$\mu_S - \frac{1}{2}\sigma_S^2 = \mu_D - \frac{1}{2}\sigma_D^2 - \ln k, \quad \epsilon_{S,t} = \epsilon_{D,t}, \text{ for each } t. \tag{7.11}$$

The second condition confirms the no-sunspots condition in Eq. (7.6) holds. It also informs us that, $\sigma_S^2 = \sigma_{SD} = \sigma_D^2$. By replacing this into the first condition confirms indeed that $\mu_S = \mu_D - \ln k = r + \sigma_D\lambda$.

Note also that by replacing the expression for the interest rate in the second of Eqs. (7.5) and the equity premium in Eq. (7.7) into Eq. (7.9), the constant k simplifies to $k \equiv \beta e^{(\gamma-1)\left(-\mu_D + \frac{1}{2}\gamma\sigma_D^2\right)}$, such that the price-dividend ratio in the log-utility, $\gamma=1$, collapses to $\beta/(1-\beta)$, as established in Chapter 3. This section provides a solution to the general CRRA case, under the additional assumption that dividends are normally distributed.

7.2.2 Extensions

Chapter 3 shows that within a IID environment, prices are convex in the dividend D_t if $\gamma > 1$, and concave if $\gamma < 1$. Eq. (7.10) reveals this property may be lost in a dynamic environment. The next chapter shows that in such an environment, the convexity of the price depends on the that of the dividend process, in the following sense: if the expected dividend growth under the risk-neutral probability is convex (resp. concave) in D_t, the price is convex (resp. concave) in D_t. In the model of this section, the expected dividend growth under the risk-neutral probability is linear in D_t, which explains the linear property in Eq. (7.10). [In progress, give economic intuition]

7.2.3 Equity premium and interest rate puzzles

> "Average excess returns on the US stock market [the equity premium] is too high to be easily explained by standard asset pricing models." Mehra and Prescott (1985)

Mehra and Prescott (1985) noted the following difficulty with the Lucas model, which gave rise to what is widely known as the *equity premium puzzle*. To be consistent with US data, the equity premium in Eq. (7.7), $\mu_S - r = \gamma\sigma_D^2$, must be an approximate 6% annualized, as explained in the next chapter. If the asset we are trying to price is literally a *consumption claim*, then, σ_D would be consumption volatility, which is very low, approximately 3.3%. For the equity premium to be high, we would need, by reverse-engineering the two equations in (7.7), a quite high value of the relative risk-aversion, say $\gamma \approx \frac{0.06}{0.033^2} \approx 55$. Section 7.4 explains this number, 55, can be improved to 35, once we also condition on the volatility of short-term bonds.

One assumption underlying the previous calculations is that aggregate dividend equals aggregate consumption, which is obviously not the case in the real world. Note, then, that dividend growth volatility is around 6%, which implies the implied γ is $17 \approx \frac{0.06}{0.06^2}$, thereby mitigating the premium puzzle. Still, the model would fail deliver realistic predictions about return volatility, as in this case, return volatility would be just 6%, by Eq. (7.11), which is less than a half of what we see in the data, as explained in the next chapter. Moreover, the model would fail predict countercyclical statistics, such as countercyclical expected returns or dividend yields. We shall return to these topics in the next chapter.

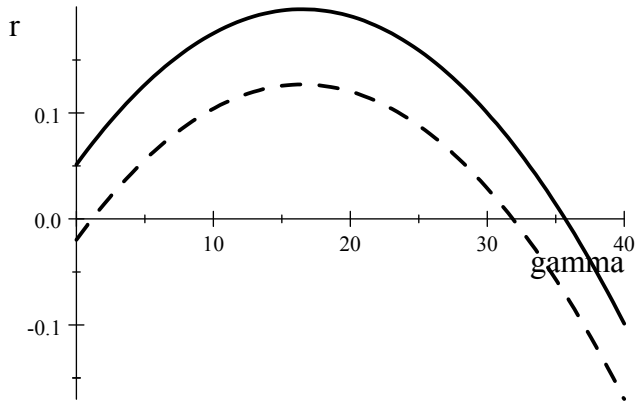
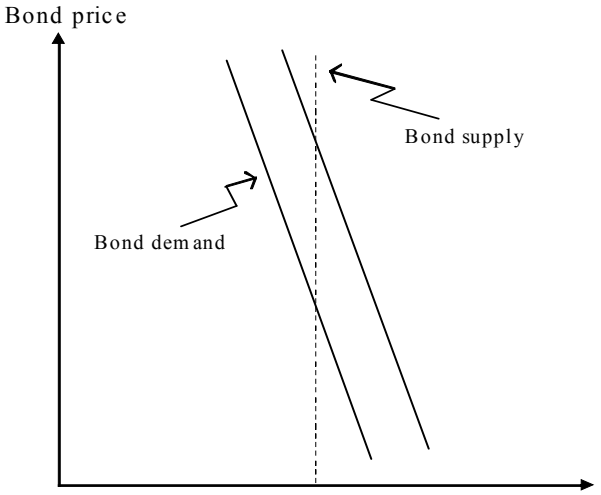


FIGURE 7.1. The risk-free rate puzzle: the two curves depict the graph $\gamma \mapsto r(\gamma) = -\ln\beta + 0.0183\cdot\gamma - \frac{1}{2}(0.0328)^2\cdot\gamma(\gamma+1)$, with $\beta = 0.95$ (solid line) and $\beta = 1.05$ (dashed

line). Even if risk aversion were to be as high as $\gamma = 30$, the equilibrium short-term rate would behave counterfactually, reaching a level as high as 10%. In order for r to be lower when γ is high, it might be required that $\beta > 1$.

The equity premium puzzle is not the only one. Even if we are willing to consider that a CRRA as large as $\gamma = 30$ is plausible, another puzzle arises—an *interest rate puzzle*. As the expression for r in equations (7.5) shows, a large value of γ can lead the interest rate to take very high values, as illustrated by Figure 7.1. Finally, related to the interest rate puzzle is an interest rate volatility puzzle. In the model of this chapter, the safe rate is constant. However, in models where both the equity premium and interest rates change over time, driven by state variables related to, say, preference shocks or market imperfections, the short-term rate is too volatile. For example, in the presence of time-varying expected dividend growth, $\mu_{D,t}$ say, the expression for the short-term rate is the same as in Eq. (7.5), but with $\mu_{D,t}$ replacing the constant μ_D, as explained in the next chapters. It is easily seen, then, that the interest rate is quite volatile for high values of γ.

This interest rate volatility puzzle relates to the assumption of a representative agent. Chapter 3 (Section 3.2.3) explains that agents with low elasticity of intertemporal substitution (EIS) have an inelastic demands for bonds. In the context of CRRA utility functions, a low EIS corresponds to a high CRRA, as EIS $= \frac{1}{\gamma}$, as explained in Chapter 3. So now suppose there is a wide-economy shock that shifts the demand for bonds, as in the following picture—for example, a shock that makes $\mu_{D,t}$ change.



An economy with a representative agent is one where the supply of bonds is fixed. The combination of a representative agent with a low EIS, then, implies a high volatility of the short-term rate, which is counterfactual. To mitigate this issue, one may consider preferences that disentangle the EIS from risk-aversion, as such as those relying on non-expected utility (Epstein and Zin, 1989, 1991; Weil, 1989), or a framework with multiple agents, where bond supply is positively sloped, as in the limited participation model of Guvenen (2009). These models are examined in Chapter 9.

7.3 Hansen-Jagannathan cup

Suppose there are n risky assets. The n asset pricing equations for these assets are,

$$1 = E\left[m_{t+1}\left(1 + R_{j,t+1}\right)\middle| F_t\right], \quad j = 1, \cdots, n.$$

Assuming $R_{j,t+1}$ is stationary, and taking the unconditional expectation of both sides of the previous equation, leaves,

$$\mathbf{1}_n = E\left[m_t\left(\mathbf{1}_n + R_t\right)\right], \quad R_t = \left(R_{1,t}, \cdots, R_{n,t}\right)^\top. \tag{7.12}$$

Next, let $\bar{m} \equiv E(m_t)$, and create a family of stochastic discount factors m_t^*, parametrized by $\bar{m}$, by projecting m on to the asset returns, as follows:

$$\text{Proj}\left(m\middle|\, \mathbf{1}_n + R_t\right) \equiv m_t^*(\bar{m}) = \bar{m} + \underset{1\times n}{\left[R_t - E(R_t)\right]^\top} \underset{n\times 1}{\beta_{\bar{m}}},$$

where[1]

$$\underset{n\times 1}{\beta_{\bar{m}}} \;=\; \underset{n\times n}{\Sigma^{-1}}\, cov\left(m, \underset{n\times 1}{\mathbf{1}_n + R_t}\right) \;= \Sigma^{-1}\left[\mathbf{1}_n - \bar{m}E\left(\mathbf{1}_n + R_t\right)\right],$$

and $\Sigma \equiv E\left[\left(R_t - E(R_t)\right)\left(R_t - E(R_t)\right)^\top\right]$. The Appendix shows that $m_t^*\left(m\right)$ is also a stochastic discounting factor, i.e.,

$$\mathbf{1}_n = E\left[m_t^*(\bar{m}) \cdot \left(\mathbf{1}_n + R_t\right)\right]. \tag{7.13}$$

We have,

$$\sqrt{var\left(m_t^*(\bar{m})\right)} = \sqrt{\beta_{\bar{m}}^\top \Sigma \beta_{\bar{m}}} = \sqrt{\left(\mathbf{1}_n - \bar{m}E\left(\mathbf{1}_n + R_t\right)\right)^\top \Sigma^{-1}\left(\mathbf{1}_n - \bar{m}E\left(\mathbf{1}_n + R_t\right)\right)}. \tag{7.14}$$

Eq. (7.14), provides the expression for the celebrated *Hansen-Jagannathan "cup"*—after the work of Hansen and Jagannathan (1991). It leads to an important tool of analysis, as the following theorem shows.

THEOREM 7.2: *Among all stochastic discount factors with fixed expectation $\bar{m}$, $m_t^*(\bar{m})$ is the one with the smallest variance.*

PROOF: Consider another discount factor indexed by $\bar{m}$, i.e. $m_t(\bar{m})$. Naturally, $m_t(\bar{m})$ satisfies $\mathbf{1}_n = E\left[m_t(\bar{m})\left(\mathbf{1}_n + R_t\right)\right]$. Moreover, by Eq. (7.13),

$$\begin{aligned}
\mathbf{0}_n &= E\left[\left(m_t(\bar{m}) - m_t^*(\bar{m})\right)\left(\mathbf{1}_n + R_t\right)\right] \\
&= E\left[\left(m_t(\bar{m}) - m_t^*(\bar{m})\right)\left(\left(\mathbf{1}_n + E(R_t)\right) + \left(R_t - E(R_t)\right)\right)\right] \\
&= E\left[\left(m_t(\bar{m}) - m_t^*(\bar{m})\right)\left(R_t - E(R_t)\right)\right] \\
&= cov\left[m_t(\bar{m}) - m_t^*(\bar{m}), R_t\right],
\end{aligned} \tag{7.15}$$

where the third line follows because $E\left[m_t(\bar{m})\right] = E\left[m_t^*(\bar{m})\right] = \bar{m}$, and the fourth line holds by $E\left[\left(m_t(\bar{m}) - m_t^*(\bar{m})\right)\right] = 0$. But $m_t^*(\bar{m})$ is a linear combination of R_t. By the previous equation,

[1] We have, $cov\left(m, \mathbf{1}_n + R_t\right) = E\left[m\left(\mathbf{1}_n + R\right)\right] - E\left(m\right)E\left(\mathbf{1}_n + R_t\right) = \mathbf{1}_n - \bar{m}E\left(\mathbf{1}_n + R_t\right)$.

then, it must be that $0 = cov\left[m_t(\bar{m}) - m_t^*(\bar{m}), m_t^*(\bar{m})\right]$. Therefore:

$$\begin{aligned} var\left[m_t(\bar{m})\right] &= var\left[m_t^*(\bar{m}) + m_t(\bar{m}) - m_t^*(\bar{m})\right] \\ &= var\left[m_t^*(\bar{m})\right] + var\left[m_t(\bar{m}) - m_t^*(\bar{m})\right] + 2 \cdot cov\left[m_t(\bar{m}) - m_t^*(\bar{m}), m_t^*(\bar{m})\right] \\ &= var\left[m_t^*(\bar{m})\right] + var\left[m_t(\bar{m}) - m_t^*(\bar{m})\right] \\ &\geq var\left[m_t^*(\bar{m})\right]. \end{aligned}$$

‖

Hansen and Jagannathan (1991) consider an extension of this result, in which the stochastic discount factor satisfies the non-negativity constraint, $m > 0$.

Consider, then, the curve, $\bar{m} \longmapsto var\left[m_t^*(\bar{m})\right]$—the cup. By Theorem 7.1, each pair $(\bar{m}, var\left[m_t(\bar{m})\right])$ predicted by any candidate model satisfying Eq. (7.12) has to lie above the cup for each possible $\bar{m}$, $(\bar{m}, var\left[m_t^*(\bar{m})\right])$. The idea is the following. Testing the validity of a model is tantamount to assess whether the Euler equations it predicts are satisfied *and* how volatile its pricing kernel is. Consider, then, Eq. (7.12), and assume that a candidate pricing kernel is parametrized by some vector $\theta \in \Theta$, such that $m_t(\bar{m};\theta) \equiv m_t(\bar{m})$ say, and is volatile enough. We want to find values of θ, say $\hat{\theta}$, such that $m_t(\bar{m};\hat{\theta})$ is a valid stochastic factor, i.e. such that it prices all the assets, just as in Eq. (7.12) and is volatile enough. Requiring it prices all the assets is actually a condition needed for Theorem 7.2 to hold—in particular, the proof shows that in this case Eq. (9.44) would hold. Naturally, the pricing kernel needs to be volatile enough as well. The issue then is to ascertain whether the values of $\hat{\theta}$ that let the pricing kernel in the cup are economically reasonable so to speak.

To illustrate, consider the Lucas' model of Section 7.2. We calculate the Hansen-Jagannathan bounds in Eq. (7.14) and ascertain whether there are parameter values of the model that would allow the stochastic discount factor to be inside the bounds in the mean and standard deviation space. The stochastic discount factor of the model is:

$$m_{t+1} = \frac{\xi_{t+1}}{\xi_t} = \exp\left(Z_{t+1}\right), \quad Z_{t+1} = -\left(r + \frac{1}{2}\lambda^2\right) - \lambda u_{D,t+1}, \quad u_{D,t+1} \equiv \frac{\epsilon_{D,t+1}}{\sigma_D}.$$

By Lemma 7.1, the first two moments of m_t are:

$$\bar{m} = E(m_t) = e^{-r} \quad \text{and} \quad \bar{\sigma}_m = \sqrt{var\left(m_t(\bar{m})\right)} = e^{-r+\frac{1}{2}\lambda^2}\sqrt{1 - e^{-\lambda^2}}, \tag{7.16}$$

where r is as in Eq. (7.5), and $\lambda = \gamma\sigma_D$, as usual. For given μ_D and σ_D^2, these two equations in (7.16) form a γ-parametrized curve in the space $(\bar{m}\text{-}\bar{\sigma}_m)$. The issue is to check whether this curve enters the Hansen-Jagannathan cup for plausible values of γ. It is not the case. Rather, we have the situation depicted in Figure 7.2. The reason the circles bend back is easily explained. When γ is low, an increase in γ leads to an increase in $\bar{\sigma}_m$ and a decrease in $\bar{m}$, because r increases with γ when γ is small. Yet as soon as γ gets sufficiently large, the interest rate decreases, due to precautionary motives, which make both $\bar{m}$ and $\bar{\sigma}_m$ increase. In general, we expect that for any model where the intertemporal elasticity of substitution is somehow confounded with relative risk-aversion, a bending pattern such that in Figure 7.2 would arise, being indicative of conflicts between intertemporal elasticity of substitution and precautionary effects relating to the demand of riskless assets.

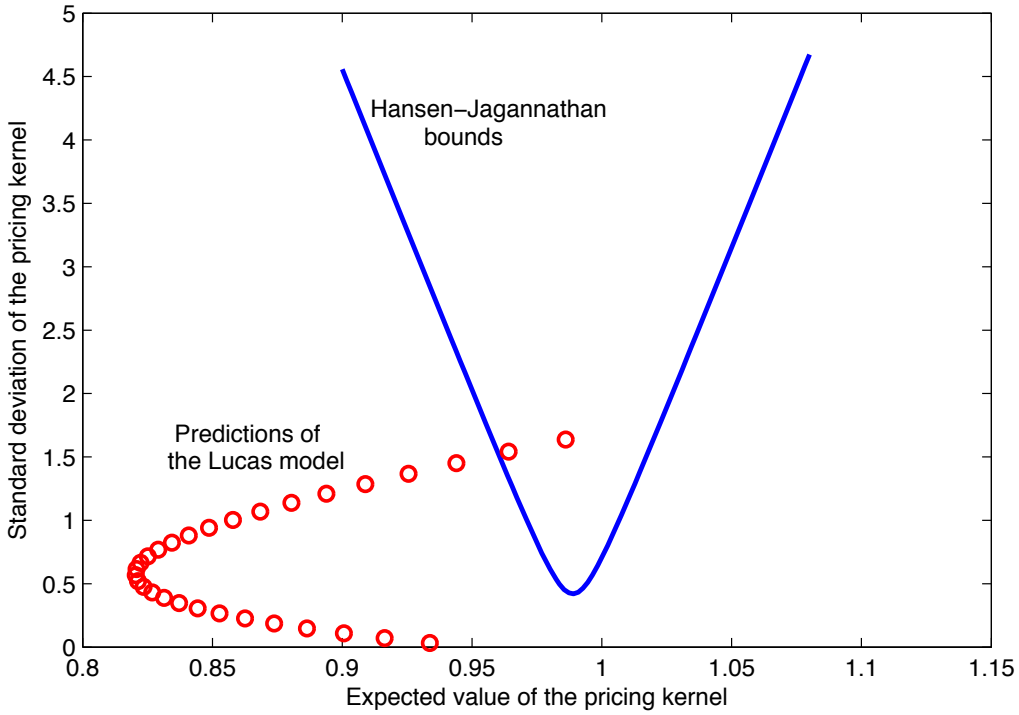


FIGURE 7.2. The solid line depicts the Hansen-Jagannathan bounds, obtained through Eq. (7.14), through aggregate stock market data and the short-term rate. The average return and standard deviation of the stock market are taken to be 0.07 and 0.14. The average short-term rate (three-month bill) and its volatility are, instead, 0.01 and 0.02. These estimates relate to the sample period from January 1948 to December 2002. The circles are predictions of the Lucas model in Eq. (7.16), with $\beta = 0.95$, $\mu_D = 0.0183$, $\sigma_D = 0.0328$ and γ ranging from 1 to 35. The two circles inside the cup are the pairs $(\bar{m}, \bar{\sigma}_m)$ in Eq. (7.16) obtained with $\gamma = 35$ and 33. Progressively lower values of γ lead the pairs $(\bar{m}, \bar{\sigma}_m)$ to lie outside the cup, nonlinearly.

The Lucas model predicts that the pricing kernel is quite moderately volatile. The following chapters discuss models with both heterogeneous agents or more general preferences, which can help boost the volatility of the pricing kernel.

7.4 Multifactor extensions

A natural way to increase the variance of the pricing kernel is to increase the number of factors. We consider two possibilities: one in which returns are normally distributed, and one in which returns are lognormally distributed. A multifactor model is naturally promising, for its potential to deliver boosted values for the risk-premium. However, it may fail deliver time-varying statistics as we shall explain.

7.4.1 Exponential affine pricing kernels

This section considers a variant of the Lucas model of Section 7.2, and determines expected returns under a different assumption regarding the returns distributions. We still maintain the hypothesis that the stochastic discount factor is exponential-Gaussian, viz

$$m_{t+1} = \exp\left(Z_{t+1}\right), \quad Z_{t+1} = -\left(r + \frac{1}{2}\lambda^2\right) - \lambda u_{D,t+1}, \quad u_{D,t+1} \sim \text{NID}\left(0,1\right),$$

where r and λ are some constants. In the absence of arbitrage,

$$1 = E(m_{t+1} \cdot \tilde{R}_{t+1}) = E\left(m_{t+1}\right) E(\tilde{R}_{t+1}) + cov(m_{t+1}, \tilde{R}_{t+1}), \quad \tilde{R}_{t+1} \equiv \frac{S_{t+1} + D_{t+1}}{S_t}.$$

By rearranging terms,[2] and using the fact that $E\left(m_{t+1}\right) = R^{-1}$,

$$E(\tilde{R}_{t+1}) - R = -R \cdot cov(m_{t+1}, \tilde{R}_{t+1}). \tag{7.17}$$

Consider the following result, which we shall use later:

LEMMA 7.3 (Stein's lemma): *Suppose that two random variables x and y are jointly normal. Then,*

$$cov\left[g\left(x\right), y\right] = E\left[g'\left(x\right)\right] \cdot cov\left(x, y\right),$$

for any function $g : E\left(\left|g'\left(x\right)\right|\right) < \infty$.

Next, suppose $\tilde{R}$ is normally distributed. This assumption is inconsistent with those underlying the model in Section 7.2, in which $\tilde{R}$ is *log*normally distributed, in equilibrium, being equal to $\ln \tilde{R} = \mu_D - \frac{1}{2}\sigma_D^2 + \epsilon_S$, where ϵ_S is normal. Howver, let us explore the asset pricing implications of this new assumption $\tilde{R}$. Because $\tilde{R}_{t+1}$ and Z_{t+1} are both normal, and $m_{t+1} = m\left(Z_{t+1}\right) = \exp\left(Z_{t+1}\right)$, we apply Lemma 7.3 and obtain,

$$cov(m_{t+1}, \tilde{R}_{t+1}) = E\left[m'\left(Z_{t+1}\right)\right] \cdot cov(Z_{t+1}, \tilde{R}_{t+1}) = -\lambda R^{-1} \cdot cov(u_{D,t+1}, \tilde{R}_{t+1}).$$

Replacing this expression for the covariance, $cov(m_{t+1}, \tilde{R}_{t+1})$, into Eq. (7.17), leaves:

$$E(\tilde{R}_{t+1}) - R = \lambda \cdot cov(u_{D,t+1}, \tilde{R}_{t+1}).$$

We now extend the previous observations to a more general setup. Consider a stochastic discount factor as a function of K factors, $m\left(\epsilon_{1t}, \cdots, \epsilon_{Kt}\right)$ say. A particularly convenient analytical assumption is that m is exponential-affine and the factors $\left(\epsilon_{i,t}\right)_{i=1}^{K}$ normal, as in the following definition:

DEFINITION 7.4 (EAPK: Exponential Affine Pricing Kernel): *Let,*

$$Z_t \equiv \phi_0 + \sum_{i=1}^{K} \phi_i \epsilon_{i,t}.$$

[2] With a portfolio return that is perfectly correlated with m, we have:

$$E_t(\tilde{R}_{t+1}^M) - \frac{1}{E_t\left(m_{t+1}\right)} = -\frac{\sigma_t\left(m_{t+1}\right)}{E_t\left(m_{t+1}\right)} \sigma_t(\tilde{R}_{t+1}^M).$$

In more general setups than the ones considered in this introductory example, both $\frac{\sigma_t\left(m_{t+1}\right)}{E_t\left(m_{t+1}\right)}$ and $\sigma_t(\tilde{R}_{t+1}^M)$ should be time-varying.

A EAPK *is a function,*

$$m_t = m(Z_t) = \exp(Z_t). \tag{7.18}$$

If $(\epsilon_{i,t})_{i=1}^K$ *are jointly normal, and each* $\epsilon_{i,t}$ *has mean zero and variance* σ_i^2, $i = 1, \cdots, K$, *the EAPK is called a* Normal EAPK (NEAPK).

It is without loss of generality to assume that each $\epsilon_{i,t}$ has a mean equal to zero in the previous definition provided that $\phi_0 \neq 0$. Next, suppose $\tilde{R}$ is normally distributed. By Lemma 7.3 and the NEAPK assumption,

$$cov(m_{t+1}, \tilde{R}_{t+1}) = cov[\exp(Z_{t+1}), \tilde{R}_{t+1}] = R^{-1} cov(Z_{t+1}, \tilde{R}_{t+1}) = R^{-1} \sum_{i=1}^{K} \phi_i cov(\epsilon_{i,t+1}, \tilde{R}_{t+1}).$$

By replacing this into Eq. (7.17) leaves the linear factor representation,

$$E(\tilde{R}_{t+1}) - R = -\sum_{i=1}^{K} \phi_i \underbrace{cov(\epsilon_{i,t+1}, \tilde{R}_{t+1})}_{\text{"betas"}}. \tag{7.19}$$

We have shown the following result:

Proposition 7.5: *Suppose that* $\tilde{R}$ *is normally distributed. Then, NEAPK* $\Longrightarrow$ *linear factor representation for asset returns.*

The APT representation in Eq. (7.19), is similar to a result in Cochrane (1996).[3] Cochrane (1996) assumes that m is affine, i.e. $m(Z_t) = Z_t$ where Z_t is as in Definition 6.1. This assumption implies that $cov(m_{t+1}, \tilde{R}_{t+1}) = \sum_{i=1}^{K} \phi_i cov(\epsilon_{i,t+1}, \tilde{R}_{t+1})$. By replacing this expression for the covariance, $cov(m_{t+1}, \tilde{R}_{t+1})$, into Eq. (7.17), leaves

$$E(\tilde{R}_{t+1}) - R = -R \sum_{i=1}^{K} \phi_i cov(\epsilon_{i,t+1}, \tilde{R}_{t+1}), \quad \text{where } R = \frac{1}{E(m)} = \frac{1}{\phi_0}.$$

The NEAPK assumption, compared to Cochrane's, carries the obvious advantage to guarantee the stochastic discount factor is strictly positive—a theoretical condition we need to rule out arbitrage opportunities.

7.4.2 Lognormal returns

Next, assume that $\tilde{R}$ is *log*normally distributed, and that the NEAPK holds. We have,

$$1 = E\left(m_{t+1} \cdot \tilde{R}_{t+1}\right) \Longleftrightarrow e^{-\phi_0} = E\left(e^{\sum_{i=1}^{K} \phi_i \epsilon_{i,t+1}} \cdot \tilde{R}_{t+1}\right). \tag{7.20}$$

Consider, first, the case $K = 1$, and let $y_t = \ln \tilde{R}_t$ be normally distributed. The previous equation can be written as,

$$e^{-\phi_0} = E\left[e^{\phi_1 \epsilon_{t+1} + y_{t+1}}\right] = e^{E(y_{t+1}) + \frac{1}{2}(\phi_1^2 \sigma_\epsilon^2 + \sigma_y^2 + 2\phi_1 \sigma_{\epsilon y})}.$$

[3] To recall why Eq. (7.19) is indeed a APT equation, suppose that $\tilde{R}$ is a n-(column) vector of returns and that $\tilde{R} = a + bf$, where f is K-(column) vector *with zero mean and unit variance* and a, b are some given vector and matrix with appropriate dimension. Then clearly, $b = cov(\tilde{R}, f)$. A portfolio π delivers $\pi^\top \tilde{R} = \pi^\top a + \pi^\top cov(\tilde{R}, f) f$. Arbitrage opportunity is: $\exists \pi : \pi^\top cov(\tilde{R}, f) = 0$ and $\pi^\top a \neq r$. To rule that out, we may show as in Part I of these *Lectures* that there must exist a K-(column) vector λ s.t. $a = cov(\tilde{R}, f)\lambda + r$. This implies $\tilde{R} = a + bf = r + cov(\tilde{R}, f)\lambda + bf$. That is, $E(\tilde{R}) = r + cov(\tilde{R}, f)\lambda$.

This is,

$$E\left(y_{t+1}\right)=-\left[\phi_0+\frac{1}{2}(\phi_1^2\sigma_\epsilon^2+\sigma_y^2+2\phi_1\sigma_{\epsilon y})\right].$$

By applying the pricing equation (7.20) to a zero coupon bond,

$$e^{-\phi_0}=E\left(e^{\phi_1\epsilon_{t+1}}\right)e^{\ln R_{t+1}}=e^{\ln R_{t+1}+\frac{1}{2}\phi_1^2\sigma_\epsilon^2},$$

which we can solve for R_{t+1}:

$$\ln R_{t+1}=-\left(\phi_0+\frac{1}{2}\phi_1^2\sigma_\epsilon^2\right).$$

The expected excess return is,

$$E\left(y_{t+1}\right)-\ln R_{t+1}+\frac{1}{2}\sigma_y^2=-\phi_1\sigma_{\epsilon y}. \tag{7.21}$$

Eq. (7.21) shows that the theory in Section 7.2 through a different angle. Apart from Jensen's inequality effects ($\frac{1}{2}\sigma_y^2$), this is indeed the Lucas model of Section 7.2 once $\phi_1=-\gamma$. This model has poor quantitative implications as discussed in Section 7.3, bound as it is to explain returns with only one "stochastic discount-factor parameter," i.e. ϕ_1.

Next consider the general case. Assume as usual that dividends are as in (7.1). To find the price function in terms of the state variable ϵ, we may proceed as in Section 7.2. In the absence of bubbles,

$$S_t=\sum_{i=1}^{\infty}E\left[\frac{\xi_{t+i}}{\xi_t}\cdot D_{t+i}\right]=D_t\cdot\sum_{i=1}^{\infty}e^{\left(\mu_D+\phi_0+\frac{1}{2}\sum_{i=1}^{K}\phi_i\left(\phi_i\sigma_i^2+2\sigma_{i,D}\right)\right)\cdot i},\quad \sigma_{i,D}\equiv cov\left(\epsilon_i,\epsilon_D\right).$$

Thus, if

$$\hat{k}\equiv\mu_D+\phi_0+\frac{1}{2}\sum_{i=1}^{K}\phi_i\left(\phi_i\sigma_i^2+2\sigma_{i,D}\right)<0,$$

then,

$$\frac{S_t}{D_t}=\frac{\hat{k}}{1-\hat{k}}.$$

There are interesting features of the model to mention. The model is cast within a multi-factor setting, and yet it predicts that the price-dividend ratio is constant, which is counterfactual, as explained in the next chapter. However, note the following facts. The first two moments of the stochastic discount factor of this model are easily determined, by relying on Lemma 7.1:

$$E\left(m_t\right)=e^{E(Z_t)+\frac{1}{2}var(Z_t)},\quad \sqrt{var\left(m_t\right)}=e^{E(Z_t)+var(Z_t)}\sqrt{1-e^{-var(Z_t)}}.$$

Therefore, we can always calibrate the parameters of this model and make sure the first two moments of the pricing kernel enter into the Hansen-Jagannathan cup. At the same time, remember, the model still predicts price-dividend ratios to be constant—that is, the model makes counterfactual predictions (constant price-dividend ratios) even when the variance of its pricing kernel is arbitrarily large.

In other words, a model satisfying Hansen-Jagannathan bounds is not definitely "good." It would rather necessitate further scrutiny. The next section illustrates a variant of the previous model in which time-variation in the risk-premiums gives rise to time varying statistics. The next chapter makes additional steps forwards, attempting to develop theoretical test conditions, which ensure that these time-varying statistics have the same cyclical properties as in the data, such as procyclicality of price-dividend ratios or countercyclical volatility.

7.5 Conditional CAPM

[In progress]

7.6 Pricing kernels and Sharpe ratios

[This section needs a considerable revision]

7.6.1 Market portfolios and pricing kernels

The market portfolio cannot be perfectly correlated with the stochastic discount factor, in general [cite some literature such as Cecchetti, Lam, and Mark (1994)]. Let $r^e_{i,t+1} = \tilde{R}_{i,t+1} - R_{t+1}$ be the excess return on a risky asset. We have:

$$0 = E_t\left(m_{t+1} r^e_{i,t+1}\right) = E_t\left(m_{t+1}\right) E_t\left(r^e_{i,t+1}\right) + \rho_{i,t} \cdot \mathrm{Std}_t\left(m_{t+1}\right) \cdot \mathrm{Std}_t\left(r^e_{i,t+1}\right),$$

where $\mathrm{Std}_t\left(u_{t+1}\right)$ denotes the standard deviation of a variable u_{t+1}, conditionally upon the information available at time t, and $\rho_{i,t} \equiv \mathrm{corr}_t(m_{t+1}, r^e_{i,t+1})$, a conditional correlation. Hence, the Sharpe Ratio, $\mathcal{S} \equiv \frac{E_t\left(r^e_{i,t+1}\right)}{\mathrm{Std}_t\left(r^e_{i,t+1}\right)}$, satisfies:

$$|\mathcal{S}| \leq \frac{\mathrm{Std}_t\left(m_{t+1}\right)}{E_t\left(m_{t+1}\right)} = \mathrm{Std}_t\left(m_{t+1}\right) \cdot R_{t+1}, \tag{7.22}$$

The highest possible Sharpe ratio is bounded. The equality holds for a hypothetical portfolio M, say, yielding excess returns perfectly conditionally negatively correlated with the stochastic discount factor, $\rho_{M,t} = -1$. We shall say of M that it is a *β-CAPM generating portfolio*. Is it also a *market portfolio*? After all, a feasible and attainable portfolio lying on the volatility bounds for the stochastic discount factor is clearly mean-variance efficient. The answer is subtle. As explained in the context of the static model of Chapter 1, the Sharpe ratio, $\mathcal{S}$, equals the slope of the Capital Market Line, and bears the interpretation of unit market risk-premium. If $\rho_{M,t} = -1$, then, by Eq. (7.22), the slope of the Capital Market Line reduces to $\frac{\mathrm{Std}_t(m_{t+1})}{E_t(m_{t+1})}$. For example, with the Lucas model in Section 7.2,

$$\frac{\mathrm{Std}_t\left(m_{t+1}\right)}{E_t\left(m_{t+1}\right)} = \sqrt{e^{\gamma^2\sigma_D^2} - 1} \approx \gamma\sigma_D.$$

In Section 7.2, we also explained that $\left(\mu_S - r\right)/\sigma_D = \gamma\sigma_D$, which is only approximately true, according to the previous relation. Not even a simple model with a *single* tree, such as that in Section 7.2, would be capable of leading to a β-CAPM generating portfolio, or a market portfolio! Indeed, for this model, we have that $E(\tilde{R}) = e^{\mu_S}$, $R = e^{-\ln\beta + \gamma\left(\mu_D - \frac{1}{2}\sigma_D^2\right) - \frac{1}{2}\gamma^2\sigma_D^2}$, and $var(\tilde{R}) = e^{2\mu_S}(e^{\sigma_D^2} - 1)$, with a Sharpe ratio equal to:

$$\mathcal{S} = \frac{1 - e^{-\gamma\sigma_D^2}}{\sqrt{e^{\sigma_D^2} - 1}}.$$

By simple computations, $\rho = -\frac{1 - e^{-\gamma\sigma_D^2}}{\sqrt{e^{\gamma^2\sigma_D^2} - 1}\sqrt{e^{\sigma_D^2} - 1}}$, which is not precisely "-1"—only approximately equal to -1, for low values of σ_D.

A further complication arises, a β-CAPM generating portfolio is not necessarily the tangency portfolio. We can show that there is another portfolio leading to the very same β-pricing relation predicted by the tangency portfolio. Such a portfolio is referred to as the *maximum correlation portfolio*, for reasons developed below. Let $\bar{R} = \frac{1}{E(m)}$. By the CCAPM in Chapter 2,

$$E\left(R^i\right) - \bar{R} = \frac{\beta_{R^i,m}}{\beta_{R_p,m}}\left(E\left(R_p\right) - \bar{R}\right),$$

where R_p is a portfolio return. Next, let $R_p = R^m \equiv \frac{m}{E(m^2)}$, which is clearly perfectly correlated with the stochastic discount factor. By results in Chapter 2,

$$E\left(R^i\right) - \bar{R} = \beta_{R^i,R^m}\left(E\left(R^m\right) - \bar{R}\right).$$

This is not yet the β-representation of the CAPM, because we have yet to show that there is a way to construct R^m as a portfolio return. In fact, there is a natural choice: pick $m = m^*$, where m^* is the minimum-variance kernel leading to the Hansen-Jagannathan bounds. Since m^* is linear in all asset retuns, R^{m^*} can be thought of as a return that can be obtained by investing in all assets. Furthermore, in the appendix we show that R^{m^*} satisfies,

$$1 = E\left(m \cdot R^{m^*}\right).$$

Where is this portfolio located? The Appendix shows that there is no portfolio yielding the same expected return with lower variance, that is, R^{m^*} is mean-variance efficient), and that:

$$E\left(R^{m^*}\right) - 1 = \frac{r - Sh}{1 + Sh} = r - \frac{1 + r}{1 + Sh}Sh < r.$$

Mean-variance efficiency of R^{m^*} and the previous inequality imply that this portfolio lies in the *lower branch* of the mean-variance efficient portfolios. And this is so because this portfolio is *positively* correlated with the *true* pricing kernel. Naturally, the fact that this portfolio is β-CAPM generating doesn't necessarily imply that it is also perfectly correlated with the true stochastic discount factor. As shown in the appendix, R^{m^*} has only the maximum possible correlation with all possible m. Perfect correlation occurs exactly in correspondence of the stochastic discount factor $m = m^*$ (i.e. when the economy exhibits a stochastic discount factor exactly equal to m^*).

PROOF THAT R^{m^*} IS β-CAPM GENERATING. The relations, $1 = E(m^* R^i)$ and $1 = E(m^* R^{m^*})$, imply:

$$E(R^i) - R = -R \cdot cov\left(m^*, R^i\right), \quad E(R^{m^*}) - R = -R \cdot cov\left(m^*, R^{m^*}\right),$$

or,

$$\frac{E(R^i) - R}{E(R^{m^*}) - R} = \frac{cov\left(m^*, R^i\right)}{cov\left(m^*, R^{m^*}\right)}.$$

By construction, R^{m^*} is perfectly correlated with m^*. Precisely, $R^{m^*} = m^* / E(m^{*2}) \equiv \zeta^{-1} m^*$, $\zeta \equiv E(m^{*2})$. Therefore,

$$\frac{cov\left(m^*, R^i\right)}{cov\left(m^*, R^{m^*}\right)} = \frac{cov\left(\zeta R^{m^*}, R^i\right)}{cov\left(\zeta R^{m^*}, R^{m^*}\right)} = \frac{cov\left(R^{m^*}, R^i\right)}{var\left(R^{m^*}\right)} = \beta_{R^i,R^{m^*}}.$$

‖

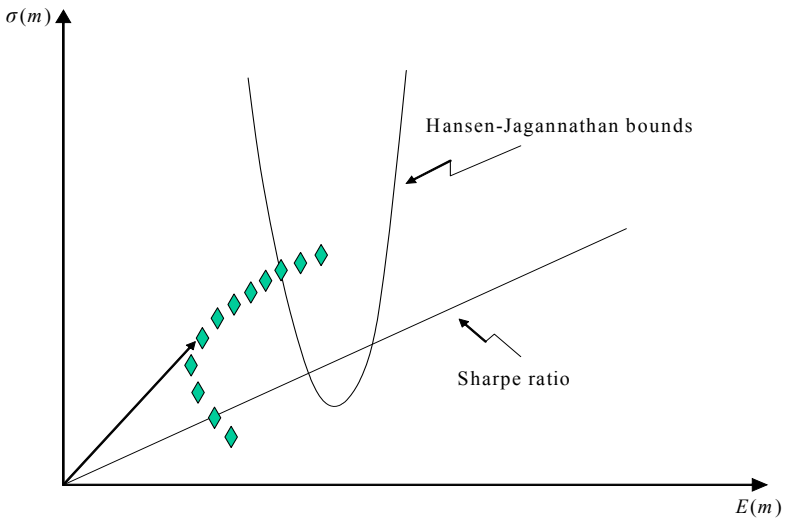
7.6.2 Pricing kernel bounds

Figure 7.2 depicts the typical situation the neoclassical asset pricing model has to face. Points ♦ are those generated by the Lucas model for various values of γ. The model has to be such that points ♦ lie *above* the observed Sharpe ratio ($\sigma(m)/E(m) \geq$ greatest Sharpe ratio ever observed in the data—Sharpe ratio on the market portfolio) *and* inside the Hansen-Jagannathan bounds. Typically, we need high values of γ to enter the Hansen-Jagannathan bounds.

There is an interesting connection between these facts and the classical mean-variance portfolio frontier described in Chapter 1. As shown in Figure 7.3, every asset or portfolio must lie inside the region bounded by two straight lines with slopes $\mp \sigma(m)/E(m)$. It must be so, as for any asset (or portfolio) priced by a stochastic discount factor m, we have that

$$\left|E(R^i) - R\right| \leq \frac{\sigma(m)}{E(m)} \cdot \sigma\left(R^i\right).$$

As seen in the previous section, the equality is only achieved by asset (or portfolio) returns that are perfectly correlated with m. A tangency portfolio such as T doesn't necessarily attain the volatility bounds for the stochastic discount factor. Moreover, the market portfolio has no reasons to lie on the volatility bound for the stochastic discount factor. As an example, for the simple Lucas model, the (only existing) asset has a Sharpe ratio, which doesn't lie on the volatility bounds for the stochastic discount factor. In a sense, the CCAPM does not need to imply the CAPM: there are necessarily no assets performing, at the same time, as market portfolios and β-CAPM generating, which are also priced consistently by the true stochastic discount factor. These conditions would only simultaneously hold if the candidate market portfolio were perfectly negatively correlated with the stochastic discount factor, which is a quite specific circumstance, the only circumstance where we can really say the CAPM is a particular case of the CCAPM. We still do not know conditions on general families of stochastic discount factors, which are consistent with the previous properties.



However, we know that there exists another portfolio, the maximum correlation portfolio, which is also β-CAPM generating. In other terms, if $\exists R_* : R_* = -\zeta m$, for some positive constant ζ, then the β-CAPM representation holds, but this doesn't necessarily mean that R_*

is also a market portfolio. More generally, if there is a return R_* that is β-CAPM generating, then,

$$\rho_{i,R^*} = \frac{\rho_{i,m}}{\rho_{R^*,m}}, \text{ all } i. \tag{7.23}$$

Therefore, we don't need an asset or portfolio return that is *perfectly* correlated with m to make the CCAPM collapse to the CAPM. All in all, the existence of an asset return that is perfectly negatively correlated with the stochastic discount factor is a sufficient condition for the CCAPM to collapse to the CAPM, not a necessary condition. The proof of Eq. (7.23) is simple. By the CCAPM,

$$E(R^i) - R = -\rho_{i,m}\frac{\sigma(m)}{E(m)}\sigma(R^i), \quad \text{and} \quad E(R^*) - R = -\rho_{R^*,m}\frac{\sigma(m)}{E(m)}\sigma(R^*).$$

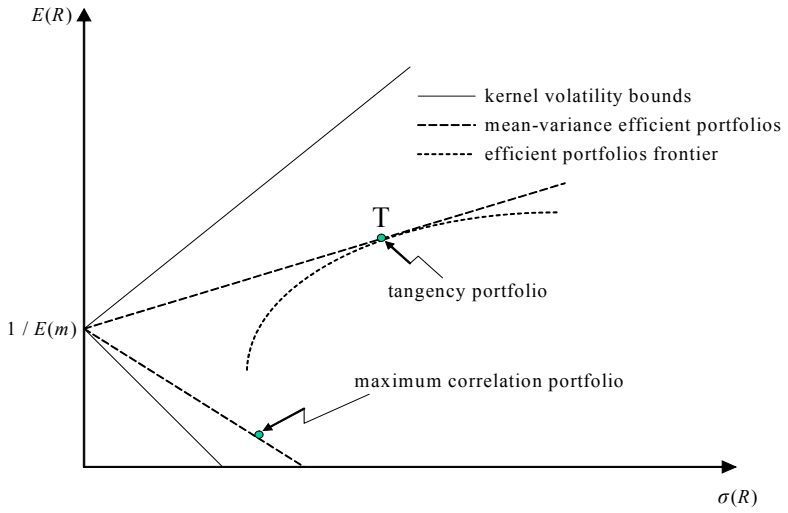
That is,

$$\frac{E(R^i) - R}{E(R^*) - R} = \frac{\rho_{i,m}}{\rho_{R^*,m}}\frac{\sigma(R^i)}{\sigma(R^*)} \tag{7.24}$$

But if R^* is β-CAPM generating,

$$\frac{E(R^i) - R}{E(R^*) - R} = \frac{cov(R^i, R^*)}{\sigma(R^*)^2} = \rho_{i,R^*}\frac{\sigma(R^i)}{\sigma(R^*)}. \tag{7.25}$$

Comparing Eq. (7.24) with Eq. (7.25) produces Eq. (7.23).



A final thought. In many pieces of applied research, we often read that because we observe time-varying Sharpe ratios on (proxies of) the market portfolio, we should also model the market risk-premium $\pi_t \equiv \sqrt{Var_t\left(m_{t+1}\right)} \Big/ E_t\left(m_{t+1}\right)$ as time-varying. While Chapter 8 explains that the evidence for time-varying risk-premiums is overwhelming, a criticism to this motivation is that π_t is only an upper bound to the Sharpe ratio of the market portfolio. On a strictly theoretical point of view, then, a time-varying π_t is neither a necessary or a sufficient condition to have time-varying Sharpe ratios, as Figure 7.3 illustrates.

7.7 Conditioning bounds

The Hansen-Jagannathan bounds in Eq. (7.14) can be improved by using conditioning information, as originally shown by Gallant, Hansen and Tauchen (1990) and in Ferson and Siegel (2003). A difficulty with these bounds is that they may diplay a finite sample bias, in that they tend to overstate the true bounds and thus reject too often a given model. Finite sample corrections are considered by Ferson and Siegel (2003). [Discuss, analytically]

Alvarez and Jermann (2005). [Discuss, analytically]

7.8 Survivorship bias?

[In progress]

7.9 Appendix

Proof of Eq. (7.13). We have,

$$\begin{aligned}
E\left[m_t^*(\bar{m})\cdot(\mathbf{1}_n+R_t)\right] &= E\left[\left(\bar{m}+(R_t-E(R_t))^\top\beta_{\bar{m}}\right)(\mathbf{1}_n+R_t)\right]\\
&= \bar{m}E\left(\mathbf{1}_n+R_t\right)+E\left[(R_t-E(R_t))^\top\beta_{\bar{m}}\left(\mathbf{1}_n+R_t\right)\right]\\
&= \bar{m}E\left(\mathbf{1}_n+R_t\right)+E\left[\left(\mathbf{1}_n+R_t\right)(R_t-E(R_t))^\top\right]\beta_{\bar{m}}\\
&= \bar{m}E\left(\mathbf{1}_n+R_t\right)+E\left[\left((\mathbf{1}_n+E(R_t))+(R_t-E(R_t))\right)(R_t-E(R_t))^\top\right]\beta_{\bar{m}}\\
&= \bar{m}E\left(\mathbf{1}_n+R_t\right)+E\left[(R_t-E(R_t))(R_t-E(R_t))^\top\right]\beta_{\bar{m}}\\
&= \bar{m}E\left(\mathbf{1}_n+R_t\right)+\Sigma\beta_{\bar{m}}\\
&= \bar{m}E\left(\mathbf{1}_n+R_t\right)+\mathbf{1}_n-\bar{m}E\left(\mathbf{1}_n+R_t\right),
\end{aligned}$$

where the last line follows by the definition of $\beta_{\bar{m}}$.

Proof that R^{m^*} can be generated by a feasible portfolio

Proof of the Equation, $1=E\left(m\cdot R^{m^*}\right)$. We have,

$$E(m\cdot R^{m^*})=\frac{E\left(m\cdot m^*\right)}{E\left[(m^*)^2\right]},$$

where

$$\begin{aligned}
E\left(m\cdot m^*\right) &= \bar{m}^2+E\left[m\left(R_t-E(R_t)\right)^\top\beta_{\bar{m}}\right]\\
&= \bar{m}^2+E\left[m\left(1+R_t\right)^\top\right]\beta_{\bar{m}}-E\left[m\left(1+E(R_t)\right)^\top\right]\beta_{\bar{m}}\\
&= \bar{m}^2+\beta_{\bar{m}}-E\left(m\right)\left[1+E(R_t)\right]^\top\beta_{\bar{m}}\\
&= \bar{m}^2+\left[\mathbf{1}_n-\bar{m}\left(1+E(R_t)\right)^\top\right]\beta_{\bar{m}}\\
&= \bar{m}^2+\left[\mathbf{1}_n-\bar{m}\left(1+E(R_t)\right)^\top\right]\Sigma^{-1}\left[\mathbf{1}_n-\bar{m}\left(\mathbf{1}_n+E(R_t)\right)\right]\\
&= \bar{m}^2+var\left(m^*\right),
\end{aligned}$$

where the last line is due to the definition of m^*.

Proof that R^{m^*} is mean-variance efficient. Let $p=(p_0,p_1,\cdots,p_n)^\top$ the vector of $n+1$ portfolio weights (here $p_i\equiv\pi^i/w$ is the portfolio weight of asset i, $i=0,1,\cdots,n$. We have,

$$p^\top\mathbf{1}_{n+1}=1.$$

The returns we consider are $\underline{r}_t=\left(\bar{m}^{-1}-1,\ r_{1,t},\cdots,r_{n,t}\right)^\top$. We denote our "benchmark" portfolio return as $r_{bt}=r^{m^*}-1$. Next, we build up an arbitrary portfolio yielding the same expected return $E(r_{bt})$ and then we show that this has a variance greater than the variance of r_{bt}. Since this portfolio

is arbitrary, the proof will be complete. Let $r_{pt} = p^\top \underline{r}_t$ such that $E(r_{pt}) = E(r_{bt})$. We have:

$$\begin{aligned} cov\left(r_{bt}, r_{pt} - r_{bt}\right) &= E[r_{bt} \cdot (r_{pt} - r_{bt})] \\ &= E[R_{bt} \cdot (R_{pt} - R_{bt})] \\ &= E\left(R_{bt} \cdot R_{pt}\right) - E\left(R_{bt}^2\right) \\ &= \frac{E[m^* \left(1 + p^\top \underline{r}_t\right)]}{E\left(m^{*2}\right)} - \frac{E\left(m^{*2}\right)}{[E\left(m^{*2}\right)]^2} \\ &= \frac{1}{E\left(m^{*2}\right)}[p^\top E\left(m^* \left(\mathbf{1}_{n+1} + \underline{r}_t\right)\right) - 1] \\ &= 0. \end{aligned}$$

The first line follows by construction since $E(r_{pt}) = E(r_{bt})$. The last line follows because

$$p^\top E\left[m^* \left(\mathbf{1}_{n+1} + \underline{r}_t\right)\right] = p^\top \mathbf{1}_{n+1} = 1.$$

Given this, the claim follows directly from the fact that

$$var\left(R_{pt}\right) = var\left[R_{bt} + \left(R_{pt} - R_{bt}\right)\right] = var\left(R_{bt}\right) + var\left(R_{pt} - R_{bt}\right) \geq var\left(R_{bt}\right).$$

Proof of the Equation, $E\left(R^{m^*}\right) - 1 = r - \frac{1+r}{1+Sh} Sh$. We have,

$$E(R^{m^*}) - 1 = \frac{\bar{m}}{E[(m^*)^2]} - 1.$$

In terms of the notation introduced in Section 7.8, m^* is:

$$m^* = \bar{m} + (a\epsilon)^\top \beta_{\bar{m}}, \;\; \beta_{\bar{m}} = \sigma^{-1}\left[\mathbf{1}_n - \bar{m}\left(\mathbf{1}_n + b\right)\right].$$

We have,

$$\begin{aligned} E[(m^*)^2] &= [\bar{m} + (a\epsilon)^\top \beta_{\bar{m}}]^2 \\ &= \bar{m}^2 + E[(a\epsilon)^\top \beta_{\bar{m}}]^2 \\ &= \bar{m}^2 + E[(a\epsilon)^\top \beta_{\bar{m}} \cdot (a\epsilon)^\top \beta_{\bar{m}}]^2 \\ &= \bar{m}^2 + E[(\beta_{\bar{m}}^\top a\epsilon)(\epsilon^\top a^\top \beta_{\bar{m}})] \\ &= \bar{m}^2 + \beta_{\bar{m}}^\top \cdot \sigma \cdot \beta_{\bar{m}} \\ &= \bar{m}^2 + [\mathbf{1}_n^\top - \bar{m}(\mathbf{1}_n^\top + b^\top)]\sigma^{-1}[\mathbf{1}_n - \bar{m}\left(\mathbf{1}_n + b\right)] \\ &= \bar{m}^2 + \mathbf{1}_n^\top \sigma^{-1} \mathbf{1}_n - \bar{m}\left(\mathbf{1}_n^\top \sigma^{-1} \mathbf{1}_n + \mathbf{1}_n^\top \sigma^{-1} b\right) \\ &- \bar{m}[\mathbf{1}_n^\top \sigma^{-1} \mathbf{1}_n + b^\top \sigma^{-1} \mathbf{1}_n - \bar{m}(\mathbf{1}_n^\top \sigma^{-1} \mathbf{1}_n + b^\top \sigma^{-1} \mathbf{1}_n + \mathbf{1}_n^\top \sigma^{-1} b + b^\top \sigma^{-1} b)] \end{aligned}$$

Again in terms of the notation of Section 7.8 ($\gamma \equiv \mathbf{1}_n^\top \sigma^{-1} \mathbf{1}_n$ and $\beta \equiv \mathbf{1}_n^\top \sigma^{-1} b$), this is:

$$E[(m^*)^2] = \gamma - 2\bar{m}\left(\gamma + \beta\right) + \bar{m}^2 \left(1 + \gamma + 2\beta + b^\top \sigma^{-1} b\right).$$

Thus, the expected return is

$$E\left(R^{m^*}\right) - 1 = \frac{E\left(m^*\right)}{E[\left(m^*\right)^2]} - 1 = \frac{\bar{m} - \gamma + 2\bar{m}\left(\gamma + \beta\right) - \bar{m}^2 \left(1 + \gamma + 2\beta + b^\top \sigma^{-1} b\right)}{\gamma - 2\bar{m}\left(\gamma + \beta\right) + \bar{m}^2 \left(1 + \gamma + 2\beta + b^\top \sigma^{-1} b\right)}.$$

Next, recall the following two definitions:

$$\bar{m} = \frac{1}{1+r}, \quad Sh = (b - \mathbf{1}_m r)^\top \sigma^{-1} (b - \mathbf{1}_m r) = b^\top \sigma^{-1} b - 2\beta r + \gamma r^2.$$

In terms of r and Sh, we have,

$$\begin{aligned} E(R^{m^*}) - 1 &= \frac{E(m^*)}{E[(m^*)^2]} - 1 \\ &= -\frac{\gamma(1+r)^2 - (1+r)(1+2\gamma+2\beta) + 1 + \gamma + 2\beta + b^\top \sigma^{-1} b}{\gamma(1+r)^2 - (1+r)(2\gamma+2\beta) + 1 + \gamma + 2\beta + b^\top \sigma^{-1} b} \\ &= \frac{r - Sh}{1 + Sh} \\ &= r - \frac{1+r}{1+Sh} Sh \\ &< r. \end{aligned}$$

This is positive if $r - Sh > 0$, i.e. if $b^\top \sigma^{-1} b - (2\beta + 1) r + \gamma r^2 < 0$, which is possible for sufficiently low (or sufficiently high) values of r.

Proof that R^{m^*} is the m-maximum correlation portfolio. We have to show that for any stochastic discount factor m, $|corr(m, R_{bt})| \geq |corr(m, R_{pt})|$. Define a ℓ-parametrized portfolio such that:

$$E[(1-\ell)R_o + \ell R_{pt}] = E(R_{bt}), \quad R_o \equiv \bar{m}^{-1}.$$

We have

$$\begin{aligned} corr(m, R_{pt}) &= corr[m, (1-\ell)R_o + \ell R_{pt}] \\ &= corr[m, R_{bt} + ((1-\ell)R_o + \ell R_{pt} - R_{bt})] \\ &= \frac{cov(m, R_{bt}) + cov(m, (1-\ell)R_o + \ell R_{pt} - R_{bt})}{\sigma(m) \cdot \sqrt{var((1-\ell)R_o + \ell R_{pt})}} \\ &= \frac{cov(m, R_{bt})}{\sigma(m) \cdot \sqrt{var((1-\ell)R_o + \ell R_{pt})}} \end{aligned}$$

The first line follows because $(1-\ell)R_o + \ell R_{pt}$ is a nonstochastic affine translation of R_{pt}. The last equality follows because

$$\begin{aligned} cov(m, (1-\ell)R_o + \ell R_{pt} - R_{bt}) &= E[m \cdot ((1-\ell)R_o + \ell R_{pt} - R_{bt})] \\ &= (1-\ell) \cdot \underbrace{E(mR_o)}_{=1} + \ell \cdot \underbrace{E(mR_{pt})}_{=1} - \underbrace{E(m \cdot R_{bt})}_{=1} \\ &= 0. \end{aligned}$$

where the first line follows because $E((1-\ell)R_o + \ell R_{pt}) = E(R_{bt})$. Therefore,

$$corr(m, R_{pt}) = \frac{cov(m, R_{bt})}{\sigma(m) \cdot \sqrt{var((1-\ell)R_o + \ell R_{pt})}} \leq \frac{cov(m, R_{bt})}{\sigma(m) \cdot \sqrt{var(R_{bt})}} = corr(m, R_{bt}),$$

where the inequality follows because R_{bt} is mean-variance efficient (i.e. $\nexists$ feasible portfolios with the same expected return as R_{bt} and variance less than $var(R_{bt})$), and then $var((1-\ell)R_o + \ell R_{pt}) \geq var(R_{bt})$, all R_{pt}.

References

Alvarez, F. and U.J. Jermann (2005): "Using Asset Prices to Measure the Persistence of the Marginal Utility of Wealth." *Econometrica* 73, 1977-2016.

Cecchetti, S., Lam, P-S. and N. C. Mark (1994): "Testing Volatility Restrictions on Intertemporal Rates of Substitution Implied by Euler Equations and Asset Returns." *Journal of Finance* 49, 123-152.

Cochrane, J. (1996): "A Cross-Sectional Test of an Investment-Based Asset Pricing Model." *Journal of Political Economy* 104, 572-621.

Epstein, L.G. and S.E. Zin (1989): "Substitution, Risk-Aversion and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework." *Econometrica* 57, 937-969.

Epstein, L.G. and S.E. Zin (1991): "Substitution, Risk-Aversion and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis." *Journal of Political Economy* 99, 263-286.

Ferson, W. E. and A. F. Siegel (2003): "Stochastic Discount Factor Bounds with Conditioning Information." *Review of Financial Studies* 16, 567-595.

Gallant, R. A., L. P. Hansen and G. Tauchen (1990): "Using the Conditional Moments of Asset Payoffs to Infer the Volatility of Intertemporal Marginal Rates of Substitution." *Journal of Econometrics* 45, 141-179.

Gordon, M. (1962): *The Investment, Financing, and Valuation of the Corporation.* Homewood, IL: Irwin.

Guvenen, F. (2009): "A Parsimonious Macroeconomic Model for Asset Pricing." *Econometrica* 77, 1711-1740.

Hansen, L. P. and R. Jagannathan (1991): "Implications of Security Market Data for Models of Dynamic Economies." *Journal of Political Economy* 99, 225-262.

Mehra, R. and E. C. Prescott (1985): "The Equity Premium: A Puzzle." *Journal of Monetary Economics* 15, 145-161.

Weil, Ph. (1989): "The Equity Premium Puzzle and the Risk-Free Rate Puzzle." *Journal of Monetary Economics* 24, 401-421.

8
Aggregate fluctuations in equity markets

8.1 Introduction

This chapter discusses empirical regularities of the aggregate equity market, how these properties relate to the business cycle, and the extent to which the neo-classical model can account for them. This chapter is, thus, a natural development of the previous. Its motivation is still to explain how existing models can help rationalize the extant empirical evidence. However, its focus now regards how the aggregate stock market relates to macroeconomic fluctuations. When are stock returns pro-cyclical? When is stock volatility countercyclical? Such are the questions this chapter addresses.

The analysis of this chapter regards how to reverse-engineer the *qualitative* behavior of economic forces that is consistent with the empirical properties of the aggregate stock market. In technical terms, we would like to identify the pricing kernels that are consistent with equity returns, equity expected returns and equity return volatility. We consider two broad, and not inconsistent, explanations. In the first, agents evaluate assets relying on time-varying discount rates. Thus, in these economies, the cyclical properties of aggregate equity markets are explained by the agents' optimal response to shocks on fundamentals. In the second class of models, expected growth is time-varying. This variability translates to fluctuations in aggregate stock market volatility. For example, time-varying expected growth may arise because agents have incomplete knowledge of the state of the economy, and try to infer it from public signals. If the agents estimate that the probability of living in the good state is time-varying, they will view expected growth as being random. This randomness becomes a source of asset volatility.

The previous properties of the pricing kernels as well as the agents' inference processes form the basis of more advanced discussions in the next chapter. Note, finally, that the models we analyze in this chapter do not necessarily lead to a resolution of the *quantitative* puzzles surveyed in the previous chapter. To illustrate, even if a given model likely leads to interesting dynamics, further scrutiny is required regarding the size of the equity premium. We may find a model predicts countercyclical volatility, as in the data, and yet this volatility can be orders of magnitude lower than in the data.

A final remark is in order concerning the very nature of aggregate stock market fluctuations. The empirical evidence reviewed in this chapter (see Section 8.2) suggests that a few key

market statistics are quite stationary in relation to their historical behavior vis-à-vis the business cycle. It is interesting, especially in light of the fact that capital markets have undergone significant changes over time, which mainly affected various aspects of their *microstructure*—say transactions technology, the price discovery process, liquidity, or volumes, to mention a few examples. How is it that the properties of the aggregate stock market reviewed in this chapter do not appear to be affected by these changes? One simple possibility is that market microstructure regards the very high frequency behavior of markets, whereas the properties studied in this chapter relate to slow, low frequency movements, which should not be too affected by market microstucture details. The models in this chapter aim to rationalize these movements, as explained. Models addressing the previous market microstructure issues are reviewed in Chapter 10.

More in detail, this chapter is organized as follows. The next section provides a succinct overview of empirical regularities of aggregate equity markets at the business cycle frequency; we explain that price-dividend ratios and stock returns are procyclical, whereas stock volatility and risk-premiums are countercyclical. Section 8.3 analyzes in deeper detail the empirical behavior of aggregate volatility, and provides intuitive explanations for it. Section 8.4 develops a framework to think about our countercyclical statistics. Section 8.5 analyzes the two classes of economies with which it illustrates the predictions of Section 8.4. The modeling approach in this chapter is based on the price-dividend ratios. Section 8.6 develops an alternative approached based on B/M ratios. [In progress]

8.2 Empirical evidence: bird's eye view

While the aggregate equity market is certainly unpredictable in the short-run as Fama showed in his seminal work (see Fama, 2014), the work of Shiller shows that it moves much more than justified by the fundamentals (see Shiller, 2014). This *excess volatility* puzzle is also associated with *predictability* in the medium-term: in bad times (say in recessions), the market requires high returns while in good times, the market appetite for risk is high. These oscillations in the investors' "mood" lead stock prices to wide and predictable fluctuations in the medium-term, that is, excess volatility.

The conclusions of Fama and those of Shiller do not contradict with each other. Fama's work centers on short-term fluctuations, whereas Shiller's focus regards the medium-term, as explained. Regarding the short-run, Chapter 10 revisits Fama's work on "market efficiency" in light of information problems. Instead, this and the following chapter analyze aggregate stock market fluctuations at a lower frequency, i.e., that of the business cycle. Note, in most of the models of this and the next chapter, the investors' "mood" is rational, that is, linked to a rational assessment of risk or uncertainty in light of preferences and beliefs.

8.2.1 Equity markets and the business cycle

The evidence of a linkage between the aggregate stock market and the business cycle is indeed both striking and well-known since long time (see, e.g., the early survey in Campbell, 2003). This section aims to streamline how aggregate stock market fluctuations relate to general macroeconomic conditions. Note that we do not aim to establish any "causality link" amongst the various variables we analyze. Rather, the analysis in this section could be best described as one delivering "descriptive statistics."

We use monthly data that cover the period from January 1948 through December 2002. We calculate yearly returns at month t as $\sum_{i=1}^{12} \tilde{R}_{t+1-i}$, where $\tilde{R}_t = \ln(\frac{S_t+D_t}{S_{t-1}})$, S_t is the S&P Composite index at month t, and D_t is the aggregate dividend calculated by Robert Shiller. Table 8.1 provides basic statistics for both raw data such as the P/D ratio, realized returns, and stock volatility and expected returns. Stock volatility is calculated as follows:

$$\text{Vol}_t \equiv \sqrt{6\pi} \cdot \bar{\sigma}_t, \quad \bar{\sigma}_t \equiv \frac{1}{12} \sum_{i=1}^{12} \left| \tilde{R}_{t+1-i} - r_{t+1-i} \right|, \tag{8.1}$$

where r_t is the risk-free rate, taken to be the one month bill return.[1] Expected returns are calculated as explained below (see Eq. (8.2)). With the exception of the P/D ratio, all figures are annualized percent.

We note the first main set of stylized facts:

> Facts I. The P/D ratio and realized returns are procylical, although variations in the business cycle conditions do not seem to be their only driving forces.

For example, Figure 8.1 reveals that the price-dividend ratio on the S&P500 declines during *all* economic slowdowns, as signaled by the recession indicator calculated by the National Bureau of Economic Research (NBER)—the NBER recessions. At the same time, during NBER expansions, price-dividend ratios seem to be driven by additional factors not necessarily related to the business cycle. For example, during the "roaring" 1960s, price-dividend ratios experienced two major drops that display the same order of magnitude as the decline at the very beginning of the "chaotic" 1970s. Realized returns follow similar patterns: see Figure 8.2.

	total		NBER expansions		NBER recessions	
	average	std dev	average	std dev	average	std dev
P/D ratio	31.99	15.88	33.21	15.79	26.20	14.89
$\ln \frac{\text{P/D}_{t+1}}{\text{P/D}_t}$	2.01	12.13	3.95	10.81	−7.28	16.79
one year returns	8.59	15.86	12.41	13.04	−9.45	15.49
real risk-free rate	1.02	2.48	1.03	2.43	0.97	2.69
excess return volatility	14.55	4.68	14.05	4.47	16.91	4.91
expected returns	8.36	3.49	8.09	3.29	9.62	4.10

TABLE 8.1. Data are sampled monthly and cover the period from January 1948 through December 2002. With the exception of the P/D ratio levels, all figures are annualized percent.

Table 8.1 also reports estimates of a key component of asset evaluation, the expected market return (i.e. the return that investors expect while investing in the stock market) at each point in

[1] The rationale behind this calculation is as follows. First, $\bar{\sigma}_t$ is an estimate of the average volatility occurring over the last 12 months. We annualize $\bar{\sigma}_t$ by multiplying it by $\sqrt{12}$. The term $\sqrt{6\pi}$ arises for the following reason. If we assume that a given return $R = \sigma u$, where σ is a positive constant and u is a standard unit normal, then $E(|R|) = \sigma\sqrt{\frac{2}{\pi}}$. The definition Vol_t in Eq. (8.1), then, follows by multiplying $\sqrt{12}\bar{\sigma}_t(\ell)$ by $\sqrt{\frac{\pi}{2}}$. This correction term, $\sqrt{\frac{\pi}{2}}$, has been suggested by Schwert (1989a) in a related context.

time of our sample. Appendix 1 describes a procedure to estimate such an unobserved variable. This chapter relies on our reconstruction of yearly expected returns, defined as

$$\hat{\mathcal{E}}_t \equiv \sum_{j=1}^{12} P_t(\hat{R}_{t+j}), \tag{8.2}$$

where $\hat{R}_{t+j}$ is the expected return at time $t+j$, and $P_t(\hat{R}_{t+j})$ denotes the projection of $\hat{R}_{t+j}$ given the information at t (based on the model described in the appendix).

The following facts summarize the cyclical properties of realized volatility and the expected returns.

> FACTS II. Stock volatility and expected returns are countercyclical. However, business cycle conditions do not seem to be the only forces explaining the swings in these variables.

That expected returns are countercyclical is a well-known fact since at least Fama and French (1989) and Ferson and Harvey (1991). Figures 8.2 through 8.4 are, indeed, suggestive of the business cycle pattern of the aggregate stock market. For example, Figure 8.3 depicts the statistical relation between stock volatility and the industrial production growth rate, which shows that stock volatility is largely countercyclical, being larger in bad times than in good.

There are, of course, exceptions. For example, stock volatility rocketed to almost 23% around the October 1987 crash—a crash occurring during one of the most enduring post-war expansion period. Countercyclical volatility is a stylized fact extensively discussed in Sections 8.3 and 8.4. An important lesson is that this empirical fact can be explained once the volatility of the P/D ratios *changes* is countercyclical. Interestingly, Table 8.1 reveals that the P/D ratios variations are more volatile in bad times than in good.

Finally, Figure 8.3 suggests that stock volatility behaves asymmetrically over the business cycle, in that it increases more in bad times than it decreases in good. This asymmetric behavior of stock volatility echoes its high frequency behavior documented at least since Glosten, Jagannathan and Runkle (1993)—stock volatility increases more when returns are negative than it decreases when returns are positive.

A third set of stylized facts links to the asymmetric behavior of the aggregate stock market fluctuations over the business cycle. We have already noted that volatility is countercyclical and behaves asymmetrically. We now turn to price multiples and expected returns.

> FACTS III. *Changes* in the P/D ratios and expected returns occur asymmetrically over the business cycle. The most severe variations in these variables occur during the contractionary phases of the business cycle.

During recessions, price multiples and expected returns move more than they do in good times. For example not only are expected returns countercyclical. On average, expected returns increase more during NBER recessions than they decrease during NBER expansions, consistently with the summary statistics of Table 8.1, and the informal pieces of evidence from Figure 8.2. We shall return to this issue of countercyclical expected returns many times in this chapter.

Similarly, not only are P/D ratios procyclical. On average, P/D ratios increase less during NBER expansions than they decrease during NBER recessions. Furthermore, this asymmetric behavior is quantitatively quite pronounced. Consider, for example, the changes in the P/D ratios: on average, their percentage (negative) changes during recessions is nearly twice as the percentage (positive) changes during expansions. Sections 8.3 and 8.4 aim to provide basic explanations of these facts.

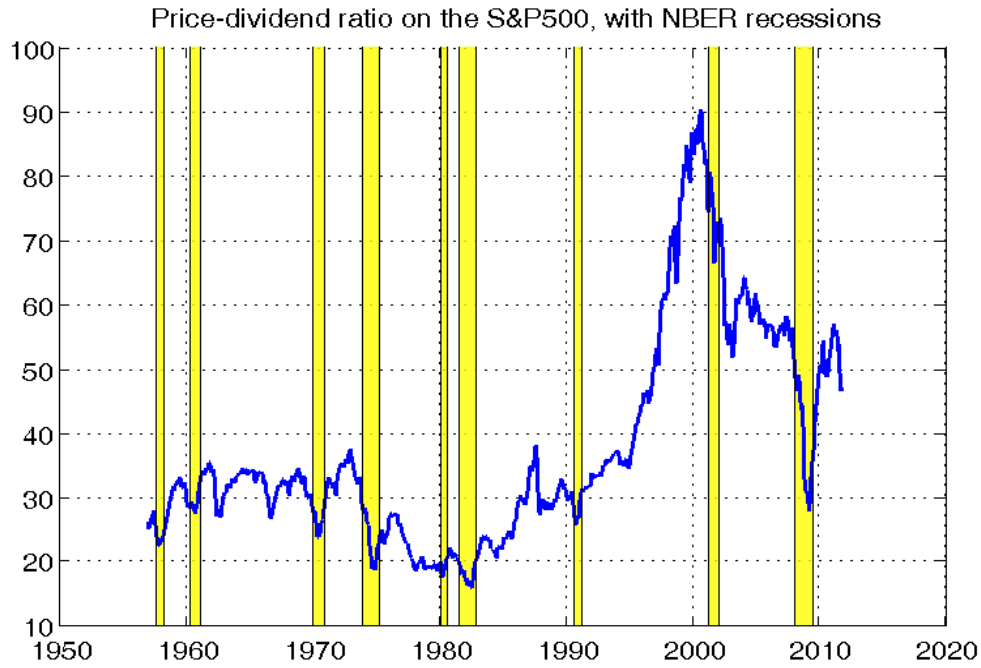


FIGURE 8.1

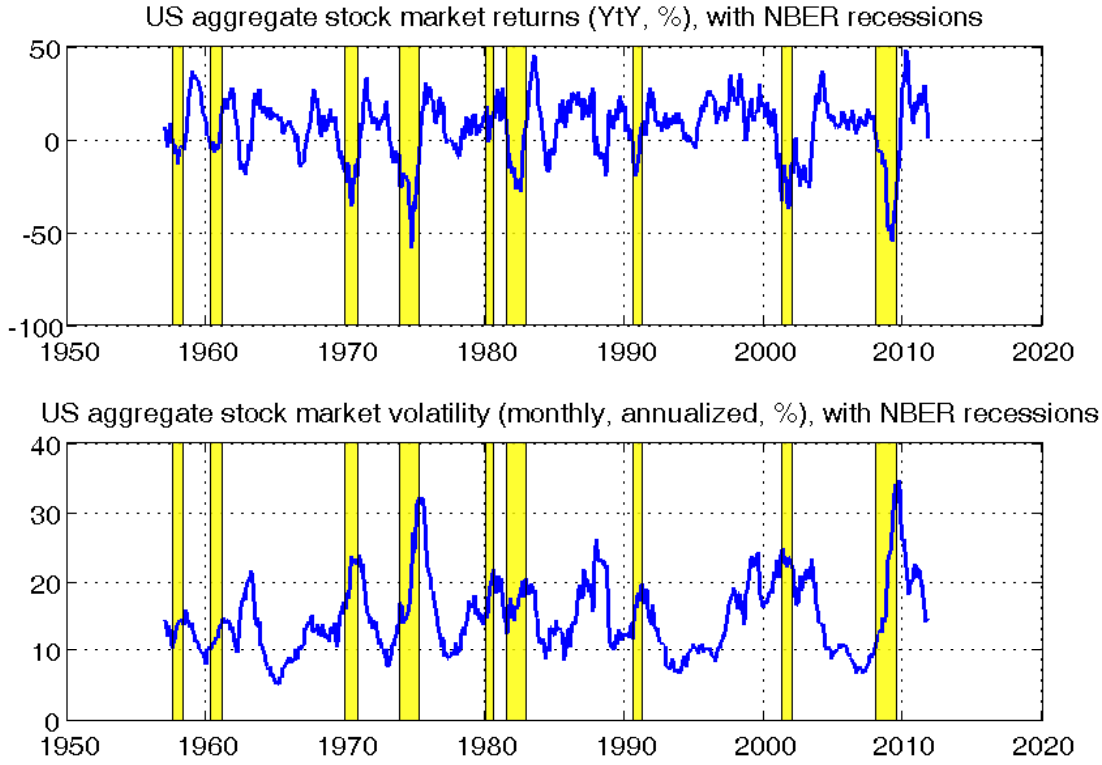


FIGURE 8.2. Monthly excess returns and volatility, in percentage, year-to-year.

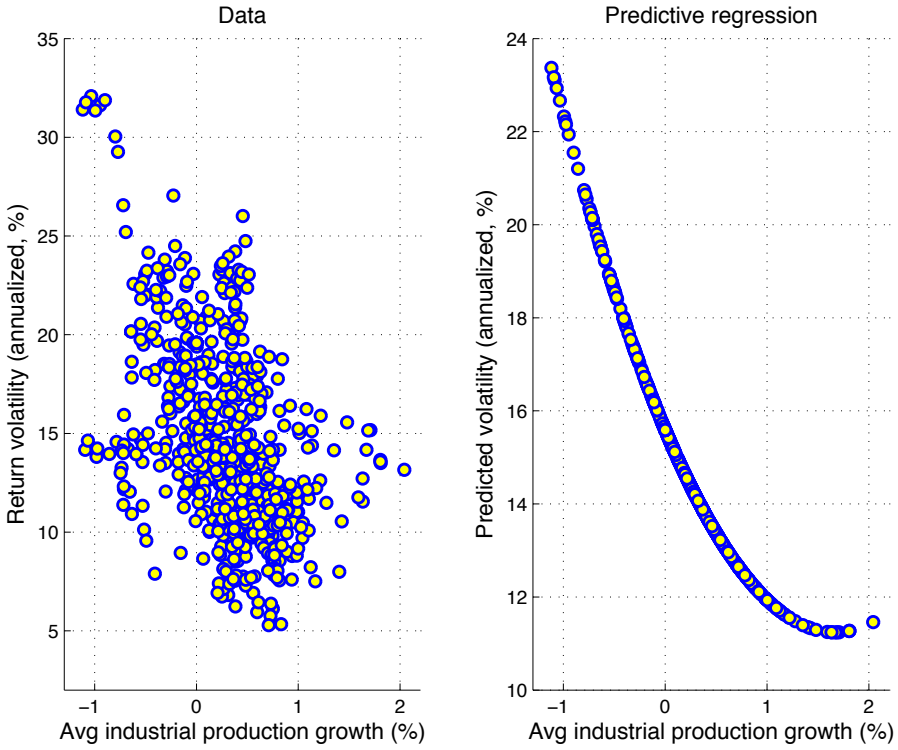


FIGURE 8.3. Stock volatility and business cycle conditions. The left panel plots stock volatility, Vol_t, against yearly (deseasoned) industrial production average growth rates, computed as $\text{IP}_t \equiv \frac{1}{12}\sum_{i=1}^{12}\text{Ind}_{t+1-i}$, where where Ind_t is the real, seasonally adjusted industrial production growth as of month t. The right panel depicts the prediction of the ordinary least squares regression: $\text{Vol}_t = \underset{(0.19)}{15.59} - \underset{(0.39)}{5.21}\cdot\text{IP}_t + \underset{(0.41)}{1.56}\cdot\text{IP}_t^2 + w_t$, where w_t is a residual term, and robust standard errors are in parenthesis. The data span the period from January 1948 to December 2002.

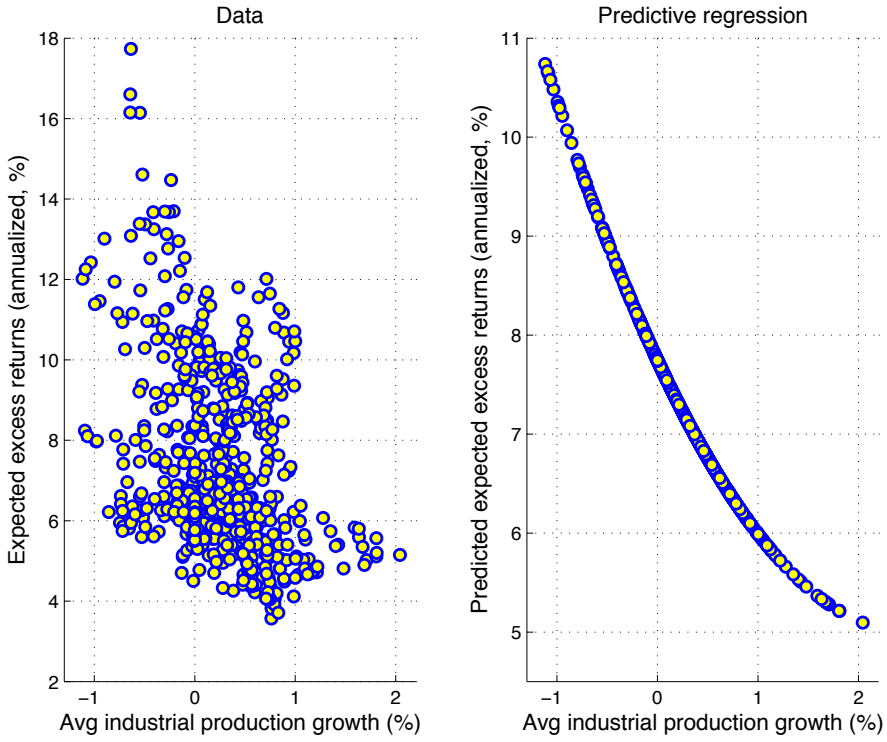


FIGURE 8.4. The left-hand side of this picture plots estimates of the expected returns (annualized, percent) ($\hat{\mathcal{E}}_t$ say) against yearly (deseasoned) industrial production average growth rates, computed as $\mathrm{IP}_t \equiv \frac{1}{12}\sum_{i=1}^{12} \mathrm{Ind}_{t+1-i}$, where where Ind_t is the real, seasonally adjusted industrial production growth as of month t. Expected returns are estimated through Eq. (8.2). The right-hand side of this picture depicts the prediction of the ordinary least squares regression: $\hat{\mathcal{E}}_t = \underset{(0.09)}{7.74} - \underset{(0.19)}{2.19} \cdot \mathrm{IP}_t + \underset{(0.19)}{0.43} \cdot \mathrm{IP}_t^2 + w_t$, where w_t is a residual term, and robust standard errors are in parenthesis. Data are sampled monthly, and span the period from January 1948 to December 2002.

8.2.2 Predictability

Facts I entail a quite intuitive consequence: price-dividend ratios convey information relating to future returns. After all, expansions are followed by recessions. Therefore, in good times, the stock market predicts that future returns will be negative. Define the excess return for the time period $[t, t+n]$ as $\tilde{R}^e_{t,t+n} \equiv \tilde{R}_{t,t+n} - r_{t,t+n}$, where $\tilde{R}_{t,t+n}$ is the asset return over $[t, t+n]$, and $r_{t,t+n}$ is the sum of the one-month Treasury bill rate, taken over $[t, t+n]$. Consider the following regressions,

$$\tilde{R}^e_{t,t+n} = a_n + b_n \times \mathrm{P/D}_t + u_{n,t}, \quad n \geq 1, \tag{8.3}$$

where u_n is a residual term. Typically, the estimates of b_n are significantly negative, with regression R^2 that increase with n. [In progress, provide these estimates]

In turn, the previous regressions imply that $E(\tilde{R}^e_{t+n}|\mathrm{P/D}_t) = a_n + b_n \times \mathrm{P/D}_t$. Thus, they suggest that price-dividend ratios are driven by expected excess returns. In this restrictive sense, countercyclical expected returns (Facts II) and procyclical price-dividend ratios (Facts I) might be two sides of the same coin. To investigate how this predictability links to business cycle

developments, consider the following regression, performed with monthly data from 1948:01 to 2002:12,

$$\tilde{R}^e_{t-12,t} = \underset{(1.04)}{14.64} - \underset{(1.37)}{9.09} \times \text{IP}_{t-12} - \underset{(2.67)}{14.27} \times \text{Infl}_{t-12} + u_t, \quad \text{with } \text{R}^2 = 11\%, \tag{8.4}$$

where robust standard errors are in parenthesis, u is a residual term, $\tilde{R}^e_{t-12,t}$ is the excess return from $t-12$ to t, IP_t is the average industrial production growth over the previous twelve months, as defined in Figure 8.3, and Infl_t is defined similarly as IP_t.

The negative signs of the coefficients in Eq. (8.4) are quite to be expected. Economic activity does display mean-reverting behavior: bad times are followed by good. But good times are those where the stock market goes up. Therefore, a slowdown in economic activity is a predictor of high returns in the future. To illustrate with a simple example, consider a case where the aggregate stock market positively links to a single state variable tracking the business cycle conditions x_t, say, such that the log of the aggregate equity index is $\ln S_t = s_0 + s_x x_t$, for two constant s_0 and s_x, and where $s_x > 0$. Assume, then, and critically, that x_t is mean-reverting, with unconditional expectation μ, speed of adjustment $\kappa > 0$, and some volatility coefficient $\sigma(x_t)$,

$$dx_t = \kappa(\mu - x_t)\, dt + \sigma(x_t)\, dW_t,$$

where W_t is a Brownian motion. Then, it is straightforward to show that $E_{t-12}\left(\ln \frac{S_t}{S_{t-12}}\right) = a_0 - a_1 x_{t-12}$, where E_t denotes the expectation conditional on the information at t, $a_0 \equiv s_x\left(1 - e^{-12\kappa}\right)\mu$ and $a_1 \equiv s_x\left(1 - e^{-12\kappa}\right)$. That is, if x_t is mean-reverting, $\kappa > 0$, and the aggregate stock market is procyclical, $s_x > 0$, expected returns negatively link to past values of x_t, i.e. $a_1 > 0$. This reasoning generalizes to a multivariate case, although the presence of feedbacks between macroeconomic variables might then dilute the contribution of each variable as a predictor of future expected returns.

The regression results in Eq. (8.4) have the same nature as that underlying Eq. (8.3): price-dividend ratios and market returns are procyclical. Note a final relation that reveals this procyclicality of market returns. At a contemporaneous level, the excess market returns are positively related to industrial production and negatively related to inflation,

$$\tilde{R}^e_{t-12,t} = \underset{(1.07)}{10.47} + \underset{(1.19)}{7.27} \times \text{IP}_t - \underset{(2.91)}{16.33} \times \text{Infl}_t + w_t, \quad \text{with } \text{R}^2 = 14\%, \tag{8.5}$$

where robust standard errors in parenthesis, and w is a residual term. Corradi, Distaso and Mele (2013) estimate a continuous time model in which the aggregate stock market is driven by industrial production, inflation and one unobserved factor. They show that in the model, the relation between the market return and the two macroeconomic factors are similar to those summarized by the linear regression in Eq. (8.5).

Finally, an apparently puzzling feature is that price-dividend ratios do not predict future dividend growth. Let $g_t \equiv \ln(D_t / D_{t-1})$. In regressions taking the following format,

$$g_{t+n} = a_n + b_n \times \text{P/D}_t + u_{n,t}, \quad n \geq 1,$$

the predictive content of price-dividend ratios is poor, and estimates of b_n might often come with a wrong sign.

To summarize, the previous regression estimates suggest that: (i) the price-dividend ratios are driven by time-varying expected returns (a proxy of the risk-premium as we shall explain), and

(ii) the role played by expected dividend growth is somewhat limited. In the next chapter (see Section 9.11), we shall explain that this view can be challenged along several dimensions. First, it seems that expected *earning growth* does help predicting price-dividend ratios. Second, the fact expected dividend growth does not seem to affect price-dividend ratios can be a property to be expected in equilibrium.

8.2.3 Risk-return trade-offs

Expected returns and stock volatility are both countercyclical and, hence, they positively relate at a business cycle frequency. Figure 8.5 provides a scatterplot of expected returns against realized volatility. It is tempting to rely on this evidence and conclude that higher expected returns arise when the market requires compensation for higher volatility—a conclusion apparently consistent with the prediction of the CAPM (see Chapter 1). This conclusion is, of course, misleading for the CAPM is wrong if volatility is random. If volatility is driven by business cycle factors, it is these very same factors to drive expected returns—volatility does, then, merely react to these factors. Let us elaborate with a simple model.

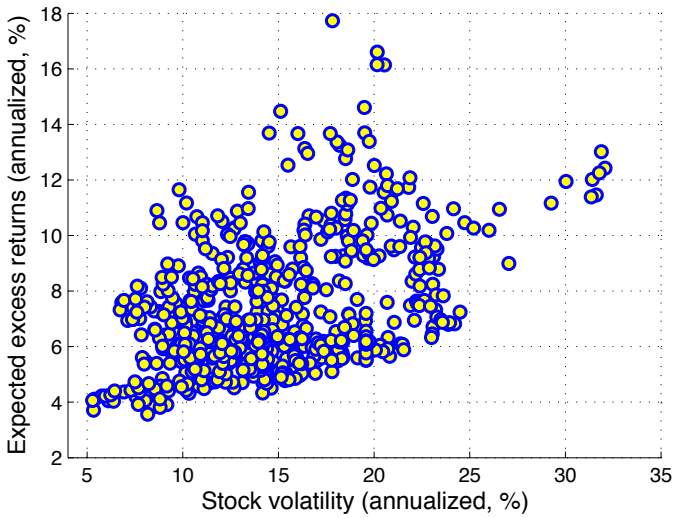


FIGURE 8.5. Expected returns and volatility.

Consider an economy with a single asset and a representative agent. The price of the asset, S_t, is solution to

$$\frac{dS_t}{S_t} = (\cdots)\, dt + \sigma_t dW_{1t},$$

and the asset return volatility σ_t is random

$$d\sigma_t^2 = (\cdots)\, dt + \xi\left(\sigma_t\right)\left(\rho dW_{1t} + \sqrt{1-\rho^2}dW_{2t}\right),$$

where W_{it} are two standard Brownian motions, $\xi\left(\sigma_t\right)$ is a given function and, finally, ρ is the instantaneous (and constant) correlation between the asset returns and volatility. Now, by results in Part I (see Eq. (5.81) in Chapter 5), we have that in equilibrium, the asset expected excess returns are

$$\mu - r = \frac{-V\mathcal{J}_{VV}}{\mathcal{J}_V}\sigma_t^2 - \sigma_t \xi \rho \frac{\mathcal{J}_{V\sigma}}{\mathcal{J}_V}, \tag{8.6}$$

where $\mathcal{J}$ is the indirect utility of a representative agent (with subscripts denoting partial derivatives) and V is his total wealth.

That is, even in this simple model, a risk-return tradeoff is not as clear-cut as the CAPM suggests: expected returns are driven by volatility but, also, by the volatility of volatility. Furthermore, this relation can be strong or weak according to parameter values. Even more fundamentally, in this simple model, volatility is taken to be exogenous. This assumption is removed in all the models of this and the following chapter. We shall study more realistic economies in which asset prices clear markets in general equilibrium, such that the dynamics of volatility are endogenous, arising as a result of the investors' optimal reaction to changes in the fundamentals. In these economies, Eq. (8.6) is not the relevant pricing equation; rather, expected excess returns link to the primitive factors affecting the investors' space choice. If any, investors require risk-premiums in relation to these factors, not realized volatility. Asset market volatility can only be understood as as a results of the forces affecting expected returns. We now proceed with explaining these forces.

8.3 Volatility: a business cycle perspective

A prominent feature of the U.S. equity market is the close connection between aggregate volatility and business cycle developments discussed in the previous section (see, e.g., Figure 8.3). Understanding the origins and implications of these facts is extremely relevant to policy makers. Indeed, if stock market volatility is countercyclical, it must necessarily be encoding information about the development of the business cycle. Policy makers could then extract the signals stock volatility brings regarding business cycle developments.

This section accomplishes three tasks. First, it mentions additional stylized facts relating stock volatility, expected returns and P/D ratios over the business cycle (in Section 8.3.1). Second, it attempts at preliminary explanations of these facts (in Section 8.3.2). Third, it investigates whether stock volatility contains any useful information about the business cycle (in Section 8.3.3).

There are topics left over from this section. For example, we do not tackle statistical issues regarding volatility measurement.[2] Nor do we consider the role of volatility in applied asset evaluation. Instead, Chapter 11 covers details about how time-varying volatility affects derivative pricing. The focus of this section is more fundamental. It explores the extent to which volatility movements could be given a wider business cycle perspective, and also highlights some of the rational mechanisms underlying them.

8.3.1 Volatility cycles

How does equity volatility relate to the business cycle? One of the first contributions to this literature is Schwert (1989a,b).[3] Schwert points out that low frequency fluctuations in equity volatility are difficult to explain through those in the volatility of other macroeconomic variables.

For example, industrial production volatility does not correlate with stock volatility. Let us calculate industrial production volatility as $\text{Vol}_{\text{G},t} \equiv \frac{1}{\sqrt{12}} \sum_{i=1}^{12} |G_{t+1-i}|$, where G_t is the real, seasonally adjusted industrial production growth rate at month t (similarly as in Eq. (8.1)).

[2] See, e.g., Andersen, Bollerslev and Diebold (2002) for an early survey regarding the many available statistical techniques to estimate volatility.

[3] Additional important contributions include Hamilton and Lin (1996) and Brandt and Kang (2004). See Corradi, Distaso and Mele (2013) for additional work and references.

Figure 8.6 plots stock volatility against $\text{Vol}_{G,t}$; it does not reveal any obvious pattern between these two variables. These results are in striking contrast with those available from Figure 8.3, where, instead, stock volatility exhibits a quite clear countercyclical behavior. More in detail, Table 8.1 reveals that stock market volatility is almost 30% higher during NBER recessions than during NBER expansions.

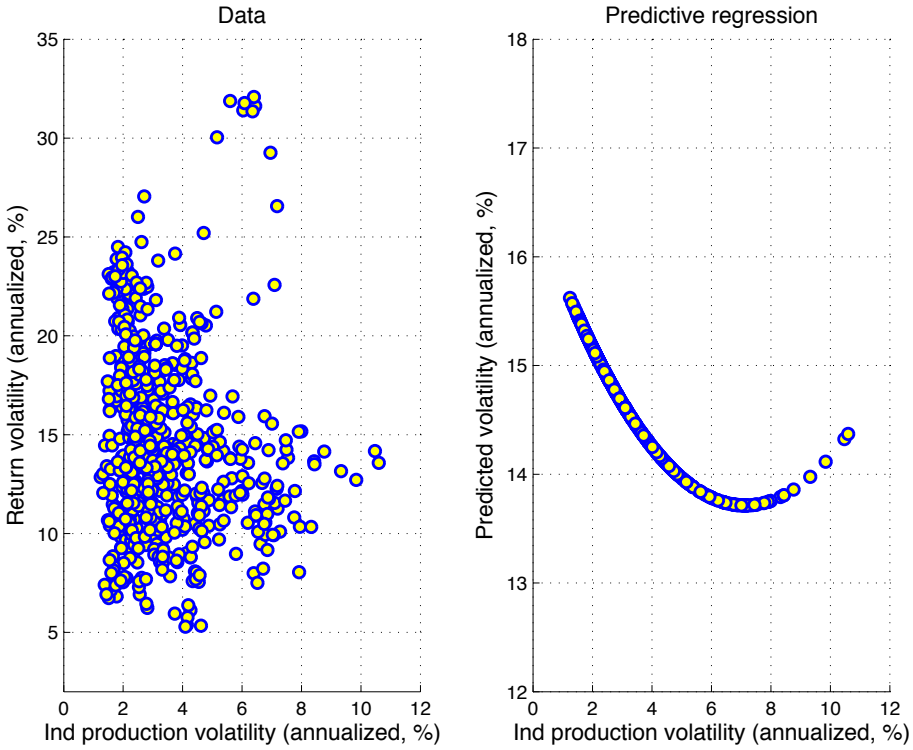


FIGURE 8.6. Return volatility and industrial production volatility. The left panel plots stock volatility, Vol_t, against industrial production volatility, $\text{Vol}_{G,t}$. The right panel of the picture depicts the prediction of the ordinary least squares regression: $\text{Vol}_t = \underset{(0.93)}{16.51} - \underset{(0.47)}{0.78} \cdot \text{Vol}_{G,t} + \underset{(0.05)}{0.05} \cdot \text{Vol}_{G,t}^2 + w_t$, where w is a residual term, and standard errors are in parenthesis. The data span the period from January 1948 to December 2002.

In fact, Schwert also shows that stock volatility *is* countercyclical. The main focus of this section is to provide a few explanations of this evidence, which supports the view equity volatility relates to the business cycle, although not precisely related to the *volatility* of other macroeconomic variables.

A seemingly unrelated but well-known stylized fact is that risk-premiums are countercyclical, as summarized by Facts II in the previous section. Particularly important are also Facts III: expected returns lower much less during expansions than they increase during recessions. With post-war data, we find that compared to an average of 8.36%, the expected returns increase by nearly 19% during recessions while they drop by a mere 3% during NBER expansions (see Table 8.1). A final stylized fact is that price-dividend ratios behave asymmetrically over the business cycle. Table 8.1 reveals that not only are price-dividend ratios pro-cyclical. Their downward changes during recessions are also more severe than their upward movements during expansions.

Table 8.1 suggests that price-dividend ratios fluctuate nearly two times more in recessions than in expansions.

How can we rationalize these facts? A simple possibility is that the economy is frequently hit by shocks that display the same qualitative behavior of return volatility, expected returns and price-dividend ratios. However, Figure 8.6 suggests this channel is unlikely. Another possibility is that the economy reacts to shocks, thanks to some mechanism endogenously related to the investors' maximizing behavior, which then activates the previous phenomena.

Section 8.3.2 puts forward explanations for countercyclical stock volatility relying on endogenous mechanisms. Section 8.3.3 provides, instead, additional empirical properties of equity volatility. The motivation is simple: because stock volatility is countercyclical, it might contain useful information about ongoing business cycle developments. The section, then, aims to provide answers to the following questions: (i) Do macroeconomic factors help explain the dynamics of stock market volatility? (ii) Conversely, what is the predictive content stock market volatility brings about the business cycle?

8.3.2 Understanding the empirical evidence

This section is a broad introduction to the main determinants of aggregate stock market fluctuations. We develop a simple example of an economy, where countercyclical volatility arises in conjunction with the property that investors' required return are (i) countercyclical, and (ii) *asymmetrically* related to business cycle developments. That is, in this economy, risk-premiums increase more in bad times than they decrease in good, consistently with the evidence discussed earlier in this chapter.

8.3.2.1 Fluctuating compensation for risk

In frictionless markets, the price of a long-lived security is simply the risk-adjusted discounted expectation of the future dividends stream. Heuristically, and other things being equal, this price increases as the expected return from holding the asset and, hence, the risk-premium, decreases. According to this mechanism, asset prices and price-dividend ratios are pro-cyclical because risk-adjusted discount rates are countercyclical.

Would countercyclical risk-premiums also lead to countercyclical volatility? We now explain that an addditional property is required: *asymmetry*. Figure 8.7 depicts a situation in which risk-premiums are countercyclical and, also, asymmetric, in that they decrease less in good times than they increase in bad, consistently with the empirical evidence.

Suppose that the economy enters a boom, in which case risk-premiums decrease, and asset prices increase, on average. During the boom, when the economy is hit by positive shocks on the fundamentals, risk-premiums decrease and asset prices increase. However, risk-premiums (and, hence, asset prices) do not change as they would during a recession—we are assuming that risk-premiums behave asymmetrically over the business cycle. Eventually, the boom ends and a recession begins. As the economy leads to a recession, risk-premiums increase and asset prices decrease. Yet, now, the negative shocks on the fundamentals lead risk-premiums to increase (and, hence, asset prices decrease) more than they moved during the boom. All in all, volatility increases on the downside. Once again, this asymmetric behavior occurs due to the assumption that risk-premiums change asymmetrically over the business cycle.[4]

[4]Mele (2007) develops a no-arbitrage framework to deal with these countercyclical issues, on which the next section is based.

The empirical evidence in Table 8.1 is supportive of these risk-premiums channel: expected returns seem to move more during recessions than during expansions. Figure 8.8 connects this asymmetric behavior of the expected returns with short-run macroeconomic fluctuations. It depicts how expected returns relate to the monthly industrial production growth, according to whether the U.S. economy is in a booming or a recessionary phase.

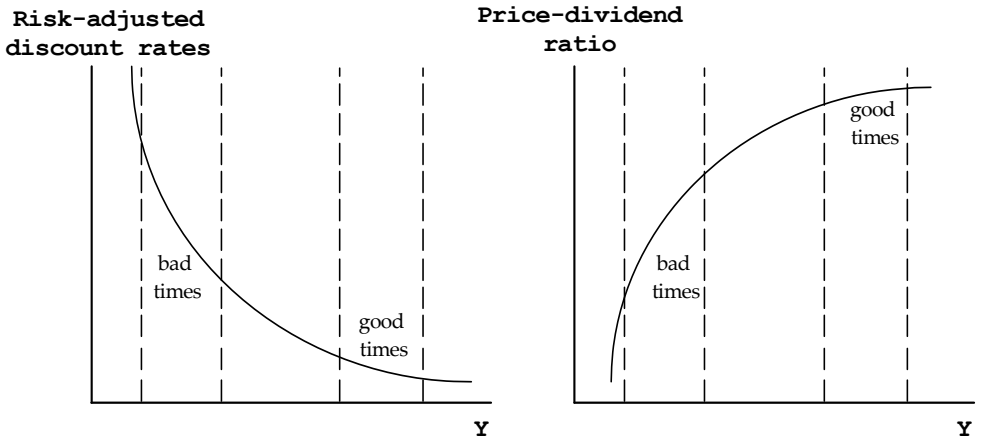


FIGURE 8.7. Countercyclical risk premiums and equity volatility

To summarize, if risk-premiums are more volatile during recessions than booms, asset prices and price-dividend ratios are more responsive to changes in economic conditions in bad times than in good, thereby leading to countercyclical volatility. These effects are precisely those we observe, as explained. The next section develops theoretical foundations to formalize these links. A key result is that countercyclical volatility is likely to arise in many models, provided the previous asymmetry in discounting is sufficiently strong. More precisely, if the asymmetry in discounting is sufficiently strong in relation to some benchmark variable tracking the business cycle conditions, the price-dividend ratio is, then, increasing and *concave* in these variables. It is this concavity feature to make stock volatility increase on the downside.

Section 8.4 provides a more comprehensive explanation of these facts, by relying on a fairly general continuous-time framework and tools relatively unusual in economics, although the intuition is still the same as that of Figure 8.7. The scope of this section is to provide a quantitative illustration of these results, based on a simple binomial tree model, which is solved in closed-form, and shown to predict a few of the stylized features of the aggregate market, surveyed in Section 8.2. Section 8.6 provides additional models that help understand the empirical evidence.

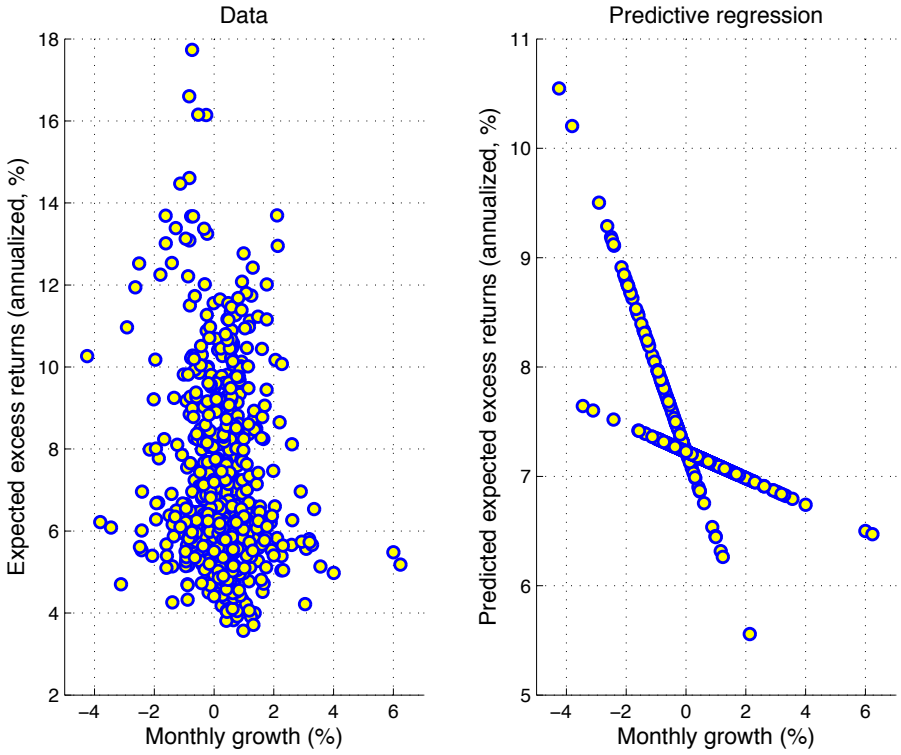


FIGURE 8.8. Expected returns and business cycle conditions. The left panel plots expected excess returns, $\hat{\mathcal{E}}_t$ in Eq. (8.2) against real (deseasoned) monthly industrial production growth, Ind_t. The right panel depicts the prediction of the ordinary least squares regression: $\hat{\mathcal{E}}_t = \underset{(0.099)}{7.225} - (\underset{(0.174)}{0.782} \cdot \mathbb{I}_{\text{Rec}_t} + \underset{(0.103)}{0.121} \cdot \mathbb{I}_{\text{Exp}_t}) \times \text{Ind}_t + w_t$, where $\mathbb{I}_{\text{Rec}_t}$ (resp., $\mathbb{I}_{\text{Exp}_t}$) is the indicator functions that takes the value one if the economy is in a NBER-recession (resp. expansion) episode and zero otherwise, w_t is a residual term, and standard errors are in parenthesis.

8.3.2.2 A simple model of countercyclical volatility

Consider an infinite horizon economy, a single asset, and a representative agent. In equilibrium, the agent consumes the dividends promised by asset. We also assume a safe asset is infinitely elastically supplied, such that the interest rate is some constant $r > 0$. In the initial state, the dividend D is equal to one (see Figure 8.9). In the second period, $D = e^{-\delta}$ ($\delta > 0$) with prob p (the bad state) or $D = e^{\delta}$ with prob $1 - p$ (the good state).

In the initial state, the agent's CRRA is $\gamma > 0$. In the good (resp., the bad) state, the agent's RRA is γ_G (resp., γ_B) > 0. In the third period, the agent receives the final payoffs of Figure 8.9, where M_S is the price of a claim to all future dividends, discounted at a RRA γ_S, with $S \in \{G, B, GB\}$ and $\gamma_{GB} = \gamma$.

This model is thus one with constant expected dividend growth, but random risk-aversion. Note that risk-aversion is being a source of "long-run risk"—once this risk is resolved, risk-aversion remains fixed at its level forever.[5]

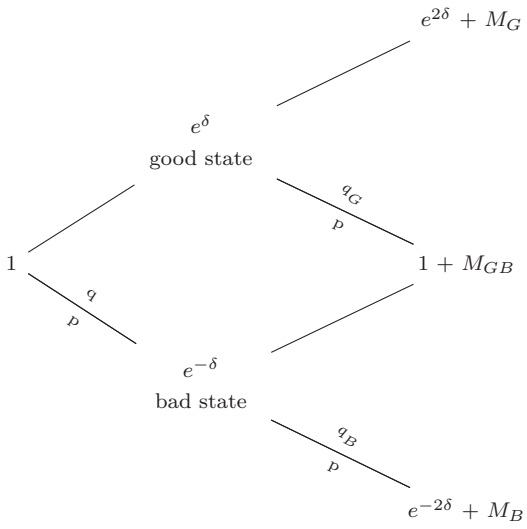


FIGURE 8.9. A model of random risk-aversion and countercyclical volatility. The dividend D is normalized to one at the initial node of the tree. With prob p, D then decreases to $e^{-\delta}$ in the bad state. The risk-neutral probability of this state is denoted as q. The risk-neutral probability of further dividend movements depends upon whether the economy is in the good or bad state (i.e., q_G or q_B). At time 3, the agent receives the dividends plus the right to the stream of all future dividends. In the upper node, this right is worth M_G (obtained through the risk-neutral probability q_G). In the central node, it is worth M_{GB} (through the risk-neutral probability q). In the lower node, it is worth M_B (through the risk-neutral probability q_B).

Table 8.2 provides calibration results for this model, based on the statistics of Table 8.1 (see Appendix 2 for details). Note that the model is calibrated using data for the aggregate dividend growth (not consumption), the volatility of which is about 6% annualized (almost

[5]The literature on long-run risks is surveyed in Chapter 9. In this literature, the expected dividend growth is affected by some unobservable and persistent factor, which generates countercyclical stock volatility, due to the assumption that dividend growth and consumption volatility are countercyclical. The model in this section leads to countercyclical volatility without the assumption that the volatility of the fundamentals is countercyclical.

twice that on consumption growth). In the Lucas' model of the previous chapter (Section 7.2), this figure would imply a RRA equal to $17 \approx \frac{0.06}{0.06^2}$ to match an equity premium of 6%. However, the Lucas model would predict that return volatility is simply dividend growth volatility, thus being equal to 6%, which is less than a half of the average volatility in the data, 14.55%. Finally the Lucas' model predicts the price-dividend ratio is constant, and therefore cannot lead to the countercyclical statistics in Table 8.1.

	Data		
	expansions	average	recessions
P/D ratio	33.21	31.99	26.20
excess return volatility	14.05	14.55	16.91
	Model calibration		
	good state	average	bad state
P/D ratio	32.50	31.81	28.15
excess return volatility	7.29	8.20	13.03
risk-adjusted rate	8.95	9.07	9.71
expected returns	10.16	11.46	18.42
implied risk-aversion	13.69	13.89	14.96

TABLE 8.2. This table reports calibration results for the infinite horizon tree model in Figure 8.9, by perfectly matching the model-implied time-3 P/D ratios to the data in the first row. The model-implied statistics in the good, average and bad state are those of time-2. The risk-adjusted rate is $r + \hat{\sigma}_0 \lambda_S$, where r is the riskless rate; $\hat{\sigma}_0$ is dividend growth volatility, and λ_S is the Sharpe ratio on gross returns in state S, determined as $\lambda_S \equiv (q_S - p) \div \sqrt{p(1-p)}$ for $S = G$ (the good state) and $S = B$ (the bad state). Finally, p is the probability of the bad state and q_S is the state-dependent risk-adjusted probability of the bad state (for $S \in \{G, B\}$). Implied risk-aversion is the coefficient RRA γ_S in the good state ($S = G$) and in the bad ($S = B$) implied by the calibrated model. The figures in the "average" column are the averages of the corresponding values in the good and bad states, averages taken under $p = 0.158$.

The model predicts that the average excess return volatility equals about 8%. Moreover, the average implied RRA is now around 13, and the average expected excess returns are high. Finally, the model of this section predicts stock volatility has swings that mimic those in the data, with levels reaching 13% in the bad state. In the bad state, however, the model overstates the expected returns by a few percentage points. Importantly, this calibration exercise illustrates the asymmetric feature of expected returns and risk aversion. In this experiment, both expected returns and risk-aversion increase more in bad times than they decrease in good.

8.3.2.3 Alternative channels

There are at least two broad mechanisms to explain aggregate stock market fluctuations, as expalined in the Introduction: (i) time-varying risk-premiums (as in this section), and (ii) time-varying expected dividend growth. Section 8.5 surveys more elaborated models than the tree of this section, aiming to rationalize countercyclical statistics based on the first channel. We

shall also examine models of "learning," in which stock volatility is time varying due to the agents' attempt to learn about the state of the economy—these models predict agents face time-varying expected dividend growth. In the next chapter, we survey many models relying on both channels, each of them focussing on particular economic mechanisms and specific predictions (e.g., idiosyncratic risk, restricted stock market participation, heterogeneous beliefs, bubbles).

8.3.3 What to do with stock market volatility?

The historical behavior of equity volatility displays a pronounced business cycle pattern. Could we exploit this pattern for the purpose of forecasting? This section considers two exercises.

In the first exercise, we forecast stock market volatility using past macroeconomic data. Stock volatility is an input to many decision making processes, ranging from portfolio selection to risk-management; understanding how it links to ongoing business cycle conditions is therefore a natural exercise. The second exercise explores whether volatility helps predict economic activity. This exercise might help decision makers (e.g., policy makers) take informed decisions.

8.3.3.1 Macroeconomic constituents of stock market volatility

Table 8.3 reports results regarding the first forecasting exercise. How does volatility link to *past* macroeconomic data? We use year-to-year industrial production growth and inflation as the two macroeconomic factors that summarize the state of the economy at any given point in time. Volatility is positively related to past growth in the medium term (say between one and two years), a finding we can easily interpret. Bad times are followed by good. Because stock market volatility is countercyclical, high growth is followed by high volatility. These explanations are similar to those put forward in Section 8.2 while elaborating on Eq. (8.4). Equity volatility is also related to past inflation, but in a more complex manner.

Figure 8.10 (top panel) depicts equity volatility and its in-sample forecasts when the regression model is fed with *past* macroeconomic data only. Naturally, the fit could be improved by providing the model with information about both past volatility and past macroeconomic data. Nevertheless, it is remarkable that the fit relying on past macro information is more than 60% better than that relying on past volatility only, as witnessed by the R^2s in Table 8.3.

Note that these results are not inconsistent with those reported by Schwert (1989). Indeed, this section relies on estimates regarding a frequency scale lower than those investigated by Schwert. More importantly, these estimate regard the linkages between stock market volatility and the *level* of macroeconomic variables, not their volatility.

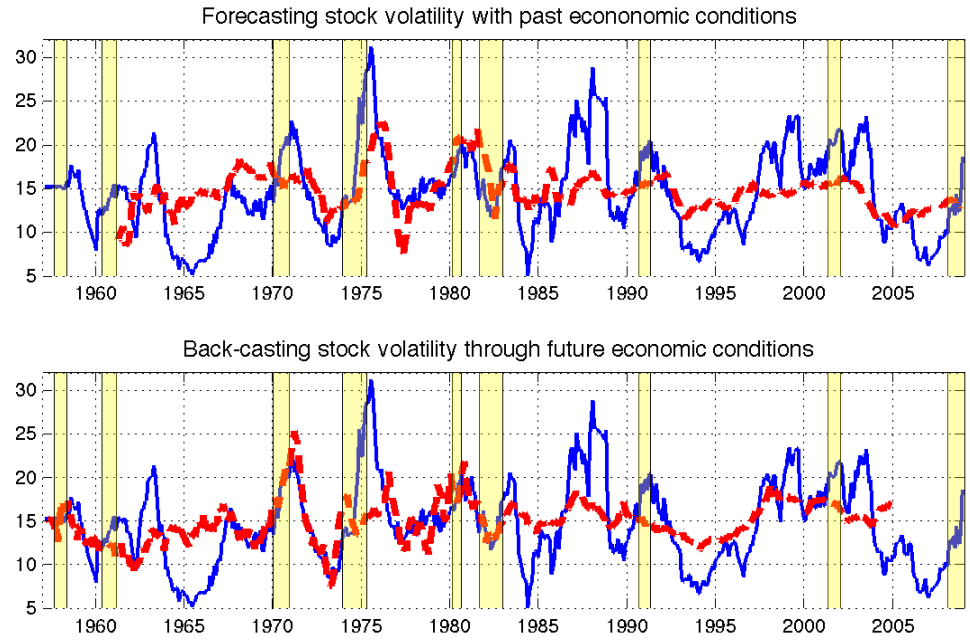


FIGURE 8.10. Predictions regarding equity market volatility. The top panel depicts stock market volatility (solid line) and its forecasts based on the sole use of *past* macroeconomic indicators (dashed), i.e., the model estimates in the second column of Table 8.3 ("Past"). The bottom panel depicts stock market volatility and its prediction based on the realization of *future* values of macroeconomic indicators, i.e., the model estimates in the third column of Table 8.3 ("Future"). Shaded areas are NBER recession and expansion episodes.

The previous findings should not be interpreted as suggesting any causality link; they could be best regarded as "descriptive statistics." They do suggest, however, that stock market volatility links to past macroeconomic developments. A natural question is how precisely it does. Once again, the previous regressions capture mere statistical relations. Yet macroeconomic factors are likely part of the evaluation leading to the very same volatility. In terms of our daily jargon, macroeconomic factors could well be determinants of the pricing kernel. That is, stock volatility links to how the price responds to shocks in the fundamentals and, hence, macroeconomic conditions, but this linkage should be determined in absence of arbitrage.

Corradi, Distaso and Mele (2013) pursue this topic in detail and build up a no-arb model that reproduces the previous predictability results. In their model, there is a no-arbitrage nexus between equity volatility and macroeconomic factors. Christiansen, Schmeling and Schrimpf (2012) and Paye (2012) provide evidence of Granger causality from past values of several macroeconomic variables to stock volatility, in out of sample experiments. However, Paye notes that we are still not able to exploit these linkages for forecasting purposes. It is an important result, as it points to the possibility that in the future, alternative data sets could do a better job than the datasets these authors are using. We make this point in Section 8.3.3.3.

The distinction between Granger causality and forecasting accuracy is indeed subtle. A set of variables could well affect the probability distribution of stock volatility—this is the definition

of Granger causality. At the same time, estimating, say, a linear regression linking past macroeconomic variables to stock volatility might not necessarily perform well. Intuitively, this relation can be subject to parameter estimation error, which increases the uncertainty sorrounding the forecasts. This uncertainty might overwhelm bias reduction gains brought by a correctly specified model, i.e., without omitted variables (macroeconomic variables). We illustrate this point in more detail below (see Section 8.3.3.4).[6]

	Past				*Future*	
Const.	10.98	10.81	4.79		Const.	17.88
Growth_{t-12}	–	−0.15	−0.002*		Growth_{t+12}	0.02
Growth_{t-24}	–	0.16	0.24		Growth_{t+24}	−0.15
Growth_{t-36}	–	0.27	0.27		Growth_{t+36}	−0.43
Growth_{t-48}	–	0.23	0.20		Growth_{t+48}	−0.31
Infl_{t-12}	–	0.76	0.63		Infl_{t+12}	1.12
Infl_{t-24}	–	−0.97	−0.92		Infl_{t+24}	−1.09
Infl_{t-36}	–	−0.45	0.62		Infl_{t+36}	−1.03
Infl_{t-48}	–	0.31	0.15*		Infl_{t+48}	0.94
Vol_{t-12}	0.36	–	0.28			
Vol_{t-24}	−0.09	–	0.03*			
Vol_{t-36}	−0.10	–	−0.02*			
Vol_{t-48}	0.08	–	0.09			
R^2	12.50	21.91	27.24		R^2	24.30

TABLE 8.3. Forecasting stock market volatility with economic activity. The left part of this table ("Past") reports OLS estimates in linear regression of one year volatility (in %) on to, *past* one year industrial production growth (in %), *past* one year month inflation (in %), and *past* stock volatility. $\text{Growth}_{t-\Delta}$ is one year industrial production growth at time $t-\Delta$, etc. Time units are months. The second part of the table ("Future") is similar, but contains coefficient estimates in linear regressions of volatility on to *future* industrial production growth and *future* inflation. Starred figures are not statistically distinguishable from zero at the 95% level. R^2 is the percentage, adjusted R^2.

8.3.3.2 Macroeconomic implications of stock market volatility

Does equity volatility also anticipate the business cycle? Table 8.3 suggests that stock volatility does indeed link to future business cycle developments. The bottom panel of Figure 8.10 depicts the predicting part of the regression in the third column of Table 8.3, a "back-casting" exercise. Fornari and Mele (2013) have actually tackled this issue in great detail, concluding that stock volatility does quite help predict the business cycle, on top of traditional indicators such as the term spread and other financial variables, both in sample and out of sample.

To illustrate, note that not only is stock volatility countercyclical, i.e., a coincident business cycle indicator. Figures 8.2 and 8.10 also seem to indicate that stock volatility tends to increase *before* recessions, a typical attribute of a *leading* indicator. Consider the following regression:

$$\sigma_t = c + \sum\nolimits_{i\in\{3,12,24,36\}} b_i\sigma_{t-i} + c_1\mathbb{I}_{t\in\mathcal{O}(\text{NBER}_t=1)} + c_2\mathbb{I}_{\text{NBER}_t=1} + \epsilon_t, \tag{8.7}$$

[6]The literature on statistical tests for Granger causality and forecasting accuracy is large. See, e.g., Clark and West (2007) for the former and Giacomini and White (2006) for the latter.

where σ_t is stock volatility at month t; $\mathbb{I}_{t\in\mathcal{O}(\text{NBER}_t=1)}$ is the indicator function that equals one in the twelve months preceding any NBER-dated recession, and zero otherwise; $\mathbb{I}_{\text{NBER}_t=1}$ is the indicator function that equals one during any NBER-dated recession, and zero otherwise; finally, ϵ_t is a residual term.

Table 8.4 reports estimates of this model parameters on a sample covering monthly data from January 1957 to September 2008. The table reports estimates for the whole sample, and two subsamples, one before the "Great Moderation," i.e., up to 1982, and another, covering the Great Moderation and ending in 2008. A value $c_1 > 0$ is indicative that stock volatility increases ahead of recessions and a value $c_2 > 0$ indicates that stock volatility also increases during recessions.

	c	b_3	b_{12}	b_{24}	b_{36}	c_1	c_2
1957-2008	3.11	0.94	-0.15	-0.01^*	-0.01^*	0.48	1.51
1957-1982	3.60	0.98	-0.24	0.02^*	-0.04	0.34^*	1.87
1983-2008	2.88	0.94	-0.09	-0.05^*	-0.01^*	1.01	1.22

TABLE 8.4. This table reports ordinary least squares estimates of the parameters in Eq. (8.7) for the post-War data and the two subsamples (i) prior and over (ii) the Great Moderation. Starred figures are not statistically distinguishable from zero at the 95% level.

It appears that especially during the Great Moderation, stock volatility does anticipate economic downturns. This issue is indeed quite a delicate one. The fact stock volatility is countercyclical does not necessarily imply it anticipates real economic activity. And even if it could, there would remain to know whether a sustained stock market volatility could really create the premises for future economic slowdowns.

Post hoc ergo propter hoc? Does aggregate stock market volatility affect investment decisions in the real sphere of the economy? Or, rather, does volatility help predict the business cycle? The policy implications of these issues are quite obvious. If volatility merely anticipates, without affecting, the business cycle, there is little policy makers can do about it, even if its forecasting power is obviously interesting per se. These themes are still unexplored at the time of writing.

[However, survey the recent "volatility paradox" ideas]

8.3.3.3 Forecasting with the wrong model

The results in this section are in-sample. It may turn out that real-time forecasts could be disappointing. One reason could be data-snooping: if we regress a variable of interest over thousands, there is a considerable chance that at least one out of those thousands nicely links to the endogenous variable, and displays a spectacular fit (in-sample). However, precisely because this fit was obtained only by chance, and not due to an economic linkages between this variable and the endogenous one, the out-of-sample performance of the model will likely disappoint.

An opposite situation can actually occur, in which a link between two variables really exists, which cannot be properly exploited for practical forecasting purposes. The intuitive reason for this difficulty is limited data. That is, we can only estimate a linkage between two variables by relying on a finite sample. Yet the finite-sample bias in the linkage estimates could turn out to be substantial, and lead to large forecasting errors. Consider the following example, a data generating process in which a variable x_t Granger causes a second one, y_t, as follows:

$$y_t = c + \beta x_t + \epsilon_t, \quad \epsilon_t \sim \text{NID}\left(0, \omega_\epsilon\right), \quad x_t \sim \text{NID}\left(\mu_x, \omega_x\right), \tag{8.8}$$

for five constants c, β, ω_ϵ, μ_x and ω_x, the parameters of the model. We assume that μ_x and ω_x are known, and consider making predictions of the variable y_t through two models.

The first model is misspecified, in that we simply neglect that x_t Granger causes y_t, i.e. $y_t = \psi + u_t$, for some constant ψ and some residual term u_t. We estimate the constant ψ of this misspecified model through ordinary least squares (OLS), obtaining:

$$\hat{\psi} - c = \beta \bar{x}_T + \bar{\epsilon}_T,$$

where $\bar{x}_T$ and $\bar{\epsilon}_T$ denote the sample averages of x_t and ϵ_t, and T is the sample size. The prediction error generated by this model for time $T+1$ is:

$$\eta_{1,T+1} \equiv y_{T+1} - \hat{\psi} = \beta \left(x_{T+1} - \bar{x}_T \right) + \epsilon_{T+1} - \bar{\epsilon}_T.$$

Note that although $\hat{\psi}$ is biased for c, even asymptotically, the predictor of this misspecified model is unbiased because $E\left(\eta_{1,T+1}\right) = 0$.

Next, condider using as a predictor, the predictive part of Eq. (8.8), obtained through the OLS estimators of c and β, say $\hat{c}$ and $\hat{\beta}$. The resulting prediction error is,

$$\begin{aligned} \eta_{2,T+1} &\equiv y_{T+1} - \hat{c} - \hat{\beta} x_{T+1} \\ &= c - \hat{c} + (\beta - \hat{\beta}) x_{T+1} + \epsilon_{T+1} \\ &= (\beta - \hat{\beta}) \left(x_{T+1} - \bar{x}_T \right) + \epsilon_{T+1} - \bar{\epsilon}_T, \end{aligned} \tag{8.9}$$

where the second equality follows by (i) $\hat{c} = \bar{y}_T - \hat{\beta} \bar{x}_T$, with $\bar{y}_T$ denoting the sample average of y_t, and (ii) Eq. (8.8), and:

$$\beta - \hat{\beta} = -\frac{Cov_T\left(x_t, \epsilon_t\right)}{V_T\left(x_t\right)},$$

and Cov_T and V_T stand for the sample covariance and variance of their arguments. The correctly specified model does, naturally, lead to an unbiased predictor, in that $E\left(\eta_{2,T+1}\right) = 0$, by the second line in Eq. (8.9).

Therefore, the two models we consider (the misspecified and the correctly specified) both lead to unbiased predictors. However, the second predictor is plagued by parameter estimation error, and might actually lead to mean-squared prediction errors higher than those generated by the first predictor, especially when the sample variance of $\beta - \hat{\beta}$ is large. In other words, for large samples, $\beta - \hat{\beta}$ is, of course, quite small, as $\hat{\beta}$ is consistent for β. In finite samples, however, this term can adversely affect the performance of the correctly specified model.

8.3.4 What did we learn?

Stock market volatility is higher in bad times than in good. Explaining this basic fact is challenging. We know very well how to model risk-premiums and how these premiums should relate to the business cycle. We are more embarrassed when it comes to explain volatility. This section explains that countercyclical volatility could arise because risk-premiums undergo large swings as the economy moves away from good states, just as the data seem to suggest.

The focus in this section relates to the *fluctuations* of aggregate stock volatility and risk premiums, not their *average levels*. Not suprisingly, the question whether these fluctuations (and their average levels) can be consistent with the neo-classical model of rational evaluation is controversial, as for many topics at the intersection of financial economics and macroeconomics,

as vividly illustrated in the early debates (see, e.g., Campbell, 2003; Mehra and Prescott, 2003). However, this section suggests that there is a potential for explaining the swings that aggregate stock volatility experiences across states of nature.

Do these theoretical insights have some additional empirical content? This section has discussed three empirical issues: (ii) the market expected returns are strongly countercyclical and asymmetrically related to macoreconomic conditions; (ii) equity volatility links to the business cycle (in-sample), although it cannot necessarily be forecast through macroeconomic variables, out-of-sample; (iii) equity volatility contains information regarding business cycle developments. We now turn to more theoretically-based explanations of the aggregate stock market fluctuations.

8.4 Rational market fluctuations

What would be needed to make rational valuation consistent with the countercyclical statistics described so far? We rely on a parsimonious framework and explain that the price-dividend ratio should play a critical role, by reacting asymmetrically to shocks in the fundamentals. Section 8.4.1 provides a decomposition of the asset returns. Section 8.4.2 develops tools of analysis with examples. In Section 8.5, we apply these tools to address the empirical issues.

8.4.1 The dynamics of asset returns

8.4.1.1 A return decomposition

Asset returns are the sum of payoffs *and* capital gains (prices). Consider the identity $\tilde{R}_{t+1} \equiv \frac{S_{t+1}+D_{t+1}}{S_t}$, or

$$\ln \tilde{R}_{t+1} \equiv g_{t+1} + \ln\left(\frac{p_{t+1}+1}{p_t}\right), \tag{8.10}$$

where $g_t \equiv \ln\left(\frac{D_t}{D_{t-1}}\right)$, the dividend growth, and $p_t \equiv \frac{S_t}{D_t}$, the price-dividend ratio. Thus, return volatility can be countercyclical because dividend growth and/or price-dividend ratio changes have countercyclical volatility.

The empirical evidence in Section 8.2 suggests it is unlikely that, at the business cycle frequency, return volatility links to the volatility of the fundamentals. Instead, empirical evidence suggests that a model should make at least two predictions regarding the price-dividend ratio. First, it needs to be volatile, and second, it needs to be more volatile in bad times than in good. For example, in an economy driven by a state variable linked to the business cycle, such as habit formation (see Section 8.5), we would require that the price-dividend ratio be increasing and *concave* in the business cycle variable, as previously explained (see Section 8.3.2). Intuitively, this property ensures stock volatility increases on the downside—the very definition of countercyclical volatility. This section aims to provide conditions under which price-dividend ratios behave in this way.

8.4.1.2 Asymmetric behavior of the price-dividend ratio

Do price multiples behave asymmetrically over the business cycle? Empirically, they do, as explained. We now rely on a simple continuous time model that leads to these asymmetric properties. We assume that dividend growth is i.i.d.,

$$\frac{dD_\tau}{D_\tau} = g_0 d\tau + \sigma_0 dW_{1\tau},$$

where g_0 and σ_0 are two constants.

We assume that the *risk-adjusted discount rates* are a function of some state variable summarizing the business cycle conditions, y_τ say. That is, and relying on notation introduced in Chapter 5 (Section 5.2.5), we assume that

$$\mathcal{R}(y_t) = r(y_t) + \sigma_0 \lambda_{\text{CF}}(y_t),$$

where $r(\cdot)$ is the short-term rate and $\lambda_{\text{CF}}(\cdot)$ is the cash-flow lambda. In the next sections and in the next chapter, we explain how agents' preferences and beliefs can lead to these discount rates. We assume that y_τ is solution to

$$dy_\tau = \varphi(y_\tau)\, d\tau + v_1(y_\tau)\, dW_{1\tau} + v_2(y_\tau)\, dW_{2\tau},$$

for some functions φ and v_i.

Given these assumptions, and under regularity conditions, the price-dividend ratio is driven by y_τ, as shown by the following expression

$$p(y_t) = \mathbb{E}\left(\int_t^\infty \frac{D_\tau^*}{D_t} \cdot e^{-\int_t^\tau \mathcal{R}(y_u) du} d\tau \,\middle|\, y_t \right), \quad \frac{D_\tau^*}{D_t} = e^{\left(g_0 - \frac{1}{2}\sigma_0^2\right)(\tau - t) + \sigma_0\left(\hat{W}_{1\tau} - \hat{W}_{1t}\right)}, \tag{8.11}$$

where $\hat{W}_{1t}$ is a standard Brownian motion under the risk-neutral probability Q, and $\mathbb{E}$ denotes the expectation under Q.

Eq. (8.11) is derived in Chapter 5 (Section 5.2.5). It suggests that the sensitivity of p with respect to y is related to the sensitivity of the risk-adjusted discount rate $\mathcal{R}$ with respect to y. Hence, whether volatility is countercyclical now depends on how the risk-adjusted discount rates change after shocks in y.

This section formalizes the previous intuitive reasoning. It shows that if $\mathcal{R}$ increases in bad times sufficiently more than it does in good, the price-dividend ratio is concave in y, thereby being more volatile in bad times than in good. This property would be consistent with the empirical evidence on the behavior of price-dividend ratio and volatility. Chapter 7 explains that additional state variables to dividends are needed to drive fluctuations in the price-dividend ratio; at the same time, Chapter 7 also explains, multifactor models are not necessarily satisfactory: there exist multifactor models such that (i) the variance of the pricing kernel increases arbitrarily with the number of factors, and yet (ii) price-dividend ratios are constant. What we really need is a discipline on how to increase the dimension of a model.

In the remainder of the chapter, we focus on two broad but key properties of models consistent with this search process: (i) "monotonicity" and (ii) "convexity" properties:

(i) *Monotonicity.* Consider the price-dividend ratio in Eq. (8.11). By Itô's lemma, stock volatility is $\sigma_0 + \frac{p'(y)}{p(y)}\text{Vol}(y)$, where $\text{Vol}(y) = \sqrt{v_1^2(y) + v_2^2(y)}$ is the volatility of y. Therefore, y can help inflate volatility if p increases with y. In turn, this monotonicity would ensure that stock volatility is strictly positive and, then, that agents' budget constraints are well-defined.

(ii.1) *Negative convexity.* Suppose as before that y correlates with the business cycle. If $\text{Vol}(y)$ is constant, stock volatility is countercyclical whenever p in Eq. (8.11) is *concave* in y, as in the simple reasoning underlying Figure 8.8. We shall study this point in detail in Section 8.5.3.

(ii.2) *Convexity.* Alternatively, suppose that expected dividend growth, g say, is stochastic (an assumption we explore in detail in Section 8.5). We shall explain that under conditions, the price-dividend ratio is a function of g, similarly as with Eq. (8.11). Now, suppose p is increasing and *convex* in g. In this case, the price-dividend ratio would displays "overreaction" to small changes in g in good times, i.e. when g is high. The empirical relevance of this point was first acknowledged by Barsky and De Long (1990, 1993), and formalized by Veronesi (1999) in a model with learning (see Section 8.5.4).

We now introduce a framework to study these issues. We need to revisit the option pricing literature on convexity of option prices and extend it to contexts with untraded risks. Chapter 11 contains additional explanations regarding these general properties of option prices that go beyond those needed for the purpose of this chapter.

8.4.2 Stocks as options

Chapter 1 (Section 1.4.4) illustrates the classic static example of a two-period market for a cash $\psi(\tilde{y})$ for some random variable $\tilde{y}$ to be paid in the second period. Assuming interest rates are zero, the fair value for this claim is $\bar{c} \equiv \mathbb{E}[\psi(\tilde{y})]$ be the premium. Appendix 5 to this chapter considers a dynamic extension of this problem in which $\tilde{y} = y_T$ for some future date T, and y is a random process, with $y_0 = y$, such that the price of the claim is now

$$c(y) = \mathbb{E}(\psi(y_T)|\, y). \tag{8.12}$$

Clearly, the two pricing problems, $\mathbb{E}(\psi(y_T)|\, y)$ and $\mathbb{E}(\psi(\tilde{y}))$, are not the same. They actually bear similarities if (i) y is the price of a traded asset; and (ii) y is a proportional process—one for which the risk-neutral distribution of $\tilde{g} \equiv \frac{y_T}{y}$ is independent of y. If these assumptions hold, the usual analysis of the static case still applies to this dynamic case. In particular, c increases after a mean-preserving spread in $\tilde{g}$ whenever ψ is convex.[7] We now examine cases in which these assumptions do not necessarily hold.

8.4.2.1 Volatility, options and convexity

If y_t is the price of a traded asset, which does not pay dividends, the drift of y_t under the risk-neutral probabilty is proportional to y_t. We shall clarify soon that in this case, the price c inherits the convexity properties from the final payoff only, ψ.

But there are risks that are not necessarily traded. In these markets, interesting nonlinearities arise. For example, Theorem 8.1 reveals that in this context, convexity of ψ is neither a necessary or a sufficient condition for convexity of c. The drift of y_t plays a crucial role.[8].

Let us consider the following problem. It is the benchmark for a number of pricing problems dealt with in the remainder of this chapter.

[7]This prediction is consistent with the celebrated Black and Scholes (1973) formula, as we further explain in Chapter 10—a point made by Jagannathan (1984, p. 429-430). As further explained in Chapter 11, Bajeux-Besnainou and Rochet (1996, Section 5), Bergman, Grundy and Wiener (1996), El Karoui, Jeanblanc-Picqué and Shreve (1998) and Romano and Touzi (1997) generalize these results to more general diffusion models, including those with stochastic volatility.

[8]Kijima (2002) produces a counterexample where convexity of option prices might break down even payoffs are convex in the underlying, and *traded*, assets. This counterexample relies on an extension of the Black-Scholes model where due to the presence of dividends, the drift of the underlying asset is concave in the asset price. Among other things, Theorem 8.1 below unveils the origins of this counterexample.

Canonical pricing problem. *Let y_τ be the solution to*:

$$dy_\tau = b\left(y_\tau\right) d\tau + a\left(y_\tau\right) d\bar{W}_\tau, \tag{8.13}$$

where $\bar{W}$ is a multidimensional $\bar{P}$-Brownian motion (for some $\bar{P}$), and b, a are some given functions. Let ψ and ρ be two twice continuously differentiable positive functions, and define

$$c\left(y,T\right) \equiv \bar{\mathbb{E}}\left(e^{-\int_0^T \rho(y_t)dt} \cdot \psi(y_T)\middle| y\right) \tag{8.14}$$

to be the price of an asset which promises to pay $\psi(y_T)$ at time T.

A simple market encompassed by this problem is one in which y_t is the price of a traded asset, and $\bar{P} = Q$, the risk-neutral probability, such that the drift in Eq. (8.13) is $b\left(y\right) = y\rho\left(y\right)$, by no-arbitrage. If y_τ is not a traded risk, $b\left(y\right) = b_0\left(y\right) - a\left(y\right)\lambda\left(y\right)$, where b_0 is the physical drift of y_τ, and λ is a risk-premium. Our canonical pricing problem now covers a number of interesting cases. For example, we may assume $\psi\left(y\right) = 1$, $\rho\left(y\right) = y$, and y is a short-term rate, in which case c is the price of a zero-coupon bond price.

Note that $\bar{P}$ is not necessarily the risk-neutral probability. Consider the following important example, a "scale-invariant" economy, that is, an economy in which "price-multiples" are independent of the dividend level. Assume that avauilable for trading is an asset that distributes a dividend D, solution to

$$\begin{cases} \dfrac{dD_\tau}{D_\tau} = m\left(y_\tau\right) d\tau + \sigma_0 dW_{1\tau} \\ dy_\tau = \varphi\left(y_\tau\right) d\tau + v_1\left(y_\tau\right) dW_{1\tau} + v_2\left(y_\tau\right) dW_{2\tau} \end{cases} \tag{8.15}$$

where W_1 and W_2 are standard rBrownian motions. This example generalizes that in Section 8.4.1.2, in that expected dividend growth, m, can be stochastic, driven as it is by the state variable y_τ.

Furthermore, note that, in this model, the distribution of $\frac{D_\tau}{D}$ does not depend on D. Processes with this property are also known as proportional processes, or Samuelson-Merton's processes. Moreover, we assume that the short-term rate r only depends on y_τ. These assumptions imply that the price-dividend ratio is only a function of the current state, similarly as in Section 8.4.1. However, due to stochastic expected dividend growth, the expression for p is slighly more general than that in Eq. (8.11). It is:

$$p\left(y\right) = \int_0^\infty c\left(y,\tau\right) d\tau, \tag{8.16}$$

where, denoting as usual the risk-adjusted discount rate with $\mathcal{R}\left(y\right)$,

$$\begin{aligned} c\left(y,\tau\right) &\equiv \mathbb{E}\left(e^{-\int_0^\tau r(y_t)dt} \cdot \frac{D_\tau}{D}\middle| y\right) \\ &= \mathbb{E}\left[e^{\int_0^\tau m(y_t)dt} \cdot e^{-\int_0^\tau \mathcal{R}(y_u)du}\left(e^{-\frac{1}{2}\sigma_0^2\tau + \sigma_0\hat{W}_{1\tau}}\right)\middle| y\right] \\ &= \bar{\mathbb{E}}\left(e^{\int_0^\tau m(y_t)dt} \cdot e^{-\int_0^\tau \mathcal{R}(y_u)du}\middle| y\right), \end{aligned} \tag{8.17}$$

and $\mathbb{E}$ is the expectation under the risk-neutral probability (Q), under which D and y satisfy[9]

$$\begin{cases} \dfrac{dD_\tau}{D_\tau} = (m(y_\tau) - \sigma_0\lambda_1(y_\tau))\,d\tau + \sigma_0 d\hat{W}_{1\tau} \\ dy_\tau = \left(\varphi(y_\tau) - \sum_{i=1}^2 v_i(y_\tau)\,\lambda_i(y_\tau)\right)d\tau + v_1(y_\tau)\,d\hat{W}_{1\tau} + v_2(y_\tau)\,d\hat{W}_{2\tau} \end{cases}$$

where $\hat{W}_{i\tau}$ are two Brownian motions under Q. The third line of Eq. (8.17) is obtained with a change of probability, and $\bar{\mathbb{E}}$ is the expectation taken under a conveniently changed probability $\bar{Q}$, defined by the Radon-Nikodym derivative,

$$\left.\frac{d\bar{Q}}{dQ}\right|_{F_\tau} = e^{-\frac{1}{2}\sigma_0^2\tau + \sigma_0\hat{W}_{1\tau}}, \tag{8.18}$$

where F_τ denotes the information set as of time τ generated by $\hat{W}_{1\tau}$. By Girsanov theorem, we have that under $\bar{Q}$,

$$\begin{array}{c} dy_\tau = \bar{\varphi}(y_\tau)\,d\tau + v_1(y_\tau)\,d\bar{W}_{1\tau} + v_2(y_\tau)\,d\bar{W}_{2\tau} \\ \text{where } \bar{\varphi}(y) \equiv \varphi(y) - \sum_{i=1}^2 v_i(y)\,\lambda_i(y) + \sigma_0 v_1(y) \end{array} \tag{8.19}$$

and $\bar{W}_i$ are Brownian motions under $\bar{Q}$.[10]

Our canonical pricing problem allows us to analyze properties of prices relating to long-lived assets, through those relating to c in Eq. (8.17), with y as in Eq. (8.19), once we set

$$\psi(y) \equiv 1; \quad \rho(y) \equiv \mathcal{R}(y) - m(y); \quad b(y) = \bar{\varphi}(y). \tag{8.20}$$

The next theorem characterizes slope and convexity properties of the price c in the canonical pricing problem.

THEOREM 8.1. *We have*:

(i) If $\psi' > 0$, then c is increasing whenever $\rho' \leq 0$. Furthermore, if $\psi' = 0$, then c is decreasing (resp. increasing) whenever $\rho' > 0$ (resp. < 0).

(ii) If $\psi'' \leq 0$ (resp. $\psi'' \geq 0$) and c is increasing, then c is concave (resp. convex) whenever $b'' < 2\rho'$ (resp. $b'' > 2\rho'$) and $\rho'' \geq 0$ (resp. $\rho'' \leq 0$). Finally, if $b'' = 2\rho'$, c is concave (resp. convex) whenever $\psi'' < 0$ (resp. > 0) and $\rho'' \geq 0$ (resp. ≤ 0).

Theorem 8.1-(i) generalizes previous results regarding monotonicity of option prices, obtained by Bergman, Grundy and Wiener (1996). By the so-called *no-crossing property* of a diffusion, y is not decreasing in its initial condition y. Therefore, c inherits the same monotonicity features of ψ if discounting does not operate adversely. This simple observation allows us to address monotonicity properties of long-lived asset prices, as we shall see in Section 8.5.

Theorem 8.1-(ii) generalizes a number of existing results on option price convexity. First, assume that ρ is constant and that y_τ is the price of a traded asset, such that $\rho' = b'' = 0$. The last part of Theorem 8.1-(ii) then says that the convexity of ψ propagates to the convexity

[9] See, for example, Huang and Pagès (1992, Theorem 3 p. 53) and Wang (1993, Lemma 1, p. 202), for regularity conditions underlying the Feynman-Kac theorem in infinite horizon settings; and Huang and Pagès (1992, Proposition 1, p. 41) for regularity conditions ensuring that the Girsanov's theorem holds in infinite horizon settings.

[10] Note a detail to arrive to Eq. (8.17): a change in probability that allows to get rid of the $e^{-\frac{1}{2}\sigma_0^2\tau + \sigma_0\hat{W}_{1\tau}}$ term arising through a positive correlation between dy and dD. Because of this correlation, the density of y under $\bar{Q}$ is right-shifted compared to the density under Q. Mele (2005, 2007) derives this representation of the price-dividend ratio.

of c. This result reproduces the findings in the literature surveyed earlier. Theorem 8.1-(ii) characterizes convexity in a more general context. Suppose, for example, that $\psi'' = \rho' = 0$, and that y_τ is not a traded risk. Then, Theorem 8.1-(ii) suggests that c inherits the convexity of the drift of y. As a final example, Theorem 8.1-(ii) extends a result in Mele (2003) relating to bond pricing: let $\psi(y) = 1$ and $\rho(y) = y$. Accordingly, c is the price of a zero-coupon bond in a short-term rate model (see Chapter 13 for details). By Theorem 8.1-(ii), c is convex whenever $b'' < 2$ (see Appendix 6 for further details and intuition on this bounding number).

Option prices rely on both discounting and nonlinearities affecting the drift of the state variables, when the underlying fundamentals are not traded (unlike stock prices). In Section 8.5, we rely on the predictions of Theorem 8.1 and analyze the price of long-lived assets. In Appendix 5, we illustrate the gist of the proofs underlying Theorem 8.1 through a series of examples. We now proceed with applying this framework to a variety of problems regarding evaluation of long-lived assets.

8.5 Time-varying discount rates or uncertain growth?

8.5.1 Two types of economies

Asset prices are what they are due to the characteristics of the pricing kernel *and* the statistical properties of dividend growth. We focus on a simplifying assumption, namely that explanations of asset prices can elaborate on either one of the two previous channels. Accordingly, the solid arrows of Figure 8.11 illustrate that once we make assumptions regarding dividend growth (e.g. i.i.d. distributions), one can then seek for pricing kernels that are consistent with the asset prices. Conversely, once we make assumptions regarding the pricing kernel, one can ask for which statistical properties of dividend growth are needed to reconcile explanations of asset prices with the security market data (dashed arrows).

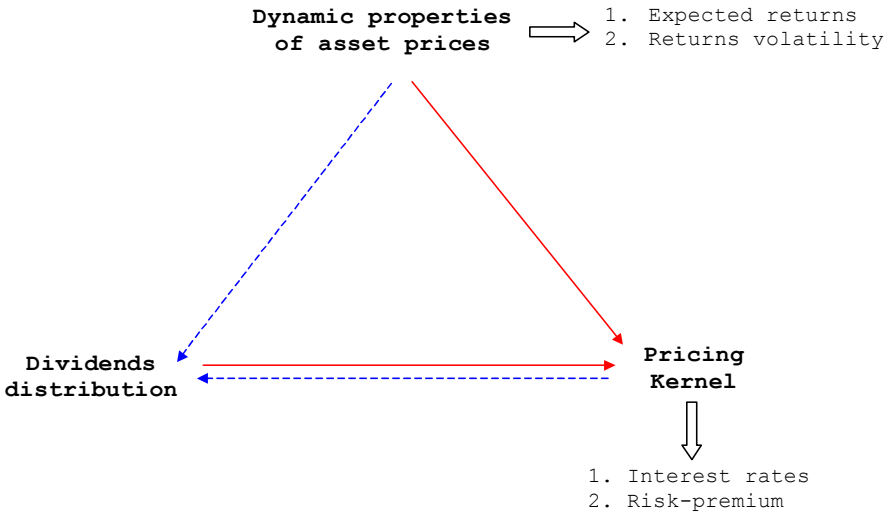


FIGURE 8.11.

This section deals with this search process while relying on methodology introduced in the previous section. We consider two economies. In the first economy, changes in the economic

fundamentals determine cyclical variations in the discount rates (in Section 8.6.1). In the second, the economic fundamentals lead to time-varying expected dividend growth (in Section 8.6.2).

We need to provide preliminary results about pricing kernels, which we need to use while illustrating these two broad classes of economies. Finally, Section 8.5.4 is an introduction to a class of hopefully analytically convenient processes we can use to model long lived asset prices.

8.5.2 Markov pricing kernels, asset returns and volatility

8.5.2.1 Pricing kernels

Motivated by the previous discussion, this section considers economies in which asset prices have high volatility due to volatile pricing kernels and stochastic expected dividend growth. We provide foundations relying on a representative agent economy. In this setting, interest rates and risk-premiums change randomly because the agent's utility depends on both consumption *and* other variables. While complete markets naturally fit in the analysis of this section (see Chapter 2), incomplete markets can sometimes be analyzed relyiong on this framework, as explained in the Chapter 9.

We extend some foundational issues in Chapter 5. Consider the stochastic discount factor in Chapter 5, $m_t(\tau) \equiv \frac{\xi_\tau}{\xi_t}$, where the pricing kernel process ξ_τ satisfies

$$\xi_\tau \equiv \xi(D_\tau, x_\tau, \tau) = e^{-\int_0^\tau \delta(D_s,x_s)ds} \Upsilon(D_\tau, x_\tau), \quad \xi_0 = 1, \tag{8.21}$$

for some functions $\delta(D,x)$ and $\Upsilon(D,x)$ and some diffusion process x_τ. We assume that δ is bounded and positive, and that Υ is twice continuously differentiable. We also assume that there is another process, y_τ, such that $y_\tau = \mathcal{Y}(x_\tau)$ for some monotonic, continuous and twice continuously differentiable function $\mathcal{Y}$, and that (D_τ, y_τ) are solutions to a slight generalization of Eqs. (8.15):

$$\begin{cases} dD_\tau = m(D_\tau, y_\tau)\, d\tau + \sigma_0(D_\tau)\, dW_{1\tau} \\ dy_\tau = \varphi(D_\tau, y_\tau)\, d\tau + v_1(D_\tau, y_\tau)\, dW_{1\tau} + v_2(D_\tau, y_\tau)\, dW_{2\tau} \end{cases} \tag{8.22}$$

The pricing kernel is solution to

$$\frac{d\xi_\tau}{\xi_\tau} = -r_\tau d\tau - \lambda_{1\tau} dW_{1\tau} - \lambda_{2\tau} dW_{2\tau}, \tag{8.23}$$

where r is the short-term rate and $[\lambda_1, \lambda_2]$ is the vector of unit risk-premiums.

By applying Itô's lemma to ξ in Eq. (8.21), and identifying the terms in Eq. (8.23), one finds that the short-term rate r and risk-premiums are both functions of the current state, in that:

$$\begin{aligned} r(D,x) &= \delta(D,x) - \tfrac{L_x \Upsilon(D,x)}{\Upsilon(D,x)} \\ \lambda_1(D,x) &= -\left(\sigma_0(D) \tfrac{\partial \ln \Upsilon(D,x)}{\partial D} + v_{x1}(D,x) \tfrac{\partial \ln \Upsilon(D,x)}{\partial x}\right) \\ \lambda_2(D,x) &= -v_{x2}(D,x) \tfrac{\partial \ln \Upsilon(D,x)}{\partial x} \end{aligned}$$

where L_x is the infinitesimal generator operator (see Chapter 5) for (D,x) and $v_{xi}(D,x)$ denote the diffusion coefficients of x_τ.

For example, consider an infinite horizon economy in which aggregate dividends are solution to Eq. (8.22), with $v_2 \equiv 0$ and $v \equiv v_1$, and a representative agent solves the following program:

$$\max_{(c_\tau)_{\tau \geq 0}} E\left(\int_0^\infty e^{-\delta\tau} u(c_\tau, x_\tau)\, d\tau\right) \quad \text{s.t} \quad V_0 = E\left(\int_0^\infty \xi_\tau c_\tau d\tau\right), \quad V_0 > 0,$$

where $\delta > 0$, the instantaneous utility u is continuous and three times continuously differentiable, and x is solution to

$$dx_\tau = \varphi_x\left(D_\tau, x_\tau\right) d\tau + v_x\left(D_\tau, x_\tau\right) dW_{1\tau}.$$

In equilibrium, optimal consumption equals aggregate dividends. In terms of ξ in Eq. (8.21), we have that $\delta\left(D, x\right) = \delta$ and $\Upsilon\left(D_\tau, x_\tau\right) = \frac{u_1(D_\tau, x_\tau)}{u_1(D_0, x_0)}$, such that $\lambda_2 = 0$, and $\lambda \equiv \lambda_1$, and

$$\begin{aligned} r(D,x) = \delta &- \frac{u_{11}\left(D,x\right)}{u_1\left(D,x\right)} m\left(D,y\right) - \frac{1}{2}\sigma_0^2\left(D\right)\frac{u_{111}\left(D,x\right)}{u_1\left(D,x\right)} \\ &- \frac{u_{12}\left(D,x\right)}{u_1\left(D,x\right)}\varphi_x\left(D,x\right) - \frac{1}{2}v_x^2\left(D,x\right)\frac{u_{122}\left(D,x\right)}{u_1\left(D,x\right)} - v_x\left(D,x\right)\sigma_0\left(D\right)\frac{u_{112}\left(D,x\right)}{u_1\left(D,x\right)} \end{aligned} \tag{8.24}$$

$$\lambda(D,x) = -\frac{u_{11}\left(D,x\right)}{u_1\left(D,x\right)}\sigma_0\left(D\right) - \frac{u_{12}\left(D,x\right)}{u_1\left(D,x\right)} v_x\left(D,x\right). \tag{8.25}$$

8.5.2.2 Expected returns and volatility under a scale-invariant assumption

We study the implications of these pricing kernels in terms of the asset expected returns and volatility. We base the derivations on the continuous time formulation of the APT model in Chapter 5 (see Section 5.2.5). These derivations rely on a *scale-invariant economy*, defined as the economy in which price-dividend ratios are independent of the dividend level. This assumption provides useful guidance for empirical work. It is easy to see that a scale-invariant economy arises in a Samuelson-Merton market, one in which the distribution of dividend growth is independent of the dividend level D: in terms of Eq. (8.22), we need to assume that (i) $m\left(D,y\right) = yD$, (ii) $\sigma_0\left(D\right) = \sigma_0 D$ (for some constant σ_0), and (iii) the drift and diffusion coefficients of y are independent of D. Furthermore, we still assume that y_t is some monotonic function $\mathcal{Y}\left(\cdot\right)$ of x_t in Eq. (8.21); accordingly, we express the price-dividend ratio in terms of y_t.

Under these assumptions, and based on derivations in Chapter 5, the asset expected returns are

$$\begin{aligned} \mathcal{E}_t &\equiv \frac{d}{d\tau} E_t\left(\frac{S_\tau}{S_t}\right)\Big|_{\tau=t} + \frac{D_t}{S_t} \\ &= \underbrace{r_t + \text{cash-flow beta}_t \times \text{cash-flow lambda}_t}_{\mathcal{R}_t \equiv \text{“risk-adjusted discount rates”}} + \underbrace{\text{price beta}_t \times \text{price lambda}_t}_{\mathcal{W}_t \equiv \text{“price premium wedge”}} \end{aligned} \tag{8.26}$$

where

$$\mathcal{R}_t = r_t + \sigma_0\lambda_1\left(y_t\right), \quad \mathcal{W}_t = \frac{p'\left(y_t\right)}{p\left(y_t\right)}\left(v_1\left(y_t\right)\lambda_1\left(y_t\right) + v_2\left(y_t\right)\lambda_2\left(y_t\right)\right).$$

Finally, returns volatility is

$$\text{Vol}\left(y_t\right) \equiv \left[\begin{array}{c} \text{Vol}_1\left(y_t\right) \\ \text{Vol}_2\left(y_t\right) \end{array}\right] \equiv \left[\begin{array}{c} \sigma_0 + v_1\left(y_t\right)\frac{p'(y_t)}{p(y_t)} \\ v_2\left(y_t\right)\frac{p'(y_t)}{p(y_t)} \end{array}\right].$$

We rely on these predictions while analyzing a number of models in the following sections as well as in Chapter 9.

8.5.3 External habit formation

Time-varying risk-premiums are a plausibly engine mechanism for asset price fluctuations. Intuitively, the very properties of asset prices must necessarily inherit those of the risk-premiums, as illustrated by Figure 8.11. Campbell and Cochrane (1999) model of external habit formation is a well-known attempt at explaining some of the empirical features outlined in Section 8.2 by incorporating time-varying risk-premiums into an economy with i.i.d. dividend growth.

Consider an infinite horizon economy in which a representative agent has undiscounted instantaneous utility:

$$u(c_\tau, x_\tau) = \frac{(c_\tau - x_\tau)^{1-\gamma} - 1}{1-\gamma}, \tag{8.27}$$

where c denotes consumption and x is a time-varying habit, or exogenous "subsistence level." In this model, the habit process is defined in a residual way, through the "surplus consumption ratio," as we now explain.

The total endowment process, D_τ, satisfies[11]

$$\frac{dD_\tau}{D_\tau} = g_0 d\tau + \sigma_0 dW_\tau. \tag{8.28}$$

A measure of distance between consumption and the level of habit is the "surplus consumption ratio," defined as

$$s_\tau \equiv \frac{c_\tau - x_\tau}{c_\tau}.$$

Note that the curvature of the instantaneous utility is inversely related to s,

$$-\frac{u_{cc}(c,x)\,c}{u_c(c,x)} = \gamma \frac{c}{c-x} = \gamma \frac{D}{D-x} \equiv \gamma s^{-1}, \tag{8.29}$$

where subscripts denote partial derivatives, the second equality is the equilibrium condition, $c_\tau = D_\tau$, and the third is the definition of the equilibrium surplus consumption ratio. By assumption, $\ln s_\tau$ is solution to:

$$d\ln s_\tau = (1-\phi)(\bar{s}_l - \ln s_\tau)d\tau + \sigma_0 l(s_\tau) dW_\tau, \tag{8.30}$$

where l is a positive function, defined below. That is, the surplus ratio s_t is driven by "output innovations" (i.e., by dW_t): the higher the output growth innovations (which lead to higher consumption in equilibrium), the higher the surplus ratio.[12]

This model of habit formation differs from previous formulations such as that of Ryder and Heal (1973), or Sundaresan and Constantinides (1990), due to three properties: (i) it is an "external" theory, in that the habit x is "aggregate," not consumption chosen by the individual, similarly as with Abel's (1990) "catching up with the Joneses" formulation, or Duesenberry's (1949) relative income model; (ii) habit responds to consumption smoothly, not to each period past consumption, as in previous models of habit formation such as that of Ferson and Constantinides (1990); (iii) it guarantees marginal utility is always positive.

[11] Campbell and Cochrane (1999) consider a discrete-time model in which the log-consumption growth is Gaussian. Eq. (8.28) is the diffusion limit of their model.

[12] One could add an additional Brownian motion in Eq. (8.30) to lower the conditional correlation between output growth and surplus ratio.

The second of the previous properties produces slow mean reversions in the price-dividend ratio and long-horizon predictability, and large predictable movements in stock volatility, three empirical features reviewed in Section 8.2.

Note that markets are complete as there is only one source of risk (the dividend in Eq. (8.28). Therefore, we can determine the Sharpe ratio in this economy relying on results in Section 8.5.2 (see Eq. (8.25)):

$$\lambda\left(D,x\right)=\frac{\gamma}{s}\left(\sigma_0-\frac{1}{D}v_x\left(D,x\right)\right), \tag{8.31}$$

where v_x is the diffusion coefficient of equilibrium habit, $x_\tau = D_\tau\left(1-s_\tau\right)$, and s_τ is solution to Eq. (8.30). By Itô's lemma, $v_x\left(D,x\right)=\left(1-s-sl(s)\right)D\sigma_0$, which replaced into Eq. (8.31), leaves:

$$\lambda(s)=\gamma\sigma_0\left(1+l(s)\right). \tag{8.32}$$

The real interest rate is, by Eq. (8.24),

$$r(s)=\delta+\gamma\left(g_0-\frac{1}{2}\sigma_0^2\right)+\gamma\left(1-\phi\right)\left(\bar{s}_l-\ln s\right)-\frac{1}{2}\gamma^2\sigma_0^2\left(1+l(s)\right)^2. \tag{8.33}$$

The third term reflects usual intertemporal substitution effects. Due to mean reversion, bad times (when s is low) are those when agents expect the very same s will improve. Therefore, in bad times, agents expect their marginal utility to decrease in the future and to compensate for this fall, they will try to decrease future consumption, compared to today, by trying to save less (or trying to borrow more), thereby pushing interest rates up. The last term is a precautionary savings term.

Campbell and Cochrane (1999) choose the function l so as to satisfy three conditions: (i) the short-term rate r is constant; and habit is predetermined both (ii) at the steady state, and (iii) near the steady state. A constant r is consistent with the empirical evidence surveyed in Section 8.2, that real interest rates are really not volatile, compared to stock returns. Making habit predetermined at and near the steady state formalizes the idea that it takes time for consumption shocks to affect habit, at least at the steady state. The Appendix shows that under these conditions, the function l is:

$$l\left(s\right)=\bar{s}^{-1}\sqrt{1+2(\bar{s}_l-\ln s)}-1, \tag{8.34}$$

where $\bar{s}=\sigma_0\sqrt{\frac{\gamma}{1-\phi}}=e^{\bar{s}_l}$. In turn, this function implies that the short-term rate in Eq. (8.33) is: $r=\delta+\gamma\left(g_0-\frac{1}{2}\sigma_0^2\right)-\frac{1}{2}\gamma(1-\phi)$.[13]

Figure 8.12 depicts the function l in Eq. (8.34), obtained using the parameter values in Campbell and Cochrane, $\gamma=2$, $\sigma_0=0.0150$, $\phi=0.870$. It is decreasing in s, and convex in s, over the empirically relevant range of variation of s.

[13]The Appendix considers a slightly more general model, in which the short-term rate is affine in $\ln s$.

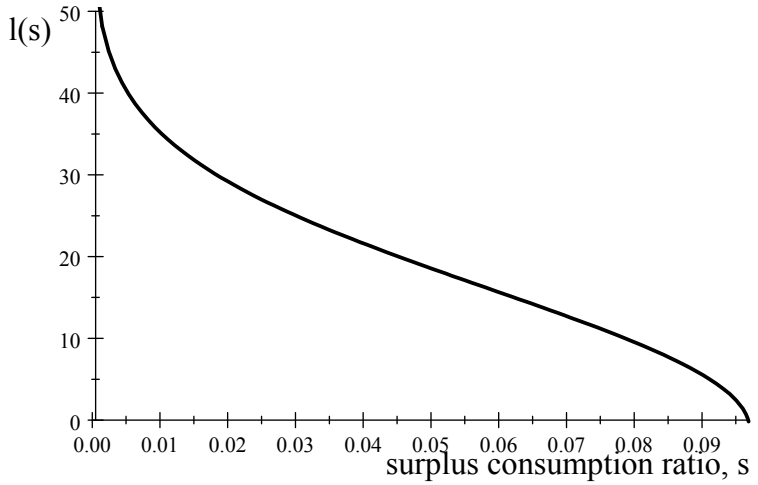


FIGURE 8.12.

These properties are inherited by the Sharpe ratio in Eq. (8.32). Precisely, the Sharpe ratio is countercyclical and moves asymmetrically over the business cycle because habit is predetermined near the steady state, and the short-term rate constant (or, at least, affine in $\ln s$, as in the Appendix).

The model makes a number of important predictions. Consider, first, the instantaneous utility in Eq. (8.27). By Eq. (8.29), relative risk aversion is equal to γs^{-1}. That is, risk aversion is countercyclical. Formally, the stochastic discount factor is $e^{-\rho t}\left(\frac{s_t}{s_0}\frac{D_t}{D_0}\right)^{-\gamma}$, $\gamma > 0$. It is countercyclical because both s and D are procyclical; moreover, it is more volatile than the standard stochastic discount factor, $e^{-\rho t}\left(\frac{D_t}{D_0}\right)^{-\gamma}$.

In terms of the expected returns, Eq. (8.26) reduces to

$$\mathcal{E}(s_t) = \mathcal{R}(s_t) + \mathcal{W}(s_t),$$

where the risk-adjusted discount rates, $\mathcal{R}(s)$, and the wedge over them, $\mathcal{W}(s)$, are given by:

$$\mathcal{R}(s) = r(s) + \sigma_0 \lambda(s), \quad \mathcal{W}(s) = \frac{p'(s)}{p(s)} v(s) \lambda(s), \tag{8.35}$$

and $v(\cdot)$ denotes the diffusion coefficient of s_t in Eq. (8.30), $v(s) = \sigma_0 s l(s)$. The mechanism is sensible. Intuitively, during economic downturns, the surplus consumption ratio s decreases and agents become more risk-averse. As a result, expected returns increase and prices decrease; moreover, the model leads to realistic risk premiums. Note that a sufficient condition for these effects to occur is that the price-dividend ratio is concave, which also ensures a countercyclical wedge, $\mathcal{W}'(s) < 0$. However, the economy is one with high risk-aversion, as on average, the calibrated model produces a value of γs^{-1} with an average value of approximately 40.

By Eq. (8.30), the log of s is a mean-reverting process. By taking logs, we are sure that s remains positive. Moreover, $\ln s$ is also conditionally heteroskedastic since its instantaneous volatility is $\sigma_0 l$. Because l is decreasing in s, and s is clearly procyclical, the volatility of $\ln s$ is countercyclical. This feature is responsible of many interesting properties of the model, such as countercyclical returns volatility.

Finally, the Sharpe ratio λ in Eq. (8.32) is made up of two components. The first is $\gamma\sigma_0$, which coincides with the Sharpe ratio predicted by the standard Gordon's (1962) model. The second is $\gamma\sigma_0 l(s)$, and arises as a compensation related to the stochastic fluctuations of the habit, $x = D(1-s)$. Therefore, λ is countercyclical due to the functional form of l. Combined with a high ϕ, this assumption leads to slowly varying, countercyclical expected returns. Finally, the model suggests that the price-dividend ratio is concave in s.[14]

We now explain the link between convexity of l and concavity of the price-dividend ratio in this model. We rely on the canonical pricing problem of Section 8.4, and appeal to Theorem 8.1. What is the price-dividend ratio in this economy? Note that the short-term rate is constant in this model, as discussed. Yet for sake of generality, assume it is state-dependent, although only a function of s (as, e.g., in the Appendix). The price-dividend ratio is then as in Eqs. (8.16)-(8.17)-(8.19), with constant growth, i.e., $m(y) = g_0$:

$$p(s) = \int_0^\infty e^{g_0\tau} \cdot \bar{\mathbb{E}}\left(e^{-\int_0^\tau \mathcal{R}(s_u)du} \middle| s_0 = s \right) d\tau, \tag{8.36}$$

where $\mathcal{R}(s)$ are the risk-adjusted discount rate in (8.35), and

$$ds_\tau = \bar{\varphi}(s_\tau)\, d\tau + v(s_\tau) d\bar{W}_\tau,$$

where

$$\bar{\varphi}(s) = \varphi(s) + (\sigma_0 - v(s))\,\lambda(s), \quad v(s) = \sigma_0 s l(s)$$
$$\varphi(s) = s\left((1-\phi)(\bar{s}_l - \ln s) + \tfrac{1}{2}\sigma_0^2 l^2(s)\right)$$

and $\bar{W}_\tau = \hat{W}_\tau - \sigma_0\tau$ is a Brownian motion under the probability $\bar{Q}$ (see Eq. (8.18)) and, finally, $\hat{W}_\tau$ is a Brownian motion under the risk-neutral probability Q.

The properties of the inner expectation in Eq. (8.36) can be analyzed relying on the canonical pricing problem of Section 8.4. Precisely, Theorem 8.1 leads to the following conclusions:

(i) *Suppose that the risk-adjusted discount rates are countercyclical, viz* $\mathcal{R}'(s) \leq 0$. *Then, the price-dividend ratio is procyclical, viz* $p'(s) > 0$.

(ii) *Suppose that the price-dividend ratio is procyclical. Then, the price-dividend ratio is also a concave function of* s *as soon as the risk-adjusted discount rates are convex in* s, *viz* $\mathcal{R}''(s) > 0$, *and* $\bar{\varphi}''(s) \leq 2\mathcal{R}'(s)$.

The previous statements impose joint restrictions on the primitives such that the price-dividend ratio is consistent with properties given in advance. The economic interpretation of the convexity of $\mathcal{R}$ is similar to that anticipated in Section 8.3 (see Figure 8.7). In terms of the Campbell-Cochrane economy, countercyclical volatility arises because $\mathcal{R}$ is decreasing and convex in the surplus consumption ratio s, such that p is concave in s.

The mechanism is the following. In bad times, consumption c_t gets close to the substistence level x_t, such that s_t is very small. Risk aversion is high as a result, and the agent becomes more reluctanct to invest in the stock market. That is, in bad times, risk-adjusted discount rates, $\mathcal{R}$, increase sharply, thus making the price-dividend ratio quite responsive to changes in the economic conditions. Instead, in good times, and again due to convexity, $\mathcal{R}$ changes relatively

[14]The solution of the model is not known in closed-form, and this property is known through a numerical solution of the model. The Appendix describes a simple method to solve the discrete-time version of this model (and related models) numerically.

less, such that the price-dividend ratio changes relatively less in response to changes in the economic conditions.

In other words, the model is such that risk aversion becomes extremely large in bad times. Technically, $\mathcal{R}$ is "sufficiently" convex in s, such that the price-dividend ratio is concave in s. Then, stock volatility increases on the downside, i.e., it is thus countercyclical, as illustrated by Figure 8.6. Note that these properties arise because the risk-adjusted rate, $\mathcal{R}$, is "sufficiently" convex in the surplus consumption ratio. The Appendix does indeed provide an upper bound to convexity that triggers these properties.[15]

One difficulty with this model is that its predictions are driven by a single state variable, the surplus consumption ratio, s_t. One implication is that the conditional correlation between consumption growth and stock returns is one. In the data, this correlation is much lower. Naturally, the model predicts this correlation is unconditionally less than one, although still too large, once compared with that in the data.

Brunnermeier and Nagel (2007) find that US investors do not change the composition of their risky asset holdings in response to changes in wealth. The authors interpret this evidence against external habit formation. Naturally, time-varying risk-premiums do not exclusively arise through external habit formation. Barberis, Huang and Santos (2001) develop a theory distinct from habit formation, which leads to time-varying risk-premiums. The next chapter explains there are many instances of economies in which risk-premiums are time-varying as a result of alternative mechanisms.

8.5.4 Large price swings as a learning induced phenomenon

We now develop models in which expected dividend growth is unobserved. In these markets, agents process available data and formulate guesses regarding the growth of the economy. These guesses lead them to face stochastic expected growth, a new state variable that has the potential to introduce interesting price dynamics. Note that while this section focusses on models with unobserved expected growth, models with observed but random expected growth have always had an interest on their own (see, e.g., the early survey of Campbell, 2003) (see Chapter 9).

8.5.4.1 The information channel

Time variation in stock volatility may also arise due to the agents' learning about the economic fundamentals. In models along these lines, public signals about the fundamentals hit the market, and agents make inference about them, thereby creating new state variables driving price fluctuations, which relate to the agents own guesses about the (unknown) state of the economic fundamentals. Timmermann (1993, 1996) provides models with exogenous discount rates and learning about the fundamentals. The effects of learning increase stock volatility beyond the extent explained by a model with known fundamentals. Brennan and Xia (2001) generalize these models to a stochastic general equilibrium. Veronesi (1999) provides a rational expectations model with learning about the fundamentals, with nonlinear effect regarding the asset price. This section provides details about the mechanisms through which learning affects asset prices in general, and stock volatility in particular.

[15] Alternatively, Mele (2007) shows that for any model in which the price-dividend ratio is driven by a diffusion variable y_t, there is a threshold y^* such that the price-dividend ratio is concave for all $y_t \leq y^*$ whenever $\lim_{y\to 0} \mathcal{R}(y) = \infty$. Note indeed that the Campbell-Cochrane model fails to satisfy restrictions (i) and (ii) over the entire range of variation of s, although then it satisfies $\lim_{s\to 0} \mathcal{R}(s) = \infty$. There exists additional models with external habit formation that lead to countercyclical volatility (see, for example, Menzly, Santos and Veronesi, 2004; Mele, 2007).

We shall assume that information about the fundamentals is incomplete, but symmetrically distributed among agents. The assumption of symmetric information might appear strong. It should not. The models in this section aim to capture the idea that markets function in a context of "incompressible" uncertainty, where agents are all unaware of the crucial aggregate, macroeconomic developments affecting asset prices. Chapter 10 reviews models with both differential and asymmetric information, which are more useful whilst thinking about the functioning of markets for individual stocks. In these markets, it is plausible to assume that agents have different information sets, and that acquire information in dedicated information markets. By contrast, it seems unrealistic to assume that one could acquire crucial information about ongoing business cycle developments and that agents are, then, asymmetrically informed about it, such that uninformed agents can learn from the asset prices: the cost of acquiring such information appears to be incommensurable.

Note that the assumption of symmetric information simplifies the analysis, as the agents do not need to base their decisions upon the observation of the equilibrium price. In a context with asymmetric information, agents can, instead, learn pieces of information other agents have, by "reading the equilibrium price," because agents with superior information impinge part of their information on the asset price, through trading, as explained in Chapter 10. This complication does not arise in the model of this section. Agents, now, need only to condition upon the realization of signals, which convey information about the fundamentals. There is no need for any agent to condition on prices, because prices merely convey the same information any such agent already has.

8.5.4.2 An introductory model of learning

Assume that the dividend D is made up of two random components:

$$D = \theta + Z, \tag{8.37}$$

where θ and Z are independently distributed, with $p \equiv \Pr(\theta = A) = 1 - \Pr(\theta = -A)$, and $\Pr(Z = A) = \Pr(Z = -A) = \frac{1}{2}$. Suppose that the "state" θ is unobserved.

How should we update our prior probability p of the "good" state after observing D? A simple application of Bayes' Theorem yields the posterior probabilities $\Pr(\theta = A|\, D_i)$ in Table 8.4. Considered as a random variable defined over the observable states D_i, the posterior probability $\Pr(\theta = A|\, D_i)$ has expectation $E\left[\Pr\left(\theta = A|\, D\right)\right] = p$ and variance $var\left[\Pr\left(\theta = A|\, D\right)\right] = \frac{1}{2}p\left(1-p\right)$. It is an inverse U-shaped function of p, and takes a value of zero exactly when the prior on the state is degenerate (zero or one).

	D_i (observable state)		
	$D_1 = 2A$	$D_2 = 0$	$D_3 = -2A$
$\Pr(D_i)$	$\frac{1}{2}p$	$\frac{1}{2}$	$\frac{1}{2}(1-p)$
$\Pr\left(\theta = A|\, D = D_i\right)$	1	p	0

TABLE 8.4. Distribution of posterior probabilities $\Pr\left(\theta = A|\, D\right)$.

The probabilities in Table 8.4 follow by a simple application of Bayes' Theorem. That is, for any partition $(E_i)_i$ of a given state space,

$$\Pr(E_i|F) = \Pr(E_i) \cdot \frac{\Pr(F|E_i)}{\Pr(F)} = \Pr(E_i) \cdot \frac{\Pr(F|E_i)}{\sum_j \Pr(F|E_j)\Pr(E_j)}. \tag{8.38}$$

Applying Eq. (8.38) to our example, we have:

$$\Pr(\theta = A|D = D_1) = \Pr(\theta = A)\frac{\Pr(D = D_1|\theta = A)}{\Pr(D = D_1)} = p\frac{\Pr(D = D_1|\theta = A)}{\Pr(D = D_1)}.$$

But $\Pr(D = D_1|\theta = A) = \Pr(Z = D_1 - A) = \Pr(Z = A) = \frac{1}{2}$. Moreover, we have that $\Pr(D = D_1) = \frac{1}{2}p$. This leaves $\Pr(\theta = A|D = D_1) = 1$. It's trivial, but one proceeds similarly while determining the other probabilities.

This simple example illustrates the main ideas underlying *Bayesian learning*. However, it leads to a nonlinear filter, $g(D) \equiv E(\theta = A|D)$, which differs from those we usually encounter in the literature (see, e.g., Chapters 8 and 9 in Liptser and Shiryaev, 2001*a*) (LS, in the sequel), where the instantaneous variance of the posterior probability changes, $d\pi$ say, is proportional to $\pi^2(1-\pi)^2$, not to $\pi(1-\pi)$.

This distinction arises due to technical reasons, notably because Z is a discrete random variable. Indeed, assume that Z has some arbitrary, but continuous density ϕ, and zero mean and unit variance. Let $\pi(D) \equiv \Pr(\theta = A|D \in dD)$. By the Bayes rule in Eq. (8.38),

$$\pi(D) = \Pr(\theta = A) \cdot \frac{\Pr(D \in dD|\theta = A)}{\Pr(D \in dD|\theta = A)\Pr(\theta = A) + \Pr(D \in dD|\theta = -A)\Pr(\theta = -A)}.$$

But $\Pr(D \in dD|\theta = A) = \Pr(Z = D - A) = \phi(D - A)$ and, similarly, $\Pr(D \in dD|\theta = -A) = \Pr(Z = D + A) = \phi(D + A)$. Therefore, simple calculations leave

$$\pi(D) - p = p(1-p)\frac{\phi(D-A) - \phi(D+A)}{p\phi(D-A) + (1-p)\phi(D+A)}. \tag{8.39}$$

That is, the variance of the "probability changes," $\pi(D) - p$, is proportional to $p^2(1-p)^2$.

We now extend these facts to a continuous time setting, by assuming Z is a Brownian motion. Precisely, we assume that the dynamic counterpart to Eq. (8.37) is,

$$D_\tau = \theta\tau + \sigma_0 Z_\tau, \tag{8.40}$$

for some constant σ_0, and Z_τ is a standard Brownian motion. We assume that the agent projects θ on to the information set generated by D_t, thereby being left with the following decomposition,

$$D_\tau = E\left(\theta|(D_t)_{t\leq\tau}\right)\tau + \sigma_0 W_\tau, \tag{8.41}$$

where W_τ is a Brownian motion with respect to the agents' information set $\sigma(D_t,\, t \leq \tau)$. In Appendix 9, we show that $\pi_t \equiv \pi(D_\tau)$ satisfies

$$d\pi_\tau = 2A\frac{\pi_\tau(1-\pi_\tau)}{\sigma_0}dW_\tau, \quad \pi(D_0) \equiv p, \tag{8.42}$$

where $g_\tau \equiv E\left(\theta|(D_t)_{t\leq\tau}\right)$ (the "filter") and $dW_\tau \equiv \sigma_0^{-1}(dD_\tau - g_\tau d\tau)$.[16]

[16]This construction is heuristic, but it can be made rigorous (see LS, Theorem. 8.1 p. 318 and Example 1 p. 371). In particular, it can be shown (LS, Theorem 7.12 p. 273) that W_τ is a Brownian motion with respect to the agents' information set $\sigma(D_t,\, t \leq \tau)$.

It is possible to show that if $\theta = A$ (resp. $\theta = -A$), then, $\lim_{\tau \longrightarrow \infty} \pi_\tau = 1$ (resp. 0) a.s. It is the Strong Law of Large Numbers for Brownian motions (e.g., Karatzas and Shreve, 1991): $\lim_{\tau \longrightarrow \infty} D_\tau = \infty$ or $-\infty$ according to whether $\theta = A$ or $\theta = -A$. Intuitively, if $\theta = A$, the Brownian noise in Eq. (8.40) will be "dominated" in the long-run, such that D_τ becomes arbitrarily large, which leaves any agent confident that $\theta = A$. In other words, in this model, the agents are able to figure out the truth in the long-run. Below, we shall specify an alternative model in which the agents can never completely learn.

Note an interesting implication of this model. The "filter" g_τ is linear in π_τ, i.e., $g_\tau = A(2\pi_\tau - 1)$. Therefore, the equilibrium in this economy with incomplete information is isomorphic in its pricing implications to that in a full information economy in which

$$\begin{cases} dD_\tau = g_\tau d\tau + \sigma_0 dW_\tau \\ dg_\tau = v(g_\tau)\, dW_\tau \end{cases}$$

where $v(g) \equiv (A - g)(g + A)/\sigma_0$.

8.5.4.3 Pricing implications

We now consider the asset pricing implications of an economy in which a representative agent solves the intertemporal consumption optimization problem

$$\max_{(c_\tau)_{\tau \geq 0}} E\left(\int_0^\infty e^{-r\tau} u(c_\tau)\, d\tau \right),$$

where r is the safe rate, assumed to be constant, and the utility is exponential, $u(c) = -e^{-\alpha c}$ and a is the CARA coefficient. The risk-premium coefficient in Eq. (8.25) is then $\hat{\lambda} \equiv \alpha\sigma_0$, a constant. The dynamics of dividends under the risk-neutral probability are

$$\begin{cases} dD_\tau = (g_\tau - \hat{\lambda}\sigma_0)d\tau + \sigma_0 d\hat{W}_\tau \\ dg_\tau = -\hat{\lambda} v(g_\tau)\, d\tau + v(g_\tau)\, d\hat{W}_\tau \end{cases} \tag{8.43}$$

where $\hat{W}$ is a Brownian motion under the risk-neutral probability.

Note that if the agent had CRRA equal to γ, then, we would conclude that in Samuelson-Merton markets, the dividend process is solution to

$$\frac{dD_\tau}{D_\tau} = (g_\tau - \sigma_0\hat{\lambda})d\tau + \sigma_0 d\hat{W}_\tau, \tag{8.44}$$

where $\hat{\lambda}$ would now equal $\gamma\sigma_0$, but g_τ would be solution to the the same equation in (8.43).

In these economies, the instantaneous volatility of the expected dividend growth, g, is inverse U-shaped. Due to risk-aversion, $\hat{\lambda} > 0$, the risk-neutral drift of g (i.e., $\hat{\lambda} v(g_\tau)$) is, then, a convex function of g. Results in Section 8.4 suggest that this property imply that the asset price is convex with respect to expected growth, g. Note that this convexity arises only once $\hat{\lambda} > 0$: only in this case would the risk-neutral drift of g be convex in g.

The economic interpretation of the convex drift of g is simple. An utility maximizing representative agent requires compensation for the dividend risk, but also for the risk regarding his estimated probability of living in the good state. In very good and in very bad times, this probability is either one or zero, such that there is not risk to be compensated for regarding this probability (i.e., $v(\cdot)$ is close to zero), whence the convexity of risk-adjustment. In the

next subsection, we develop more intuition on these risk-adjustments, relying on the portfolio choices of a representative agent.

The implications of this convexity is that prices undergo large swings in good times. After a string of repeated positive news on growth, the agents find it very likely that the true dividend drift is $\theta = A$: the higher is D_τ, the more likely it is that the asset is "good." In these states of nature, the agents do not feel exposed to errors on π_τ, and therefore require little risk-aversion corrections regarding g_τ (only regarding realized dividend growth, dD_τ) and so asset prices go up both because dividends go up *and* because their increases signals that the asset is good (which reduces their exposure to the risk of not observing the asset quality).[17] Note that, in bad times, the agents are not exposed to these errors either, such that asset prices fall, albeit "moderately" so. That is, asset prices do go down because dividends are going down; however, their fall is somehow mitigated because the fall in dividends signals the asset is bad and thereby reduces the agents exposure to the risk of not observing the asset quality. This convexity property may lead to fluctuations that an econometrician (say) could interpret as being generated by a bubble even if no bubbles are present whatsoever.

Finally, it is instructive to provide the expression for expected returns and volatility predicted by this model, based on the analysis of Section 8.5.2. For example, assuming that the representative agent has CRRA equal to γ, we have that $\hat{\lambda} = \sigma_0 \gamma$, such that expected returns are

$$\mathcal{E}_t = \underbrace{r_t + \sigma_0^2 \gamma}_{\mathcal{R}_t} + \underbrace{\frac{p'(g_t)}{p(g_t)} \sigma_0 \gamma v(g_t)}_{\mathcal{W}_t} \tag{8.45}$$

and volatility is

$$\text{Vol}_t = \sigma_0 + \frac{p'(g_t)}{p(g_t)} v(g_t). \tag{8.46}$$

Below, we provide a more rigorous analysis of these convexity properties relying on two famous (and more general) models of learning. Prior to this, we provide additional intuition regarding the previous mechanisms based on agents' portfolio choices.

8.5.4.4 With Bellman

The previous explanations regarding risk-aversion corrections can be illustrated while solving the agent's dynamic programming problem. In particular, we shall show, it is the hedging component of the agent's portfolio choice to determine how much the agent is willing to invest in the asset in periods of uncertainty.

[In progress]

8.5.4.5 Convexity again, and two models of learning

The model in Eqs. (8.43) can be considered as a special case of that considered by Veronesi (1999), in which an infinitely lived agent has constant *absolute* risk aversion equal to $\alpha > 0$, and observes realizations of D, generated by:

$$dD_\tau = \theta d\tau + \sigma_0 dZ_{1\tau}, \tag{8.47}$$

where $Z_{1\tau}$ is a Brownian motion, and θ is the expected dividend change, supposed to follow a two-state $(\underline{\theta}, \bar{\theta})$ Markov chain. (See, also, David, 1997, for a related model.) In Eqs. (8.43), the

[17] In terms of Eqs. (8.44), the drift of dividend growth is tilted to the left due the risk-adjustment term, $\sigma_0 \hat{\lambda}$. Moreover, we also have that g_τ is tilted to the left, due to the negative drift, $-\hat{\lambda} v(g_\tau)$. The effects of $\hat{\lambda} v(g_\tau)$ are small in very good and very poor times, bacause $v(\cdot)$ is small in these cases. The effects of $\sigma_0 \hat{\lambda}$ are independent of the state of the economy.

dividend is a "type," in that once nature draws θ, this θ is there forever. In Eqs. (8.47), instead, θ is allowed to change, according to a Markov chain, as explained.

The key aspect in this economy is that the expected dividend change, θ, is unobserved. As a result of this lack of knowledge, the agent attempts to learn about the state in which he is living, through Bayesian learning. The resulting economy is one generalizing that in the previous section. In this economy, the price is the same as that in a full information economy in which the dividend is solution to

$$\begin{cases} dD_\tau = g_\tau d\tau + \sigma_0 dW_\tau \\ dg_\tau = k\left(\bar{g} - g_\tau\right) d\tau + v\left(g_\tau\right) dW_\tau \end{cases} \tag{8.48}$$

where $v\left(g\right) = (\bar{\theta} - g)(g - \underline{\theta})/\sigma_0$, and $k, \bar{g}$ are some positive constants. Note that while the diffusion terms in Eq. (8.43) and in Eq. (8.48) have the same functional form, the expected dividend change is a martingale in Eq. (8.43), and mean-reverting in Eq. (8.48), under the *physical* probability—i.e. when $\hat{\lambda} = 0$ in Eq. (8.43).

These differences exists because in Eq. (8.43), θ is drawn at time 0 and stays the same forever, whereas in Eq. (8.48), θ is a Markov chain. In other words, in Eq. (8.47), the agent will never figure out the truth, and g_τ is mean-reverting toward its average rate.

Eqs. (8.48) amount to a special case of Eqs. (8.22). Therefore, by Eq. (8.25), the risk-premium is constant, and equal to $\lambda = \alpha\sigma_0$ such that the dynamics of dividends under the risk-neutral probability are:

$$\begin{cases} dD_\tau = \left(g_\tau - \alpha\sigma_0^2\right) d\tau + \sigma_0 d\hat{W}_\tau \\ dg_\tau = \left(k(\bar{g} - g_\tau) - \alpha\sigma_0 v\left(g_\tau\right)\right) d\tau + v\left(g_\tau\right) d\hat{W}_\tau \end{cases} \tag{8.49}$$

Veronesi (1999) also assumes the riskless asset is infinitely elastically supplied, and therefore that the interest rate r is a constant.

Let us analyze the properties of the equilibrium price predicted by this model. It is easy to see that given Eq. (8.49), the asset price is:

$$S(D, g) = \mathbb{E}\left(\int_0^\infty e^{-r\tau} D_\tau d\tau \middle| D, g\right) = \int_0^\infty C\left(D, g, \tau\right) d\tau,$$

where

$$C\left(D, g, \tau\right) = e^{-r\tau}\left(D - \alpha\sigma_0^2\tau\right) + G\left(g, \tau\right), \quad \text{and} \quad G\left(g, \tau\right) \equiv e^{-r\tau}\int_0^\tau \mathbb{E}\left(g_u \middle| g\right) du. \tag{8.50}$$

The conditional expectation, $\mathbb{E}\left(g_u \middle| g\right)$ can be read as a special case of the canonical price in Eq. (8.14), namely, for $\rho = 0$ and $\psi(g) = g$. By Theorem 8.1-(ii), $\mathbb{E}\left(g_u \middle| g\right)$ is convex in g whenever the drift of g in Eq. (8.49) is convex. This condition always holds true because $\alpha > 0$. That is, the conditional expectation of g in Eq. (8.50), inherits the same second order properties (convexity) of this drift function.

The economics behind these convexities is the same as that behind Eqs. (8.43) or (8.44). Prices are convex in the expected dividend growth because of a kind of "speculative enthusiasm": after a sequence of high realized growth, investors believe that the likelihood is high they are living in good times, *and* that the likelihood is low they are making mistakes in their assessment. Thus, asset prices increase fueled by low risk-premium discounts. Note that these low premiums regard the uncertainty the agents face while assessing the world in which they live. Their "speculative enthusiasm" is rational, so to speak, i.e., not determined by "animal spirits."

These properties hold as we are assuming that the riskless asset is infinitely elastically supplied. When investors' demand for the safe assets affects the interest rate, the interest rate becomes a function of the expected growth, g. Thus when g increases, the increase in the asset price is mitigated by the fact that relatively less savings are now being made, reflecting the fact that more resources are likely to be available in the future. The interest rate increase. Formally, by Eq. (8.24), the short-term rate is $r(g) = \delta + \alpha g - \frac{1}{2}\alpha^2\sigma_0^2$, such that the asset price is, now,

$$S_r(D,g) \equiv \mathbb{E}\left(\int_0^\infty e^{-\int_0^\tau r(g_s)ds} D_\tau d\tau \middle| D, g\right) = \int_0^\infty C_r(D,g,\tau)\, d\tau,$$

where

$$C_r(D,g,\tau) = (D - \alpha\sigma_0^2\tau)\, B(g,\tau) + G_r(g,\tau)$$
$$B(g,\tau) \equiv \mathbb{E}\left(e^{-\int_0^\tau r(g_s)ds}\middle| g\right), \quad G_r(g,\tau) \equiv \mathbb{E}\left(\int_0^\tau e^{-\int_0^\tau r(g_s)ds} g_u \middle| g\right)$$

By Theorem 8.1, the two functions, $B(g,\tau)$ and $G_r(g,\tau)$, are *affine* in g if the risk-neutral drift of g in Eq. (8.49), $b(g) \equiv (k(\bar{g} - g) - \alpha\sigma_0 v(g))$, and the short-term rate $r(g)$, are such that $b''(g) = 2r'(g)$. A simple calculation reveals this is indeed the case, with $b''(g) = 2\alpha$. That is, if interest rates are determined endogenously, the asset price is no longer convex in the dividend expected growth. It is unpleasant arithmetics. We know the short-term rate should not fluctuate too much, compared to stocks. Assuming it is constant as Veronesi (1999) did, is reasonable and leads to interesting properties, which would be destroyed under the counterfactual circumstance the short-term rate is allowed to considerably fluctuate.

Interestingly, this property that prices are linear in g, once interest rates are endogeneous, is not specific to cases where dividend is as in Eqs. (8.48) and agents have CARA. Veronesi (2000) shows that convexity properties are lost even in proportional markets (i.e. markets where the distribution of dividend growth is independent of the initial level of dividends) and representative agents with CRRA. In fact, Appendix 10 shows that there are no linear signal structures and representative agent CRRA economies with *complete securities markets* supporting the convexity property.

In the previous model, the asset price reacts asymmetrically to changes in the expected dividend growth, due to the assumption that agents learn about a *discrete* state space in a *continuous-time* economy. We now describe another model, analyzed by Brennan and Xia (2001), in which the agents learn about a *continuous* state in a *continuous-time* economy. In this model, an infinitely lived agent has CRRA preferences, and observes D, solution to:

$$\frac{dD_\tau}{D_\tau} = \hat{g}_\tau d\tau + \sigma_0 dZ_{1\tau},$$

and the expected dividend growth, $\hat{g}_\tau$, is unobserved. Rather than assuming $\hat{g}_\tau$ to be on a countable number of states, Brennan and Xia postulate that it is an Ornstein-Uhlenbeck process:

$$d\hat{g}_\tau = k(\bar{g} - \hat{g}_\tau)d\tau + \sigma_1 dZ_{1\tau} + \sigma_2 dZ_{2\tau},$$

where $\bar{g}$, σ_1 and σ_2 are positive constants.

The agent learns about $\hat{g}_\tau$ through Bayes' rule, similarly as in the previous section. It can be shown that if the agent has a Gaussian prior on $\hat{g}_0$ with variance α_*^2 defined below, the price is $S(D,g)$, where D satisfies

$$\begin{cases} \dfrac{dD_\tau}{D_\tau} = g_\tau d\tau + \sigma_0 dW_{1\tau} \\ dy_\tau = k(\bar{g} - g_\tau)d\tau + v dW_{1\tau} \end{cases}$$

where σ_0, k, $\bar{g}$ and v are constants.[18,19] By Eq. (8.25), the risk-premium is constant, and equal to $\lambda = \gamma\sigma_0$, where γ is the CRRA coefficient, and by Eq. (8.24), the short-term rate is linear in g. Therefore, and by the same reasoning leading to Eq. (8.36), the price-dividend ratio is independent of D, and is given by:

$$p(g) = \int_0^\infty e^{-\sigma_0\lambda\tau}\bar{\mathbb{E}}\left(e^{\int_0^\tau (g_u - r(g_u))du}\middle| g\right) d\tau, \tag{8.51}$$

where

$$dg_\tau = (\varphi(g_\tau) + (\sigma_0 - \lambda)\, v(g_\tau))\, d\tau + v(g_\tau)\, d\bar{W}_\tau, \tag{8.52}$$

and $\bar{W}_\tau$ is a $\bar{P}$-Brownian motion.[20] The two functions, $\varphi(g)$ and $v(g)$, are momentarily left unspecified, as we wish to provide general results within this model.

Under regularity conditions, monotonicity and convexity properties are inherited by the inner expectation in Eq. (8.51). Precisely, in the notation of the canonical pricing problem, we have that $\rho(g) \equiv -g + r(g) + \sigma_0\lambda$ and $b(g) \equiv \varphi(g) + (\sigma_0 - \lambda)\, v(g)$. Therefore, by Theorem 8.1, we have:

(i) *If $r'(g) < 1$, the price-dividend ratio is increasing in the expected dividend growth, g.*

(ii) *Suppose that the price-dividend ratio is increasing in g. Then, it is also convex in g whenever $r''(g) < 0$ and $\frac{d^2}{dg^2}(\varphi(g) + (\sigma_0 - \lambda)\, v(g)) \geq -2 + 2r'(g)$.*

These conditions are satisfied by Brennan and Xia (2001). [...] Explain that the inner expectation in Eq. (8.51) is one for an affine model to be introduced and explained in full detail in Chapter 13. Finally, expected returns and returns volatility have the same expression as in Eqs. (8.45) and (8.46). [Determine and analyze these expected returns and return volatility] [In progress]

8.5.5 Linearity-generating processes

The focus of the previous sections is a search for pricing kernels that make asset pricing models qualitatively consistent with countercyclical statistics—a search relying on theoretical test conditions, those emanating from Theorem 8.1 in Section 8.4.

We can actually use Theorem 8.1 for another (somehow surprising) purpose: a search for asset pricing models that have a closed-form solution. The idea is simple. Theorem 8.1-(ii) provides conditions under which price-dividend ratios can be either concave or convex in the state variables driving them. Specifically, consider the expression for the price-dividend ratio in Eqs. (8.16)-(8.17)-(8.19). In terms of the canonical pricing problem of Section 8.4, we set the expressions in Eqs. (8.20), reported here for convenience,

$$\psi(y) \equiv 1, \quad \rho(y) \equiv \mathcal{R}(y) - m(y), \quad b(y) = \bar{\varphi}(y),$$

[18] That is, in terms of Eqs. (8.22), $m(D,g) = gD$, $\sigma_0(D) = \sigma_0 D$, $\varphi(g) = k(\bar{g} - g)$, $v_2 = 0$. Moreover, the constant v satisfies $v \equiv v_1 \equiv v_1(\alpha_*) = \sigma_1 + \frac{\alpha_*}{\sigma_0}$, where α_* is the positive solution to $v_1(x) = \sigma_1^2 + \sigma_2^2 - 2kx$.

[19] Brennan and Xia (2001) actually consider a slightly more general model, where consumption and dividends differ. They derive a model with a reduced-form identical to that in this example. In the calibrated model, Brennan and Xia found that the variance of the filtered $\hat{g}$ is higher than the variance of the expected dividend growth in an economy with complete information. The results on γ^* in this example can be obtained through an application of Theorem 12.1 in Liptser and Shiryaev (2001) (Vol. II, p. 22). They generalize results in Gennotte (1986) and are a special case of results in Detemple (1986). Both Gennotte and Detemple did not emphasize the impact of learning on the pricing function.

[20] By Girsanov's theorem, $\bar{W}_\tau = \hat{W}_\tau - \sigma_0\tau$ and $\hat{W}_\tau = W_\tau + \lambda\tau$ are Brownian motions under $\bar{P}$ and under Q. That is, $\bar{W}_\tau = W_\tau + (\lambda - \sigma_0)\tau$, whence Eq. (8.52).

where as usual, $\mathcal{R}(y) \equiv r(y) + \sigma_0 \lambda_{\text{CF}}(y)$ and, by Eq. (8.19),

$$\bar{\varphi}(y) \equiv \varphi(y) - \sum_{i=1}^{2} v_i(y)\,\lambda_i(y) + \sigma_0 v_1(y),$$

and $\varphi(y)$ denotes the drift of y under the physical probability. Then, given that $\psi' = \psi'' = 0$, we have that by Theorem 8.1-(i),

$$c \text{ is: } \begin{array}{c} \text{If } \bar{\varphi}'' = 2(\mathcal{R}' - m') \\ \left\{ \begin{array}{lll} \text{concave in } y & \text{if } (\mathcal{R}'' - m'') \geq 0 & \text{and } \psi'' < 0 \\ \text{convex in } y & \text{if } (\mathcal{R}'' - m'') \leq 0 & \text{and } \psi'' > 0 \end{array} \right. \end{array} \tag{8.53}$$

regardless of whether c is increasing in y.

It is quite unlikely a closed-form solution exists for the price-dividend ratio when one of these conditions is satisfied. Yet how about assuming none of the two conditions for concavity or convexity hold true? In this case, obviously, price-dividend ratios are neither concave or convex, so they must necessarily be affine in the state variables! To summarize, Eqs. (8.53) suggest that the price-dividend ratios are affine in y if

$$\bar{\varphi}'' = 2(\mathcal{R}' - m') \quad \text{and} \quad \mathcal{R}'' - m'' = 0. \tag{8.54}$$

Gabaix (2009) is the first to note that price-dividend ratios are affine in the state variables driving them, should the drift of these state variables be quadratic. His remarks are consistent with Theorem 8.1, and the two conditions in (8.54). In fact, Gabaix develops a unified theory of "linearity-generating processes," which generalizes the single state variable framework underlying Theorem 8.1 and the related conditions in in (8.54).

To illustrate these facts, consider a model that fits the class of linearity-generating processes, one of external habit formation by Menzly, Santos and Veronesi (2004) (MSV, henceforth). In this model, a representative agent maximizes,

$$U = E\left(\int_0^\infty e^{-\delta t} \ln(c_t - x_t)\, dt\right), \tag{8.55}$$

where x_t is external habit. Relative risk-aversion equals the inverse of the surplus consumption ratio, $\frac{1}{s_t}$, with $s_t = \frac{c_t - x_t}{c_t}$, which in equilibrium equals $\frac{D_t - x_t}{D_t}$, where D is consumption endowment.

MSV assume that the surplus consumption ratio is a continuous-time autoregressive process, solution to,

$$d\left(\frac{1}{s_t}\right) = \beta\left(\frac{1}{\bar{s}} - \frac{1}{s_t}\right) dt - \alpha\left(\frac{1}{s_t} - \frac{1}{\upsilon}\right) \sigma_0 dW_t, \tag{8.56}$$

for some constants β, $\bar{s}$, ω and υ. It can be shown that if α is small enough, $s_t \in (0, \upsilon)$. Ljungqvist and Uhlig (2000) utilize similar assumptions to model productivity shocks in their model of catching up with the Joneses.

It is easy to verify that this model does satisfy the two conditions in (8.54). First, note that Eq. (8.56) implies that the surplus consumption ratio satisfies,

$$ds_t = s_t\left(\beta\left(1 - \tfrac{s_t}{\bar{s}}\right) + \alpha^2\sigma_0^2\left(1 - \tfrac{s_t}{\upsilon}\right)^2\right) dt + \alpha\sigma_0 s_t\left(1 - \tfrac{s_t}{\upsilon}\right) dW_t,$$

such that,

$$\bar{\varphi}(s) = s\left(\beta\left(1-\tfrac{s}{\bar{s}}\right)+\alpha^2\sigma_0^2\left(1-\tfrac{s}{\upsilon}\right)^2\right)+\left(\sigma_0-\lambda_{\text{CF}}(s)\right)\alpha\sigma_0 s\left(1-\frac{s}{\upsilon}\right).$$

Note that the drift function of the surplus ratio under $\bar{Q}$ is quadratic, a property that would crucially lead to the first condition in (8.54) to be satisfied.

By results in Section 8.5.1, the market Sharpe ratio is $\lambda(D,x)=\frac{1}{s}(\sigma_0-\frac{v(D,x)}{D})$, where $v(D,x)$ is the instantaneous volatility of the habit level, $x=D(1-s)$ and, by Itô's lemma, equals $v(D,x)=D\sigma_0(1-s)-D\text{Vol}(s)$, such that, by Itô's lemma again, $\text{Vol}(s)=\alpha\sigma_0 s\left(1-\frac{s}{\upsilon}\right)$. Therefore, the market Sharpe ratio equals,

$$\lambda_{\text{CF}}(s)\equiv\lambda(D,x)=\sigma_0\left(1+\alpha\left(1-\frac{s}{\upsilon}\right)\right),$$

which is countercyclical. Finally, again by results in Section 8.5.1, we can also infer that:

$$r(s)=\delta+g_0-\sigma_0^2+\beta\left(1-\frac{s}{\bar{s}}\right)-\sigma_0^2\alpha\left(1-\frac{s}{\upsilon}\right).$$

Therefore, we have that,

$$\bar{\varphi}''(s)=-2\frac{\beta}{\bar{s}},\quad \mathcal{R}'(s)=-\frac{\beta}{\bar{s}},\quad m'(s)=m''(s)=0.$$

As a result, the two conditions in (8.54) are satisfied, and the price-dividend ratio is affine in s. We can check that the price-dividend ratio is affine in s, through a direct computation. Denote the instantaneous utility with $u(c,x)\equiv\ln(c-x)$, with $c=D$. By the usual asset price representation, we have that the price-dividend ratio, $p(s_0)$, is,

$$\begin{aligned}
p(s_0) &= E\left(\int_0^\infty e^{-\delta t}\frac{u_c(c_t,x_t)}{u_c(c_0,x_0)}\frac{c_t}{c_0}dt\right)\\
&= s_0\cdot\int_0^\infty e^{-\delta t}E\left(\frac{1}{s_t}\right)dt\\
&= s_0\cdot\int_0^\infty e^{-\delta t}\left(\frac{1}{\bar{s}}+e^{-\beta t}\left(\frac{1}{s_0}-\frac{1}{\bar{s}}\right)\right)dt\\
&= \frac{1}{\delta+\beta}+\frac{\beta}{\delta\bar{s}(\delta+\beta)}s_0.
\end{aligned}$$

Figure 8.13 depicts the price-dividend ratio as a function of the current surplus consumption ratio, using the following parameter values, $\rho=0.04$, $\beta=0.15$, and $\bar{s}=0.03$.

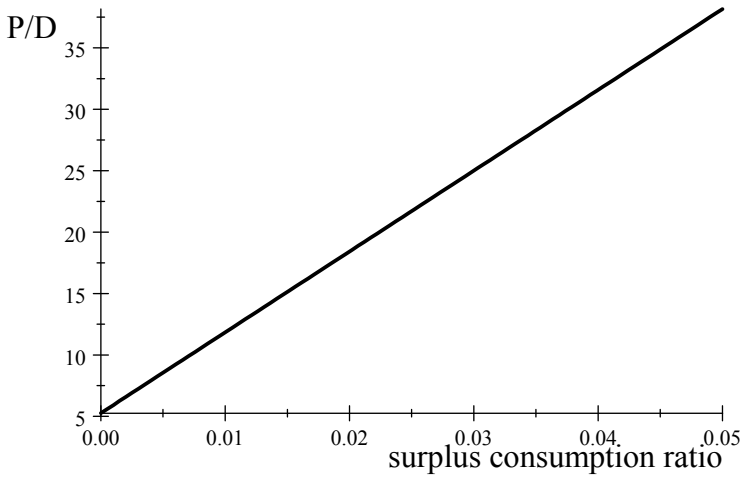


FIGURE 8.13. Price-dividend ratio for the aggregate consumption claim predicted by the Menzly, Santos and Veronesi (2004) model of external habit formation.

As an example of a simple case of a linearity-generating process, consider a model with a constant market Sharpe ratio λ and a constant short-term rate r, but with stochastic dividend growth,

$$\frac{dD_t}{D_t} = g_\tau d\tau + \sigma_0 dW_\tau,$$

where the expected dividend growth, y_τ, is solution to,

$$dg_\tau = -g_\tau^2 d\tau + vW_\tau,$$

such that,

$$m(g) = g, \quad \mathcal{R}(g) = r + \sigma_0\lambda, \quad \bar{\varphi}(g) = -g_\tau^2 + (\sigma_0 - v)\lambda.$$

It is straightforward to see that the two conditions in (8.54) hold true—the price-dividend ratio is affine in g. Indeed, assuming that the price-dividend ratio is independent of D, we have that it satisfies the following differential equation,

$$\left(-g^2 + \sigma_0 v\right) p' + \frac{1}{2} v^2 p'' + 1 = (r - g)\, p. \tag{8.57}$$

Let us conjecture, now, that the price-dividend ratio is affine in g, i.e. there are two constant α and β such that

$$p(g) = \alpha_0 + \alpha_1 g.$$

Replacing the previous expression into Eq. (8.57) allows us to pin down the two expressions for α_0 and α_1 such that the solution for the price-dividend ratio is,

$$p(g) = \frac{r + g}{r^2 - \sigma_0 v}.$$

8.6 Modeling market-to-book ratios

[In progress]

8.7 Appendix 1: Estimation of the market expected return

We estimate yearly expected excess returns by relying on the following model

$$\tilde{R}_t - r_t = a + \phi(\tilde{R}_{t-1} - r_{t-1}) + b^\top p_{t-1} + \epsilon_t, \tag{8A.1}$$

where a, ϕ, and b are coefficients (b is two-dimensional), p_t is a vector containing default and term spreads, and ϵ_t is a residual term, assumed to be independent and identically normally distributed. Note that Eq. (8A.1) could be cast, equivalently, as:

$$\begin{cases} \tilde{R}_t - r_t = \hat{R}_t + (1 - \phi L)^{-1} \epsilon_t \\ \hat{R}_t = a + \phi \hat{R}_{t-1} + b^\top p_{t-1} \end{cases} \tag{8A.2}$$

where L is the lag operator, $Lx_t \equiv x_{t-1}$, such that yearly expected returns could be defined just as in Eq. (8.2) in the main text, with $P_t(\hat{R}_{t+j})$ denoting the projection of $\hat{R}_{t+1}$ based on model (8A.2).

To calculate the projection terms in Eq. (8.2) obviously necessitates specifying the dynamics of p_t in (8A.2). We proceed, alternatively, as follows. First, we estimate Eq. (8A.1). Second, we reconstruct a time series of monthly expected returns, $\hat{R}_t$ in the second of Eqs. (8A.2) given the estimates of a, ϕ and b. Third, we fit an AR(1) model to the reconstructed series $\hat{R}_t$, and calculate projections on this series to calculate $\hat{\mathcal{E}}_t$ in Eq. (8.2).

8.8 Appendix 2: Calibration of the tree in Section 8.3

SOLUTION AND CALIBRATION OF THE MODEL. We provide details regarding the calibration results in Table 8.2 of the main text. In the first step, we estimate the two parameters p and δ of the dividend process. Let G be the yearly dividend gross growth rate. We calibrate p and δ by a perfect matching of the model expected dividend growth, $\mu_D \equiv E(G) = pe^{-\delta} + (1-p)e^{\delta}$ and dividend variance, $\sigma_0^2 \equiv var(G) = \left(e^{\delta} - e^{-\delta}\right)^2 p(1-p)$, to their sample counterparts $\hat{\mu}_D = 1.0594$ and $\hat{\sigma}_0 = 0.0602$ obtained with US aggregate dividend data. The result is $(p, \delta) = (0.158, 0.082)$. Given these calibrated values of (p, δ), we fix $r = 1.0\%$, and proceed to calibrate the probabilities q, q_B and q_G.

To calibrate (q, q_B, q_G), we need an explicit expression for all the payoffs at each node. We rely on the price of the claim, M_S, which we obtain in closed-form, as follows. For each state $S \in \{G, B, GB\}$, M_S is solution to,

$$\frac{M_S}{D_S} = e^{-r}\mathbb{E}_S\left(\frac{D'_S}{D_S} + \frac{M'_S}{D'_S}\frac{D'_S}{D_S}\right), \tag{8A.3}$$

where $\mathbb{E}_S(\cdot)$ is the expectation taken under the risk-neutral probability q_S in state S, $S \in \{G, B, GB\}$, and $q_{GB} = q$, $D_G = e^{2\delta}$, $D_B = e^{-2\delta}$, $D_{GB} = 1$, and D'_S and M'_S are the dividend and the price of the claim as of the next period. Since risk-aversion is constant from the third period on, the price-dividend ratio is constant as well, from the third period on, which implies that $\frac{M_S}{D_S} = \frac{M'_S}{D'_S}$. Replacing this equality into Eq. (8A.3), and solving for M_S, yields,

$$M_S = D_S \frac{q_S e^{-\delta} + (1-q_S)e^{\delta}}{e^r - \left[q_S e^{-\delta} + (1-q_S)e^{\delta}\right]}, \quad S \in \{G, B, GB\}. \tag{8A.4}$$

We calibrate $(q_G, q_B, q_{GB} = q)$ to make the "average" price-dividend (P/D henceforth) ratio M_{GB}, the "good" P/D ratio $\frac{M_G}{e^{2\delta}}$ and the "bad" P/D ratio $\frac{M_B}{e^{-2\delta}}$ in Eq. (8A.4) perfectly match the average P/D ratio, the average P/D ratio during NBER expansion periods, and the average P/D ratio during NBER recession periods (i.e. 31.99, 33.21 and 26.20, from Table 8.1). Given $(p, \delta, r, q, q_B, q_G)$, we compute the P/D ratios in states G and B. For example, the price of the asset in state B is, $P_B = e^{-r}[q_B\left(e^{-2\delta} + M_B\right) + (1-q_B)(1+M_{GB})]$. Given P_B, we compute the log-return in the bad state as $\ln(\frac{\tilde{\Pi}}{P_B})$, where

$$\tilde{\Pi} = \begin{cases} e^{-2\delta} + M_B & \text{with probability } p \\ 1 + M_{GB} & \text{with probability } 1-p \end{cases}$$

Then, we compute the return volatility in state B. The P/D ratios, the expected log-return and return volatility in state G are computed similarly. (Please notice that volatilities under p and under $\{q_S\}_{S\in\{G,B,GB\}}$ are not the same.)

Next, we recover the risk-aversion parameter γ_S in the three states $S \in \{G, B, GB\}$ implied by the previously calibrated probabilities q, q_G and $q = q_{GB}$. As we shall show below, the relevant formula to use is,

$$\frac{q_S}{p} = \frac{e^{\gamma_S \delta}}{pe^{\gamma_S \delta} + (1-p)e^{-\gamma_S \delta}}, \quad S \in \{G, B, GB\}. \tag{8A.5}$$

The values for the "implied" risk-aversion parameter in Table 8.2 are obtained by inverting Eq. (8A.5) for γ_S, given the calibrated values of (p, δ, q_S, q_G).

Finally, we determine the risk-adjusted discount rate as $r + \hat{\sigma}_0 \lambda_S$, where λ_S is the Sharpe ratio, which we shall show below to equal,

$$\lambda_S = \frac{q_S - p}{\sqrt{p(1-p)}}, \quad S \in \{G, B, GB\}. \tag{8A.6}$$

Proof of Eq. (8A.5). We only provide the derivation of the risk-neutral probability q_B, since the proofs for the expressions of the risk-neutral probabilities q_G and $q = q_{GB}$ are nearly identical. In equilibrium, the Euler equation for the stock price at the "bad" node is,

$$P_B = \beta E\left[\frac{u_B'(\tilde{D}_S)}{u_B'\left(e^{-\delta}\right)}\left(\tilde{D}_S + M_S\right)\right] = \beta E\left[\tilde{G}_S^{-\gamma_B}\left(\tilde{D}_S + M_S\right)\right], \quad S \in \{B, GB\}, \tag{8A.7}$$

where: (i) β is the discount rate; (ii) the utility function for consumption C is state dependent and equal to, $u_B(C) = C^{1-\gamma_B}/(1-\gamma_B)$; (iii) $E(\cdot)$ is the expectation taken under the physical probability p; and (iv) the dividend $\tilde{D}_S$ and the gross dividend growth rate $\tilde{G}_S$ are either $\tilde{D}_B = e^{-2\delta}$ and $\tilde{G}_B = \frac{e^{-2\delta}}{e^{-\delta}} = e^{-\delta}$ with probability p, or $\tilde{D}_{GB} = 1$ and $\tilde{G}_{GB} = \frac{1}{e^{-\delta}} = e^{\delta}$ with probability $1-p$.

In the model, the asset is elastically supplied or, equivalently, there exists a storage technology with a fixed rate of return equal to $r = 1\%$. Let us derive the agent's private evaluation of this asset. The Euler equation for the safe asset is,

$$e^{-r_B} = \beta E(\tilde{G}_S^{-\gamma_B}) = \beta \sum_{S\in\{B,GB\}} p_S \tilde{G}_S^{-\gamma_B}, \tag{8A.8}$$

where the safe interest rate, r_B, is state dependent, $p_B = p$ and $p_{GB} = 1-p$. Therefore,

$$q_B = \beta e^{r_B} p \tilde{G}_B^{-\gamma_B}, \quad 1 - q_B = \beta e^{r_B}(1-p)\tilde{G}_{GB}^{-\gamma_B} \tag{8A.9}$$

is a probability distribution. In fact, by plugging q_B and $1-q_B$ into Eq. (8A.7), one sees that it is the risk-neutral probability distribution. To obtain Eq. (8A.5), note that by Eq. (8A.8), $\beta e^{r_B} = 1/E(\tilde{G}_S^{-\gamma_B})$, which replaced into Eq. (8A.9) yields,

$$\frac{q_B}{p} = \frac{\tilde{G}_B^{-\gamma_B}}{E(\tilde{G}_S^{-\gamma_B})}.$$

Eq. (8A.5) follows by the definition of $\tilde{G}_S$ given above.

Proof of Eq. (8A.6). Let e^{μ} the gross expected return of the risky asset. The asset return can take two values: e^{R_l} with probability p, and e^{R_h} with probability $1-p$, and $R_h > R_l$. Therefore, for each state, we have that:

$$e^{\mu} = pe^{R_l} + (1-p)e^{R_h}, \quad e^{r} = qe^{R_l} + (1-q)e^{R_h}, \tag{8A.10}$$

where we have omitted the dependence on the state S to alleviate the presentation. The standard deviation of the asset return is $\text{Std}_R = \left(e^{R_h} - e^{R_l}\right)\sqrt{p(1-p)}$. The Sharpe ratio is defined as

$$\lambda = \frac{e^{\mu} - e^{r}}{\text{Std}_R}.$$

By substracting the two equations in (8A.10),

$$q = p + \frac{e^{\mu} - e^{r}}{\left(e^{R_h} - e^{R_l}\right)\sqrt{p(1-p)}}\sqrt{p(1-p)} = p + \lambda\sqrt{p(1-p)},$$

from which Eq. (8A.6) follows immediately. Note, also, that in terms of this definition of the Sharpe ratio, the risk-neutral expectation of the dividend growth is, $\mathbb{E}(G) = E(G) - \lambda\sigma_0$.

8.9 Appendix 3: Asset prices in a multifactor model

Consider a reduced-form model, where an asset price, S_i say, is a twice-differentiable function of a number of factors, $S_i = S_i(y)$, $i = 1, \cdots, m$, and $y = [y_1, \cdots, y_d]^\top$ is a vector of factors. We assume that the i-th asset pays off an instantaneous dividend $D_i = D_i(y)$, and that y is a diffusion process:

$$dy_t = \varphi(y_t)\, dt + v(y_t)\, dW_t,$$

where φ is d-valued, v is $d \times d$ valued, and W is a d-dimensional Brownian motion. We assume the number of assets does not exceed the number of factors, $m \leq d$, consistently with the framework in Chapter 5. By Itô's lemma:

$$\frac{dS_i}{S_i} = \frac{LS_i}{S_i} dt + \frac{\overbrace{\nabla S_i}^{1\times d}}{S_i} \overbrace{v}^{d\times d} dW,$$

where LS_i is the infinitesimal operator. Let r_t be the instantaneous short-term rate. By no-arbitrage, and under regularity conditions, there exists a measurable d-vector process λ, the vector of unit prices of risk associated with the fluctuations of the factors, such that,

$$\begin{bmatrix} \frac{LS_1}{S_1} - r + \frac{D_1}{S_1} \\ \vdots \\ \frac{LS_m}{S_m} - r + \frac{D_m}{S_m} \end{bmatrix} = \underbrace{\sigma}_{m\times d} \underbrace{\lambda}_{d\times 1}, \qquad \text{where } \sigma = \begin{bmatrix} \frac{\nabla S_1}{S_1} \\ \vdots \\ \frac{\nabla S_m}{S_m} \end{bmatrix} \cdot v. \tag{8A.11}$$

To simplify, we take this economy to be Markov, assuming that, $r_t \equiv r(y_t)$ and $\lambda_t \equiv \lambda(y_t)$.

Eqs. (8A.11) add up to a system of m uncoupled partial differential equations, and the solution is one no-arbitrage price system, assuming no bubbles.

8.10 Appendix 4: Arrow-Debreu PDEs

We develop an interesting connection. Note that by Eq. (8A.11), S is solution to,

$$LS + D = rS + (S_D \sigma_0 D + S_y v_1)\,\lambda_1 + S_y v_2 \lambda_2, \quad \forall (D, y) \in \mathbb{R} \times \mathbb{R}. \tag{8A.12}$$

Under regularity conditions, the Feynman-Kac representation of the solution to Eq. (8A.12) is,

$$S\,(D, y) = \int_0^\infty C\,(D, y, \tau)\,,$$

where,

$$C\,(D, y, \tau) = \mathbb{E}\left(e^{-\int_0^\tau r(y_t)dt} \cdot D_\tau \middle| \, D, y\right) = E\,(m_\tau \cdot D_\tau |\, D, y)\,,$$

and m is the stochastic discount factor: $m_\tau = \frac{\xi_\tau}{\xi_0}$, $\xi_0 = 1$.

Alternatively, we represent the price S under the physical probability. Given the previous assumptions, we have that ξ necessarily satisfies,

$$\frac{d\xi_\tau}{\xi_\tau} = -r\,(y_\tau)\,d\tau - \lambda_1\,(y_\tau)\,dW_{1\tau} + \lambda_2\,(y_\tau)\,dW_{2\tau}. \tag{8A.13}$$

Next, define the undiscounted "Arrow-Debreu adjusted" asset price process, defined as:

$$w(D, y) \equiv \Upsilon(D, y) \cdot S(D, y),$$

where $\Upsilon(D, y)$ is as in Eq. (8.21),

$$\Upsilon\,(D_\tau, y_\tau) = \xi_\tau e^{\int_0^\tau \delta(D_s, y_s)ds}.$$

By results in Section 8.4.2, we know that the following price representation holds true:

$$S_\tau \xi_\tau = E\left(\int_\tau^\infty \xi_s D_s ds\right), \quad \tau \geq 0.$$

Under regularity conditions, the previous equation can then be understood as the unique Feynman-Kac stochastic representation of the solution to the following partial differential equation

$$Lw(D, y) + f(D, y) = \delta(D, y) w(D, y), \quad \forall (D, y) \in \mathbb{R} \times \mathbb{R},$$

where $f \equiv \Upsilon D$. Eq. (8A.12) then follows by the definition of $Lw_\tau \equiv \frac{d}{ds} E\,[\Upsilon S]\big|_{s=\tau}$.

8.11 Appendix 5: Increasing risk and convexity

8.11.1 The maximum principle

Consider the differential equation:

$$\frac{dx_\tau}{d\tau} = \phi_\tau,\ \tau \in (t,T), \quad x_T = 0, \quad \text{sign}\left(\phi_\tau\right) = \text{constant on } \tau \in (t,T), \tag{8A.14}$$

where ϕ is such that x is bounded on (t,T). We claim that

$$\text{sign}\left(x_t\right) = -\ \text{sign}\left(\phi_t\right). \tag{8A.15}$$

Figure 8A.1 provides intuition underlying Eq. (8A.15). Consider the following heuristic arguments. Note that

$$0 = x_T = x_t + \int_t^T \phi_\tau d\tau \Longrightarrow x_t = -\int_t^T \phi_\tau d\tau,$$

such that Eq. (8A.15) holds.

Next, suppose that x_τ still satisfies Eq. (8A.14), but that at the same time, x_τ is some function of a state variable y_τ and time,

$$x_\tau = f\left(y_\tau, \tau\right),$$

where y_τ satisfies:

$$\frac{dy_\tau}{d\tau} = b_\tau, \quad \tau \in (t,T).$$

Assuming enough regularity conditions, we have that

$$\frac{dx}{d\tau} = \left(\frac{\partial}{\partial\tau} + L\right) f, \quad \tau \in (t,T), \tag{8A.16}$$

where $Lf = \frac{\partial f}{\partial y} b$. Therefore, comparing Eq. (8A.14) with Eq. (8A.16) leaves:

$$\left(\frac{\partial}{\partial\tau} + L\right) f = \phi, \quad \tau \in (t,T). \tag{8A.17}$$

Because $x_T = f(y_T, T) = 0$, the solution to Eq. (8A.17) is

$$f(y_t, t) = -\int_t^T \phi_\tau d\tau. \tag{8A.18}$$

Therefore, and extending Eq. (8A.15), we have the following result. Suppose that $x_T = f(y_T, T) = 0$, and that $\text{sign}\left(\phi_\tau\right) = \text{constant on } \tau \in (t,T)$. Thenn, by Eq. (8A.18):

$$\text{sign}\left(f\left(y_t, t\right)\right) = -\ \text{sign}\left(\phi_t\right).$$

These results can be extended to stochastic differential equations. Consider a more general version of Eq. (8A.17), one that typical arises in asset models with Brownian motions:

$$0 = \left(\frac{\partial}{\partial\tau} + L - k\right) G + D, \quad \tau \in (t,T). \tag{8A.19}$$

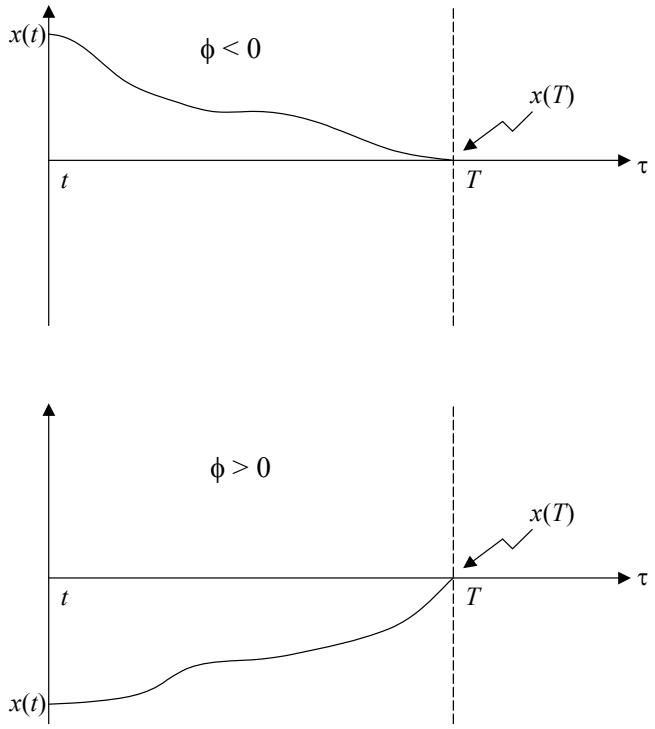


FIGURE 8A.1. Illustration of the maximum principle for ordinary differential equations

Let

$$y_\tau \equiv e^{-\int_t^\tau k_u du} G_\tau + \int_t^\tau e^{-\int_t^u k_s ds} D_u du.$$

We claim that under regularity conditions, if Eq. (8A.19) holds, y is then a martingale. Indeed,

$$\begin{aligned}
dy_\tau &= -k_\tau e^{-\int_t^\tau k_u du} G_\tau d\tau + e^{-\int_t^\tau k_u du} dG_\tau + e^{-\int_t^\tau k_u du} D_\tau d\tau \\
&= -k_\tau e^{-\int_t^\tau k_u du} G_\tau d\tau + e^{-\int_t^\tau k_u du} \left[\left(\frac{\partial}{\partial \tau} + L \right) G_\tau \right] d\tau + e^{-\int_t^\tau k_u du} D_\tau d\tau + dM_\tau \\
&= e^{-\int_t^\tau k_u du} \left[-k_\tau G_\tau + \left(\frac{\partial}{\partial \tau} + L \right) G_\tau + D_\tau \right] d\tau + dM_\tau \\
&= dM_\tau
\end{aligned}$$

where M_τ is a local martingale, and the last equality holds due to Eq. (8A.19). But if y is a martingale,

$$y_t = G_t = E_t\left(y_T\right) = E_t\left(e^{-\int_t^T k_u du} G_T\right) + E_t\left(\int_t^T e^{-\int_t^\tau k_s ds} D_\tau d\tau\right),$$

where E_t is the expectation taken given the information at time t. Therefore, conclude as follows. Assume that $G_T = 0$ and that $\text{sign}\left(D_\tau\right) = \text{constant}$ on $\tau \in (t, T)$. Then, $\text{sign}\left(y_\tau\right) = \text{sign}\left(D_t\right)$ for each τ.

We are now ready to use the maximum principle to study a few problems arising in this chapter.

8.11.2 Increasing risk

The following result studies the relation between contingent claims and the volatility of the fundamentals in a setting slightly more general than that in Eq. (8.12), viz

$$c(x,T-s)\equiv\mathbb{E}\left(e^{-\int_s^T\rho(x_t)dt}\cdot\psi(x_T)\middle|\,x_s=x\right). \tag{8A.20}$$

where x is a diffusion process with drift b and diffusion a. We have:

THEOREM 8A.1. (Increasing risk) *Consider two economies A and B with two fundamental volatilities a_A and a_B and let $\pi_i(x)\equiv a_i(x)\cdot\lambda^i(x)$ and $\rho_i(x)$ $(i=A,B)$ the corresponding risk-premium and discount rate. If $a_A>a_B$, the price c^A in economy A is lower than the price price c^B in economy B whenever for all $(x,\tau)\in\mathbb{R}\times[0,T]$,*

$$V(x,\tau)\equiv-\left(\rho_A(x)-\rho_B(x)\right)c^B(x,\tau)-\left(\pi_A(x)-\pi_B(x)\right)c_x^B(x,\tau)+\frac{1}{2}\left(a_A^2(x)-a_B^2(x)\right)c_{xx}^B(x,\tau)<0. \tag{8A.21}$$

If X is the price of a traded asset, $\pi_A=\pi_B$. If in addition ρ is constant, c is decreasing (increasing) in volatility whenever it is concave (convex) in x. This phenomenon is related to the "convexity effect" discussed in the main text. If X is not a traded risk, two additional effects are activated. The first relates to a discounting adjustment, and is apparent through the first term in the definition of V. The second effect reflects risk-premiums adjustments and corresponds to the second term.

PROOF OF THEOREM 8A.1. The function $c(x,T-s)$ in Eq. (8A.20) is solution to

$$0=-c_2(x,T-s)+L^*c(x,T-s)-\rho(x)c(x,T-s), \tag{8A.22}$$

subject to the boundary condition $c(x,0)=\psi(x)$, where $L^*c(x,u)=\frac{1}{2}a^2(x)\,c_{xx}(x,u)+b(x)c_x(x,u)$ and subscripts denote partial derivatives. Clearly, c^A and c^B are both solutions to Eq. (8A.22) but with different coefficients. Let $b_A(x)\equiv b_0(x)-\pi_A(x)$. The price difference $\Delta c(x,\tau)\equiv c^A(x,\tau)-c^B(x,\tau)$ is solution to

$$0=-\Delta c_2(x,T-s)+\frac{1}{2}\sigma_B^2(x)\Delta c_{xx}(x,T-s)+b_A(x)\Delta c_x(x,T-s)-\rho_A(x)\Delta c(x,T-s)+V(x,T-s),$$

with $\Delta c(x,0)=0$ and V is as in Eq. (8A.21) of the theorem. The result follows by the maximum principle. ■

8.11.3 Proof of Theorem 8.1

By differentiating Eq. (8A.22) two times with respect to x, we find that that $c^{(1)}(x,\tau)\equiv c_x(x,\tau)$ and $c^{(2)}(x,\tau)\equiv c_{xx}(x,\tau)$ are solutions to

$$\begin{aligned}0=&-c_2^{(1)}(x,T-s)+\frac{1}{2}a^2(x)c_{xx}^{(1)}(x,T-s)+\left(b(x)+\frac{1}{2}(a^2(x))'\right)c_x^{(1)}(x,T-s)\\&-\left(\rho(x)-b'(x)\right)c^{(1)}(x,T-s)-\rho'(x)c(x,T-s),\end{aligned}$$

and

$$\begin{aligned}0=&-c_2^{(2)}(x,T-s)+\frac{1}{2}a^2(x)c_{xx}^{(2)}(x,T-s)+\left(b(x)+(a^2(x))'\right)c_x^{(2)}(x,T-s)\\&-\left(\rho(x)-2b'(x)-\frac{1}{2}(a^2(x))''\right)c^{(2)}(x,T-s)\\&-\left(2\rho'(x)-b''(x)\right)c^{(1)}(x,T-s)-\rho''(x)c(x,T-s),\end{aligned}$$

subject to the boundary conditions $c^{(1)}(x,0) = \psi'(x)$ and $c^{(2)}(x,0) = \psi''(x)$. By results reviewed in Appendix 5 to this chapter, $c^{(1)}(x, T-s) > 0$ (resp. < 0) whenever $\psi'(x) > 0$ (resp. < 0) and $\rho'(x) < 0$ (resp. > 0). This completes the proof of Part (i) of the theorem. The proof of Part (ii) is obtained similarly. ■

REMARK 7A.2. (Alternative proof) An alternate proof can directly rely on the convexity of the *payoff function*, and a result due to Hajek's (1985). This result says that if ψ is increasing and convex, and x_1 and x_2 are two diffusion processes, both starting off from the same origin, with integrable drifts b_1 and b_2 and volatilities a_1 and a_2, then, $E[\psi(x_{1\tau})] \leq E[\psi(x_{2\tau})]$, whenever $b_1(t) \leq b_2(t)$ and $a_1(t) \leq a_2(t)$ for all $t \in (0, \tau)$. This result generalizes classic comparison theorems (e.g., Karatzas and Shreve, 1991, p. 291-295), where ψ is an increasing function and $a_1 \equiv a_2$.

REMARK 7A.3. (Bounds on convexity) An inspection of the proof of Theorem 8.1 reveals that concavity of prices holds under a less restrictive conditions than that given in Part (ii) of the theorem. We would only need that,

$$-\left(2\rho'(x) - b''(x)\right) c^{(1)}(x, T-s) - \rho''(x) c(x, T-s) < 0,$$

or that,

$$\rho''(x) > -\left(2\rho'(x) - b''(x)\right) \frac{c^{(1)}(x, T-s)}{c(x, T-s)}. \tag{8A.23}$$

Assuming that $\rho' < 0$, such that $\frac{c^{(1)}}{c} > 0$, the inequality (8A.23) imposes a theoretical upper bound to convexity of discount rates such that prices are concave.

8.11.4 A digression on a "macro-asset" derivative

We discuss an example regarding a highly conceptual and abstract asset, a "macro-asset" option, i.e., one asset that delivers payoffs linked to aggregate outcomes. We illustrate a few facts Theorem 8.1 predicts. Let c_t be the aggregate consumption process. The owner of the option has the right to receive a twice-differentiable payoff $\psi(c_T)$ at some date T, where ψ is increasing and convex. We assume that c_t is solution to:

$$\frac{dc_t}{c_t} = g_t dt,$$

and that consumption growth satisfies

$$dg_t = \varphi(g_t)\, dt + v(g_t)\, dW_t,$$

where φ and v are some well-behaved functions, and W is a standard Brownian motion. Let $p(c, g, t)$ be the option price when the state of the economy at t is $(c_t, g_t) = (c, g)$. We assume that p is as differentiable as needed below, that interest rates are constant, and that all agents are risk-neutral.

By the usual connection between partial differential equations and conditional expectations (see Chapter 5), the price $p(c, g, t)$ is solution to:

$$0 = \frac{\partial p}{\partial t} + gcp_c + \frac{1}{2}v^2 p_{gg} + \varphi p_g - rp, \quad \text{for all } c, g \text{ and } t \in [0, T), \tag{8A.24}$$

with boundary condition $p(c, g, T) = \psi(c)$ for all c, g (subscripts denote partial differentiation).

First, we study monotonicity of $p(c, g, t)$ with respect to both c and g, following two approaches. The first approach relies on the so-called *no-crossing property* of a diffusion process. Note that

$$p(c, g, t) = e^{-r(T-t)} E\left[\psi\left(c_t \cdot e^{\int_t^T g_u du}\right) \middle| c_t = c,\ g_t = g\right]. \tag{8A.25}$$

Since ψ is increasing, p is increasing in c. Furthermore, the no-crossing property of g implies that g_u is increasing in the initial condition g_t. Therefore, $p\left(c,g\right)$ is also increasing in g.

To analyze convexity of p with respect to g, differentiate Eq. (8A.24) and its boundary condition with respect to c, and find that $w \equiv p_c$ is solution to:

$$0 = \frac{\partial w}{\partial t} + gcw_c + \frac{1}{2}v^2 w_{gg} + \varphi w_g - (r-g)w, \quad \text{for all } c,g \text{ and } t \in [0,T),$$

with boundary condition $w(c,g,T) = \psi'(c)$, all c,g. The Feynman-Kac representation of the solution to the previous equation is:

$$p_c(c,g,t) = e^{-r(T-t)}E\left(e^{\int_t^T g_u du} \cdot \psi'(c_T) \middle| c_t = c,\ g_t = g \right), \tag{8A.26}$$

which is positive, by the assumption that $\psi' > 0$. This confirms the monotonicity properties established previously through no-crossing arguments. So when is $p(c,g,t)$ *convex* in g? By differentiating Eq. (8A.24) with respect to g, one obtains that $u \equiv p_g$ is solution to:

$$0 = \frac{\partial u}{\partial t} + gcu_c + \frac{1}{2}v^2 u_{gg} + (\varphi + \frac{1}{2}(v^2)')u_g - (r-\varphi')u + cp_c, \quad \text{for all } c,g \text{ and } t \in [0,T), \tag{8A.27}$$

with boundary condition $u(c,g,T) = 0$, all c,g. By the Feynman-Kac representation theorem,

$$p_g(c,g,t) = E\left(\int_t^T e^{-r(u-t)+\int_t^u \varphi'(g_s)ds} c_u p_c(c_u,g_u,u)du \middle| c_t = c,\ g_t = g \right).$$

By Eq. (8A.26), $p_c > 0$. Hence, p is increasing in g. We can now apply Theorem 8.1 and conclude that p is strictly convex in g whenever the drift function of g is weakly convex. Indeed, by differentiating Eq. (8A.27) with respect to g, we obtain that $\omega \equiv p_{gg}$ is solution to:

$$\frac{\partial \omega}{\partial t} + gc\omega_c + \frac{1}{2}v^2\omega_{gg} + (\varphi + (v^2)')\omega_g - (r - 2\varphi' - \frac{1}{2}(v^2)'')\omega + k, \quad \text{for all } c,g \text{ and } t \in [0,T), \tag{8A.28}$$

where

$$k(c,g,t) \equiv 2cp_{cg}(c,g,t) + \varphi''(g)p_g(c,g,t), \tag{8A.29}$$

and boundary condition $\omega(c,g,T) = 0$ for all c,g.

By Eq. (8A.26), we have that $p_c\left(c,g,t\right) = e^{-r(T-t)}E_t\left(\frac{c_T}{c_t}\psi'(c_T)\right)$. Thus p_c is increasing in g by the assumption that ψ is increasing and convex, and the no-crossing property of a diffusion, by which g_u is increasing in the initial condition g_t. Therefore, $p_{cg} > 0$. Furthermore, $p_g > 0$. Therefore, $k(c,g,t) > 0$ whenever $\varphi''(g) \geq 0$. By the Feynman-Kac theorem, then, p is *convex* in g whenever $\varphi''(g) \geq 0$.

The previous conclusions can hold even with a concave payoff function, say $\psi\left(c\right) = \ln c$. In this case, Eq. (8A.26) implies that $p_c(c,g,t) = e^{-r(T-t)}\frac{1}{c}$, such that the function k in Eq. (8A.29) collapses to, $k(c,g,t) = \varphi''(g)p_g(c,g,t)$. That is, the price $p\left(c,g,t\right)$ is convex (resp. concave) in g whenever φ is convex (concave) in g. Note, then, that the price is linear in g whenever $\varphi'' = 0$, as it can easily be verified by replacing $\psi(c) = \ln c$ into Eq. (8A.25), leaving:

$$p(c,g,t) = e^{-r(T-t)}\ln c + e^{-r(T-t)}E\left(\int_t^T g_u du \middle| g_t = g \right).$$

8.12 Appendix 6: Dynamics of habit in Campbell and Cochrane (1999)

We derive Eq. (8.34), by making a slightly more general assumption that the short-term rate we wish to come up with, is affine in $\ln s$, meaning that, the last two terms in Eq. (8.33) sum up to,

$$\gamma\left(1-\phi\right)\left(\bar{s}_l-\ln s\right)-\frac{1}{2}\gamma^2\sigma_0^2\left(1+l\left(s\right)\right)^2=-\text{const.}+b(\bar{s}_l-\ln s), \tag{8A.30}$$

for some b, and where const. is to be determined. The working paper version of Campbell and Cochrane (1999) considers exactly this case.

Define the log of the surplus ratio as

$$s_{l,\tau}\equiv\ln\left(1-e^{x_{l,\tau}-c_{l,\tau}}\right), \tag{8A.31}$$

where $s_l\equiv\ln s$, $x_l\equiv\ln x$ and $c_l\equiv\ln c$, and consider its first-order Taylor's expansion around the steady state $\overline{x_l-c_l}\equiv E\left(x_{l,\tau}-c_{l,\tau}\right)$,

$$s_{l,\tau}\approx\bar{s}_l+\left(1-\frac{1}{\bar{s}}\right)\left(x_{l,\tau}-c_{l,\tau}-\overline{x_l-c_l}\right), \tag{8A.32}$$

where $\bar{s}_l\equiv\ln\left(1-e^{\overline{x_l-c_l}}\right)$ and $\bar{s}\equiv e^{\bar{s}_l}$. Consider the discrete-time version of Eq. (8.30),

$$s_{l,t+1}-\bar{s}_l\approx\phi\left(s_{l,t}-\bar{s}_l\right)+l\left(s_t\right)\left(c_{l,t+1}-c_{l,t}-\left(g_0-\frac{1}{2}\sigma_0^2\right)\right), \tag{8A.33}$$

where $s_{l,t}\equiv s_l\left(\tau\right)$. Replacing Eq. (8A.32), evaluated at $\tau=t$ and $\tau=t+1$, into both sides of the previous approximation, and rearranging terms, leaves:

$$x_{l,t+1}-c_{l,t+1}-\overline{x_l-c_l}=\phi\left(x_{l,t}-c_{l,t}-\overline{x_l-c_l}\right)+\frac{l\left(s_t\right)}{1-\frac{1}{\bar{s}}}\left(c_{l,t+1}-c_{l,t}-\left(g_0-\frac{1}{2}\sigma_0^2\right)\right). \tag{8A.34}$$

The function l in Eq. (8.34) is found by imposing the following three conditions, where the first is a slight generalization of that mentioned in the main text, and the remaining two are the last two conditions in the main text:

- First, the short-term rate in Eq. (8.33) is affine in $\ln s$, i.e. Eq. (8A.30) holds, such that:

$$l\left(s\right)=\sqrt{2\frac{1}{\gamma^2\sigma_0^2}\left(\gamma\left(1-\phi\right)-b\right)\left(\bar{s}_l-\ln s\right)+\frac{2}{\gamma^2\sigma_0^2}\text{const.}}-1. \tag{8A.35}$$

- Second, habit is predetermined at the steady state, meaning that $x_{l,t+1}$ does not change with $c_{l,t+1}$, which by Eq. (8A.34), it does not, when:

$$l\left(\bar{s}\right)=\frac{1}{\bar{s}}-1. \tag{8A.36}$$

  Evaluating Eq. (8A.35) at the steady state $\bar{s}$, and using the previous condition delivers, $\frac{2}{\gamma^2\sigma_0^2}\text{const.}=\frac{1}{\bar{s}^2}$, such that Eq. (8A.35) is,

$$l\left(s\right)=\sqrt{2\frac{1}{\gamma^2\sigma_0^2}\left(\gamma\left(1-\phi\right)-b\right)\left(\bar{s}_l-\ln s\right)+\frac{1}{\bar{s}^2}}-1. \tag{8A.37}$$

- Third, habit is predetermined near the steady state, meaning that,

$$\frac{d}{ds_l}\left(\frac{dx_l}{dc_l}\right)\Big|_{s=\bar{s}_l}=0. \tag{8A.38}$$

We, then, need to find the dynamics of $x_{l,t+1}$, expressed as a function of $c_{l,t+1}$. By the definition of the log-surplus consumption ratio in Eq. (8A.31), we have that $x_{l,t+1}=\ln\left(1-e^{\sigma\left(c_{l,t+1}\right)}\right)+c_{l,t+1}$, where $\sigma\left(c_{l,t+1}\right)\equiv s_{l,t+1}$ and $s_{l,t+1}$ is as in Eq. (8A.33), such that, using Eq. (8A.33):

$$\frac{dx_{l,t+1}}{dc_{l,t+1}}=1-\frac{1}{e^{-\sigma\left(c_{l,t+1}\right)}-1}\sigma'\left(c_{l,t+1}\right)=1-\frac{l_o\left(s_{l,t}\right)}{e^{-s_{l,t+1}}-1},$$

where we have set $l_o\left(s_l\right)\equiv l\left(s\right)$. Therefore, Eq. (8A.38) is: $\frac{d}{ds_l}\left(1-\frac{l_o(s_l)}{e^{-s_l}}\right)\Big|_{s=\bar{s}_l}=0$, which leaves, after simple computation, and using Eq. (8A.36),

$$l_o'\left(\bar{s}_l\right)=-\frac{1}{\bar{s}}.$$

By taking the derivative in Eq. (8A.37), and replacing into the left hand side of the previous equation, and solving for $\bar{s}_l$, yields,

$$\bar{s}=\sigma_0\sqrt{\frac{\gamma^2}{\gamma\left(1-\phi\right)-b}},$$

which is the expression of the main text, for $b=0$. By replacing this expression of $\bar{s}$ into Eq. (8A.37), leaves Eq. (8.34) in the main text.

Finally, note that, now, the expression of the short-term rate can be found after simple computations:

$$r\left(s\right)=\delta+\gamma\left(g_0-\frac{1}{2}\sigma_0^2\right)-\frac{1}{2}\left(\gamma\left(1-\phi\right)-b\right)+b\left(\bar{s}-\ln s\right).$$

8.13 Appendix 7: An algorithm to simulate discrete-time pricing models

Consider the pricing equation,

$$S = E\left[m \cdot \left(S' + D'\right)\right], \quad m = \beta \frac{u_c\left(D', x'\right)}{u_c\left(D, x\right)} = \beta \left(\frac{s'}{s}\right)^{-\gamma} \left(\frac{D'}{D}\right)^{-\gamma}.$$

The price-dividend ratio, $p \equiv S/D$ say, satisfies:

$$p = E\left[m \frac{D'}{D}\left(1 + p'\right)\right], \quad \frac{D'}{D} = e^{g_0 + w}.$$

The previous equation is a functional equation in $p\left(s\right)$, say:

$$p(s) = E\left[g\left(s', s\right)\left(1 + p\left(s'\right)\right) \middle| s\right], \quad g\left(s', s\right) = \beta \left(\frac{s'}{s}\right)^{-\gamma} \left(\frac{D'}{D}\right)^{1-\gamma}.$$

A numerical solution can be implemented as follows. Create a grid and define $p_j = p\left(s_j\right)$, $j = 1, \cdots, N$, for some N. We have,

$$\left[\begin{array}{c} p_1 \\ \vdots \\ p_N \end{array}\right] = \left[\begin{array}{c} b_1 \\ \vdots \\ b_N \end{array}\right] + \left[\begin{array}{ccc} a_{11} & \cdots & a_{N1} \\ \vdots & \ddots & \vdots \\ a_{1N} & \cdots & a_{NN} \end{array}\right] \left[\begin{array}{c} p_1 \\ \vdots \\ p_N \end{array}\right],$$

$$b_i = \sum_{j=1}^{N} a_{ji}, \quad a_{ji} = g_{ji} \cdot p_{ji}, \quad g_{ji} = g\left(s_j, s_i\right), \quad p_{ji} = \Pr\left(s_j \middle| s_i\right) \cdot \Delta s,$$

where Δs is the integration step, $s_1 = s_{\min}$, $s_N = s_{\max}$, $s_{\min}$ and $s_{\max}$ are the boundaries in the approximation, and $\Pr\left(s_j \middle| s_i\right)$ is the transition density from state i to state j - in this case, a Gaussian transition density. Let $p = \left[p_1 \ \cdots \ p_N\right]^\top$, $b = \left[b_1 \ \cdots \ b_N\right]^\top$, and let A be a matrix with elements a_{ji}. The solution is,

$$p = \left(I - A\right)^{-1} b. \tag{8A.39}$$

The model can be simulated in the following manner. Let $\underline{s}$ and $\bar{s}$ be the boundaries of the underlying state process. Fix $\Delta s = \dfrac{\bar{s} - \underline{s}}{N}$. Draw states. State s^* is drawn. Then,

1. If $\min\left\{s^* - \underline{s}, \bar{s} - s^*\right\} = s^* - \underline{s}$, let k be the smallest integer close to $\frac{s^* - \underline{s}}{\Delta s}$. Let $s_{\min} = s^* - k\Delta s$, and $s_{\max} = s_{\min} + N \cdot \Delta s$.
2. If $\min\left\{s^* - \underline{s}, \bar{s} - s^*\right\} = \bar{s} - s^*$, let k be the biggest integer close to $\frac{s^* - \underline{s}}{\Delta s}$. Let $s_{\max} = s^* + k\Delta s$, and $s_{\min} = s_{\max} - N \cdot \Delta s$.

The previous algorithm avoids interpolations, and ensures that during the simulations, p is computed in correspondence of exactly the state s^* that is drawn. Precisely, once s^* is drawn, we proceed to the following two steps: (i) create the corresponding grid $s_1 = s_{\min}$, $s_2 = s_{\min} + \Delta s, \cdots, s_N = s_{\max}$ according to the previous rules; and (ii) compute the solution from Eq. (8A.39). In this way, one has $p\left(s^*\right)$ at hand—the simulated P/D ratio when state s^* is drawn.

8.14 Appendix 8: Heuristic details of learning in continuous time

We derive Eq. (8.42). The probability the dividend process in Eq. (8.40) has the "good" drift $\theta = A$, conditional on the observation of $(D_t)_{t\leq\tau}$ is,

$$\pi(\tau, D_\tau) = \frac{p\phi(D_\tau - A\tau)}{p\phi(D_\tau - A\tau) + (1-p)\phi(D_\tau + A\tau)} = \frac{p}{p + (1-p)e^{-\frac{2AD_\tau}{\sigma_0^2}}},$$

where the first equality follows by a straightforward generalization of the arguments leading to Eq. (8.39), and the second holds by the assumption that $\sigma_0 Z_\tau$ is Gaussian with mean zero and variance $\sigma_0^2\tau$. Hence, we have that $\pi(\tau, D_\tau)$ is a function of the dividend only. To simplify notation, let $\pi(D_\tau) \equiv \pi(\tau, D_\tau)$. We have,

$$\pi'(D) = 2\pi^2(D)\frac{1-p}{p}e^{-\frac{2AD}{\sigma_0^2}}\frac{A}{\sigma_0^2}, \quad \pi''(D) = \pi'(D)\frac{2A}{\sigma_0^2}\left(\frac{1-p}{p}2\pi(D)e^{-\frac{2AD}{\sigma_0^2}} - 1\right)$$

Note that,

$$\frac{1-\pi(D)}{\pi(D)} = \frac{(1-p)e^{-\frac{2AD}{\sigma_0^2}}}{p}, \quad \text{and} \quad g_\tau = E\left(\theta|(D_t)_{t\leq\tau}\right) = A(2\pi(D_\tau) - 1),$$

such that,

$$\pi'(D) = \frac{2A}{\sigma_0^2}\pi(D)(1-\pi(D)), \quad \pi''(D) = -\pi'(D)\frac{2}{\sigma_0^2}g. \tag{8A.40}$$

Next, apply Itô's lemma to the function $\pi(D_\tau)$, where D_τ is solution to Eq. (8.41). By the expressions in in Eq. (8A.40),

$$\begin{aligned} d\pi(D_\tau) &= \pi'(D_\tau)dD_\tau + \frac{1}{2}\pi''(D_\tau)dD_\tau^2 \\ &= \left(\pi'(D_\tau)g_\tau + \frac{1}{2}\pi''(D_\tau)\sigma_0^2\right)d\tau + \pi'(D_\tau)\sigma_0 dW_\tau \\ &= \frac{2A}{\sigma_0}\pi(D_\tau)(1-\pi(D_\tau))dW_\tau. \quad \blacksquare \end{aligned}$$

8.15 Appendix 9: Linear regime-switching economies

We prove a claim made in Section 8.5.3. Consider a complete markets economy where dividends, consumption, and signals (D_t, c_t, a_t) satisfy:

$$\begin{pmatrix} dD_t/\,D_t \\ dc_t/\,c_t \\ da_t \end{pmatrix} = \begin{pmatrix} \theta \\ \bar{g} \\ \theta \end{pmatrix} dt + \begin{pmatrix} \bar{\sigma}_0 \\ \bar{\sigma} \\ \bar{\sigma}_a \end{pmatrix} dZ_t, \qquad \begin{array}{lcl} \bar{\sigma}_0 & \equiv & (\sigma_0 \;\; 0 \;\; 0) \\ \bar{\sigma} & \equiv & (\sigma_1 \;\; \sigma_2 \;\; \sigma_3) \\ \bar{\sigma}_a & \equiv & (\sigma_4 \;\; \sigma_5 \;\; \sigma_6) \end{array}$$

and $Z = (Z_1 \;\; Z_2 \;\; Z_3)^\top$ denotes a vector standard Brownian motion, with θ being a two-state $(\bar{\theta}, \underline{\theta})$ Markov chain, and $\bar{g}, \sigma_i$ are constants. Let $\sigma_3\sigma_5 \neq \sigma_2\sigma_6$. Then, there are *no* CRRA representative agent equilibria in which price-dividend ratios are convex in expected dividend growth. To demonstrate this claim, we apply the filtering results of Liptser and Shiryaev (2001) (Vol. I), and find that the previous economy is isomorphic to one in which,

$$\begin{pmatrix} dD_t/\,D_t \\ dc_t/\,c_t \\ dy_t \end{pmatrix} = \begin{pmatrix} y_t \\ \bar{g} \\ k\,(y^* - y_t) \end{pmatrix} dt + \begin{pmatrix} \bar{\sigma}_0 \\ \bar{\sigma} \\ \left(\bar{\theta} - y_t\right)(y_t - \underline{\theta}) \cdot \bar{\sigma}_y \end{pmatrix} dW_t,$$

where $W = (W_1 \;\; W_2 \;\; W_3)^\top$ is a vector standard Brownian motion, $k, y^* > 0$, and $\bar{\sigma}_y$ is some vector satisfying $\bar{\sigma} \cdot \bar{\sigma}_y = 0$. Standard arguments lead that in equilibrium, the short-term rate, r, is constant and $\lambda_i \propto \sigma_i$ $(i = 1, 2, 3)$. The claim follows by $\bar{\sigma} \perp \bar{\sigma}_y$ and Theorem 8.1.

8.16 Appendix 10: Bond price convexity revisited

Consider a short-term rate process r_τ, and let $u(r_0, T)$ be the price of a bond expiring at time T when the current short-term rate is r_0:

$$u(r_0, T) = \mathbb{E}\left(e^{-\int_0^T r_\tau d\tau} \middle| r_0 \right).$$

As pointed out in Section 8.6, Theorem 8.1-(ii) implies that in scalar diffusion models of the short-term rate, such as those dealt with in Chapter 13, one has $u_{11}(r_0, T) < 0$ whenever $b'' < 2$, where b is the risk-neutralized drift of r. This result, obtained by Mele (2003), can be proved through the Feynman-Kac representation of u_{11}, and a similar proof can be used to show Theorem 8.1-(ii). This appendix provides a more intuitive derivation under a set of simplifying assumptions. By Eq. (6) p. 685 in Mele (2003),

$$u_{11}(r_0, T) = \mathbb{E}\left[\left(\left(\int_0^T \frac{\partial r_\tau}{\partial r_0} d\tau \right)^2 - \int_0^T \frac{\partial^2 r_\tau}{\partial r_0^2} d\tau \right) e^{-\int_0^T r_\tau d\tau} \right].$$

Hence $u_{11}(r_0, T) > 0$ whenever

$$\int_0^T \frac{\partial^2 r_\tau}{\partial r_0^2} d\tau < \left(\int_0^T \frac{\partial r_\tau}{\partial r_0} d\tau \right)^2. \tag{8A.41}$$

To keep the presentation simple, assume r_τ is solution to:

$$dr_\tau = b(r_\tau)dt + a_0 r_\tau dW_\tau,$$

where a_0 is a constant. We have,

$$\frac{\partial r_\tau}{\partial r_0} = \exp\left(\int_0^\tau b'(r_u)du - \frac{1}{2}a_0^2\tau + a_0 W_\tau \right) \quad \text{and} \quad \frac{\partial^2 r_\tau}{\partial r_0^2} = \frac{\partial r_\tau}{\partial r_0}\left[\int_0^\tau b''(r_u)\frac{\partial r_u}{\partial r_0} du \right]. \tag{8A.42}$$

Therefore, if $b'' < 0$, then $\partial^2 r_\tau / \partial r_0^2 < 0$, and by the inequality in (13.63), $u_{11} > 0$.

This result can be improved. Suppose that $b'' < 2$, instead of $b'' < 0$. By the second of Eqs. (8A.42),

$$\frac{\partial^2 r_\tau}{\partial r_0^2} < 2\frac{\partial r_\tau}{\partial r_0} \int_0^\tau \frac{\partial r_u}{\partial r_0} du,$$

and consequently,

$$\int_0^T \frac{\partial^2 r_\tau}{\partial r_0^2} d\tau < 2\int_0^T \frac{\partial r_\tau}{\partial r_0}\left(\int_0^\tau \frac{\partial r_u}{\partial r_0} du \right) d\tau = \left(\int_0^T \frac{\partial r_u}{\partial r_0} du \right)^2,$$

which is the inequality in (13.63).

References

Abel, A.B. (1990): "Asset Prices under Habit Formation and Catching Up with the Joneses." *American Economic Review Papers and Proceedings* 80, 38-42.

Andersen, T. G., T. Bollerslev and F.X. Diebold (2002): "Parametric and Nonparametric Volatility Measurement." Forthcoming in Aït-Sahalia, Y. and L. P. Hansen (Eds.): *Handbook of Financial Econometrics.*

Bajeux-Besnainou, I. and J.-C. Rochet (1996): "Dynamic Spanning: Are Options an Appropriate Instrument?" *Mathematical Finance* 6, 1-16.

Barberis, N., M. Huang and T. Santos (2001): "Prospect Theory and Asset Prices." *Quarterly Journal of Economics* 116, 1-53.

Barsky, R.B. and J.B. De Long (1990): "Bull and Bear Markets in the Twentieth Century." *Journal of Economic History* 50, 265-281.

Barsky, R.B. and J.B. De Long (1993): "Why Does the Stock Market Fluctuate?" *Quarterly Journal of Economics* 108, 291-311.

Bergman, Y.Z., B.D. Grundy, and Z. Wiener (1996): "General Properties of Option Prices." *Journal of Finance* 51, 1573-1610.

Black, F. and M. Scholes (1973): "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81, 637-659.

Brandt, M.W. and Q. Kang (2004): "On the Relationship between the Conditional Mean and Volatility of Stock Returns: A Latent VAR Approach." *Journal of Financial Economics* 72, 217-257.

Brennan, M.J. and Y. Xia (2001): "Stock Price Volatility and Equity Premium." *Journal of Monetary Economics* 47, 249-283.

Brunnermeier, M.K. and S. Nagel (2007): "Do Wealth Fluctuations Generate Time-Varying Risk Aversion? Micro-Evidence on Individuals' Asset Allocation." Forthcoming in *American Economic Review.*

Campbell, J.Y. (2003): "Consumption-Based Asset Pricing." In: Constantinides, G.M., M. Harris and R. M. Stulz (Editors): *Handbook of the Economics of Finance* (Volume 1B: Chapter 13), 803-887.

Campbell, J.Y., and J.H. Cochrane (1999): "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior." *Journal of Political Economy* 107, 205-251.

Christiansen, C., M. Schmeling and A. Schrimpf (2012): "A Comprehensive Look at Financial Volatility Prediction by Economic Variables." *Journal of Applied Econometrics* 27, 956-977.

Clark, T.E. and K.D. West (2007): "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models." *Journal of Econometrics* 138, 291-311.

Constantinides, G.M. (1990): "Habit Formation: A Resolution of the Equity Premium Puzzle." *Journal of Political Economy* 98, 519-543.

Corradi, V., W. Distaso and A. Mele (2013): "Macroeconomic Determinants of Stock Volatility and Volatility Premiums." *Journal of Monetary Economics* 60, 203-220.

David, A. (1997): "Fluctuating Confidence in Stock Markets: Implications for Returns and Volatility." *Journal of Financial and Quantitative Analysis* 32, 427-462.

Detemple, J.B. (1986): "Asset Pricing in a Production Economy with Incomplete Information." *Journal of Finance* 41, 383-391.

Duesenberry, J.S. (1949): *Income, Saving, and the Theory of Consumer Behavior*. Cambridge, Mass.: Harvard University Press.

El Karoui, N., M. Jeanblanc-Picqué and S.E. Shreve (1998): "Robustness of the Black and Scholes Formula." *Mathematical Finance* 8, 93-126.

Fama, E.F. and K.R. French (1989): "Business Conditions and Expected Returns on Stocks and Bonds." *Journal of Financial Economics* 25, 23-49.

Fama, E.F. (2014): "Nobel Lecture: Two Pillars of Asset Pricing." *American Economic Review* 104, 1467-1485.

Ferson, W.E. and C.R. Harvey (1991): "The Variation of Economic Risk Premiums." *Journal of Political Economy* 99, 385-415.

Fornari, F. and A. Mele (2013): "Financial Volatility and Real Economic Activity." *Journal of Financial Management, Markets and Institutions* 1, 155-198.

Gabaix, X. (2009): "Linearity-Generating Processes: A Modelling Tool Yielding Closed Forms for Asset Prices." Working paper New York University.

Gennotte, G. (1986): "Optimal Portfolio Choice Under Incomplete Information." *Journal of Finance* 41, 733-746.

Giacomini, R. and H. White (2006): "Tests of Conditional Predictive Ability." *Econometrica* 74, 1545-1578.

Glosten, L., R. Jagannathan and D. Runkle (1993): "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks." *Journal of Finance* 48, 1779-1801.

Gordon, M. (1962): *The Investment, Financing, and Valuation of the Corporation*. Homewood, IL: Irwin.

Hajek, B. (1985): "Mean Stochastic Comparison of Diffusions." *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete* 68, 315-329.

Hamilton, J.D. and G. Lin (1996): "Stock Market Volatility and the Business Cycle." *Journal of Applied Econometrics* 11, 573-593.

Huang, C.-F. and Pagès, H. (1992): "Optimal Consumption and Portfolio Policies with an Infinite Horizon: Existence and Convergence." *Annals of Applied Probability* 2, 36-64.

Jagannathan, R. (1984): "Call Options and the Risk of Underlying Securities." *Journal of Financial Economics* 13, 425-434.

Karatzas, I. and S.E. Shreve (1991): *Brownian Motion and Stochastic Calculus.* Berlin: Springer Verlag.

Kijima, M. (2002): "Monotonicity and Convexity of Option Prices Revisited." *Mathematical Finance* 12, 411-426.

Liptser, R. S. and A. N. Shiryaev (2001): *Statistics of Random Processes.* Berlin, Springer-Verlag. [2001a: *Vol.* I (*General Theory*). 2001b: *Vol.* II (*Applications*).]

Ljungqvist, L. and H. Uhlig (2000): "Tax Policy and Aggregate Demand Management under Catching Up with the Joneses." *American Economic Review* 90, 356-366.

Mehra, R. and E.C. Prescott (2003): "The Equity Premium in Retrospect." In Constantinides, G.M., M. Harris and R. M. Stulz (Editors): *Handbook of the Economics of Finance* (Volume 1B, chapter 14), 889-938.

Mele, A. (2003): "Fundamental Properties of Bond Prices in Models of the Short-Term Rate." *Review of Financial Studies* 16, 679-716.

Mele, A. (2005): "Rational Stock Market Fluctuations." WP FMG-LSE.

Mele, A. (2007): "Asymmetric Stock Market Volatility and the Cyclical Befhavior of Expected Returns." *Journal of Financial Economics* 86, 446-478.

Menzly, L., T. Santos and P. Veronesi (2004): "Understanding Predictability." *Journal of Political Economy* 111, 1, 1-47.

Paye, B.S. (2012): **"**Déjà Vol: Predictive Regressions for Aggregate Stock Market Volatility Using Macroeconomic Variables." *Journal of Financial Economics* 106, 527-546.

Romano, M. and N. Touzi (1997): "Contingent Claims and Market Completeness in a Stochastic Volatility Model." *Mathematical Finance* 7, 399-412.

Rothschild, M. and J. Stiglitz (1970): "Increasing Risk: I. A Definition." *Journal of Economic Theory* 2, 225-243.

Rothschild, M. and J. Stiglitz (1971): "Increasing Risk: II. Its Economic Consequences." *Journal of Economic Theory* 5, 66-84.

Ryder, H.E. and G.M. Heal (1973): "Optimal Growth with Intertemporally Dependent Preferences." *Review of Economic Studies* 40, 1-33.

Schwert, G.W. (1989a): "Why Does Stock Market Volatility Change Over Time?" *Journal of Finance* 44, 1115-1153.

Schwert, G.W. (1989b): "Business Cycles, Financial Crises and Stock Volatility." *Carnegie-Rochester Conference Series on Public Policy* 31, 83-125.

Shiller, R.J. (2014): "Nobel Lecture: Speculative Asset Prices." *American Economic Review* 104, 1486-1517.

Sundaresan, S.M. (1989): "Intertemporally Dependent Preferences and the Volatility of Consumption and Wealth." *Review of Financial Studies* 2, 73-89.

Timmermann, A. (1993): "How Learning in Financial Markets Generates Excess Volatility and Predictability in Stock Prices." *Quarterly Journal of Economics* 108, 1135-1145.

Timmermann, A. (1996): "Excess Volatility and Return Predictability of Stock Returns in Autoregressive Dividend Models with Learning." *Review of Economic Studies* 63, 523-577.

Veronesi, P. (1999): "Stock Market Overreaction to Bad News in Good Times: A Rational Expectations Equilibrium Model." *Review of Financial Studies* 12, 975-1007.

Veronesi, P. (2000): "How Does Information Quality Affect Stock Returns?" *Journal of Finance* 55, 807-837.

Wang, S. (1993): "The Integrability Problem of Asset Prices." *Journal of Economic Theory* 59, 199-213.

9
Macrofinance

9.1 Introduction

This chapter discusses models that aim to address the empirical puzzles surveyed in the previous two chapters. The most prominent are the "equity premium puzzle," its pattern over the business cycle as well as that of aggregate stock volatility. Because these puzzles mainly relate to the behavior of aggregate variables, it is natural to attempt to provide explanations hinging upon macroeconomic variables such as consumption or output, whence the title of this chapter.

The equity premium puzzle is the difficulty of the neoclassical model to predict expected returns that are quantitatively consistent with those in the data. In particular, Chapter 7 explains that an implausibly high level of risk-aversion is needed to reconcile models with data. Moreover, a high risk-aversion implies a low elasticity of intertemporal substitution and, hence, an implausibly high volatility of the interest rates, which gives rise to the "interest rate puzzle."

In the early attempts to address the equity premium puzzle, the assumption of a representative agent with CRRA preferences was replaced with that of a representative agent with non-expected utility. In the non-expected utility framework, risk-aversion can be understood independently of the elasticity of intertemporal substitution. This approach, described in Section 9.2, does not necessarily lead to address the equity premium puzzle. We shall explain that within the non-expected utility model, a possible resolution of the puzzle requires that the price-dividend ratio be affected by a number of state variables. One instance of state variables identified in the literature is a long-run risk, defined as a mechanism capable of turning even a small shock into an economic damage perduring for years. For example, it has been argued, expected consumption growth is highly persistent, such that shocks to expected growth have long-lasting implications. While the model has the potential to address the equity premium puzzle, we shall explain that there are nuances regarding the realism of this precise mechanism.

Section 9.3 explores channels that address the equity premium puzzle based on a variant of habit formation. The external habit formation model reviewed in Chapter 8 relies on the existence of a representative agent with high risk-aversion. Alternatively, one can consider economies with heterogenous agents, in which each agent has a consumption reference that he wants to benchmark to ("catching up with the Joneses"). The economy is heterogeneous in that each agent displays a different curvature of his utility function. In bad times, when markets

are down, the less risk-averse agents are those who suffer the most, reflecting their previous relatively more aggressive investment decisions. Thus, their relative wealth and, hence, social weight decreases. That is, the aggregate risk-aversion increases in bad times, leading to the countercyclical mechanisms explained in the previous chapter. This model has the potential to explain the equity premium, but is also subject to a number of caveats.

Section 9.4 describes economies in which agents can be hit by idiosyncratic shocks, a risk that they cannot hedge against through security trading. How is it that idiosyncratic shocks could affect asset prices? The key assumption of the models we survey is that although agents are ex-ante the same, they will be affected by shocks that have different amplitude. A job loss is one important instance that illustrates how these models work. Recessions do not necessarily affect agents in the same way. Some agents can be hurt more than others. The possibility of a job loss can, then, induce agents to act prudently while investing in the stock market. That is, risk-premiums are countercyclical. This explanation can lead to a potential resolution of the equity premium puzzle, theoretically at least.

While idiosyncratic risk is a clear explanation, its empirical implications do not seem to be entirely exhaustive. Note that a natural hedge against idiosyncratic risk is self-insurance, that is, the ability to save in good times to cope with adversities possibly occurring in bad. Agents might actually eliminate a large portion of idiosyncratic risk by insuring themselves while having access to capital markets, thereby making idiosyncratic risk practically irrelevant to the explanation of the equity premium. For idiosyncratic risk to really matter, we would need to observe a large and persistent idiosyncratic risk, or capital markets transactions to be so expensive to prevent agents from implementing self-insurance plans. But empirically, idiosyncratic risks do not appear to be as large and persistent, and market transaction costs are not as large, as required by our standard models with idiosyncratic risk. Naturally, there are historical instances in which idiosyncratic risk may have a better appeal: the Great Recession occurring after the 2007 crisis is one teaching us idiosyncratic risks could be quite persistent.

Section 9.5 considers an alternative channel, market incompleteness. Economies with incomplete markets are not necessarily consistent with a sizeable equity premium. If agents have comparable access to equity markets, they share the same level of risk, such that the premium they require is similar to that in economies with complete markets. But market incompleteness has the potential to resolve the puzzle, when a large fraction of the agents do not participate. The mechanism is simple. If the markets are being shut down to a large proportion of agents, the agents who participate are concerned for the aggregate macroeconomic risk they bear. This concern leads them to require a sizeable premium.

Sections 9.6 and 9.7 deal with economies where agents have differences in beliefs and also uncertainty (as opposed to risk) regarding the fundamentals. [In progress]

Section 9.8 deals with issues arising within production-based economies. In these economies, consumption is endogenous, and an increase in the agents risk aversion might actually lead to a decreased consumption volatility. One additional important difficulty in these economies is that capital supply is infinitely elastic, such that the price of capital is quite smooth. To increase capital price volatility, we need hindrances in the capital formation process, such as the presence of adjustment costs, or an added volatility of the demand for capital, obtained for example when agents have habit formation over their consumption plans. Both rigidities in the capital formation process and volatility in the demand of capital are needed, in order to explain the equity premium.

Section 9.9: In progress.

Section 9.10 presents a simple model to assess a very old hypothesis in financial economics: the extent to which equity volatility can be explained by firms leverage.

Section 9.11 reviews recent models capable to explain the cross-section of asset returns, relying on multiple trees.

Section 9.12 surveys predictions that the previous models make about the yield curve. Chapter 13 contains additional models and explanation of the yield curve and its relation to macroeconomic developments.

Section 9.13 deals with an intriguing topic: what do financial economists and macroeconomists have really in common? Granted, in many of the models surveyed in this chapter, we aim to understand asset prices in a context of the business cycle. However, these models are built upon a revamp of many of the assumptions underlying the neo-classical paradigm. Yet macroeconomists do not necessarily seem to acknowledge our asset pricing lessons. Are macroeconomists mistaken? Or is there a case for a modern version of a *dichotomy* between the real and the financial spheres of the economy? A simple model shows there is a potential for the hypothesis of a separation between finance and macroeconomics. However, this potential is seriously undermined once we allow financial markets to feed back the economy.

Section 9.14 concludes the chapter and reviews explanations of macroeconomic developments relying on such asset price feedbacks. [In progress]

9.2 Non-expected utility

The standard intertemporal additive separable utility function confounds intertemporal substitution effects and attitudes towards risk. This property is problematic. Epstein and Zin (1989, 1991) and Weil (1989) consider a class of recursive, but not necessarily expected utility, preferences. In this section, we present some details of this approach, without insisting on the theoretic underpinnings (such as those in Epstein and Zin, 1989). We provide basic definitions and derivations for this class of preferences, and focus on their asset pricing implications.

9.2.1 Recursive formulations

Consider the following standard case, in which a decision maker displays additive preferences, in that his continuation utility V_t satisfies:

$$V_t = u(c_t) + \beta E_t(V_{t+1}) \equiv u(c_t) + \beta u(\hat{v}_{t+1}), \tag{9.1}$$

where u is a von Neumann-Morgenstern utility function, and $\hat{v}_{t+1}$ is the certainty equivalent in terms of consumption, i.e., $\hat{v}_{t+1} : u(\hat{v}_{t+1}) = E_t(V_{t+1})$. Next, let us define the certainty equivalent continuation utility expressed in consumption units, as $v_t : V_t = u(v_t)$, such that we can write Eq. (9.1) as:

$$v_t = W(c_t, \hat{v}_{t+1}) \equiv u^{-1}(u(c_t) + \beta u(\hat{v}_{t+1})),$$

where W is the *aggregator*, and where by construction, $\hat{v}_{t+1}$ is also the certainty-equivalent utility at $t+1$, defined as $u(\hat{v}_{t+1}) = E_t[u(v_{t+1})]$.

Consider, for example, the standard CRRA utility function, in which case the aggregator is:

$$W(c, \hat{v}) = \left(c^{1-\gamma} + \beta \hat{v}^{1-\gamma}\right)^{\frac{1}{1-\gamma}}.$$

In this formulation, the decision maker attitude vis-à-vis risk is encoded into the certainty equivalent $\hat{v}_{t+1}$ through the utility u and, as is well-known (see Chapter 3), in this specific

example, the CRRA γ is exactly equal to the reciprocal of the elasticity of intertemporal substitution (EIS). But there are no economic reasons for this coincidence. An alternative assumption regarding the aggregator W is that CRRA and EIS are not tied up as in the previous example. Consider the following aggregator:

$$W(c,\hat{v}) = (c^{\rho} + \beta \hat{v}^{\rho})^{1/\rho} \quad \text{and} \quad u(\hat{v}) = \frac{\hat{v}^{1-\gamma}}{1-\gamma}, \tag{9.2}$$

for three positive constants ρ, γ and β. In this formulation, risk-attitudes for static wealth gambles have still the classical CRRA flavor. Precisely, we say that γ is the CRRA for static wealth gambles, and $\psi \equiv (1-\rho)^{-1}$ is the EIS. We still have

$$\hat{v}_{t+1} = u^{-1}\left[E_t\left(u\left(v_{t+1}\right)\right)\right] = \left(E_t\left(v_{t+1}^{1-\gamma}\right)\right)^{\frac{1}{1-\gamma}},$$

and the parametrization for W in Eq. (9.2) then implies that

$$v_t = \left(c_t^{\rho} + \beta\left(E_t(v_{t+1}^{1-\gamma})\right)^{\frac{\rho}{1-\gamma}}\right)^{\frac{1}{\rho}}. \tag{9.3}$$

Naturally, in the absence of uncertainty, $v_t^{\rho} = c_t^{\rho} + \beta v_{t+1}^{\rho}$, another clear illustration that ψ is the EIS. Also, it is straightforward to verify that as soon as the CRRA equals the reciprocal of the EIS, i.e., $\gamma = 1-\rho$, v_t in Eq. (9.3) collapses to the standard intertemporal additive utility, $v_t^{1-\gamma} = E\left(\sum_{n=0}^{\infty} \beta^n c_{t+n}^{1-\gamma}\right)$.

Let us return to the general case. Let us recall that v_t is the certainty equivalent continuation utility expressed in consumption units; in some cases (e.g., in Section 9.13), it is both convenient and intuitive to formulate problems in terms of the continuation utility $V_t = u(v_t)$, such that we can write Eq. (9.3) as:

$$V_t = \frac{1}{1-\gamma}\left(c_t^{\rho} + \beta\left((1-\gamma)E_t(V_{t+1})\right)^{\frac{\rho}{1-\gamma}}\right)^{\frac{1-\gamma}{\rho}}. \tag{9.4}$$

In Appendix 1, we shall rely on Eq. (9.4) while deriving the asset pricing implications of portfolio choices within a representative agent framework.

9.2.2 Testable restrictions

Let us define cum-dividend wealth as $x_t \equiv \sum_{i=1}^{m}(P_{it} + D_{it})\theta_{it}$. In Appendix 1, we show that x_t evolves as follows

$$x_{t+1} = (x_t - c_t)\,\omega_t^{\top}(\mathbf{1}_m + \mathbf{r}_{t+1}) \equiv (x_t - c_t)(1 + r_{M,t+1}), \tag{9.5}$$

where ω is the vector of proportions of wealth invested in the m assets, $\mathbf{r}_{t+1}$ is the vector of asset returns, with any component i being equal to, $r_{it+1} \equiv \frac{P_{it+1}+D_{it+1}-P_{it}}{P_{it}}$, and $r_{M,t}$ is the return on the market portfolio, defined as

$$r_{M,t+1} = \sum_{i=1}^{m} r_{it+1}\omega_{it}, \quad \omega_{it} \equiv \frac{P_{it}\theta_{it+1}}{\sum_i P_{it}\theta_{it+1}},$$

where P_{it} and D_{it} are the price and the dividend of asset i at time t: the usual convention is that the portfolio θ_{it+1} is chosen at time t. Note that the portfolio decision also includes a

riskless asset, which can be interpreted by restricting the dividend on any of the previous assets to zero, as usual.

Consider an economy in which the underlying state is some Markov process y. We consider stationary consumption and investment plans. Accordingly, utility is a function $V(x,y)$ when current wealth is x and the state is y. By Eq. (9.4),

$$V(x_t,y_t)=\frac{1}{1-\gamma}\max_{c,\omega}\left(c^{\rho}+\beta\left(\left(1-\gamma\right)E_t(V\left(x_{t+1},y_{t+1}\right))\right)^{\frac{\rho}{1-\gamma}}\right)^{\frac{1-\gamma}{\rho}}. \tag{9.6}$$

In Appendix 1, we show that the first order conditions for the representative agent lead to the following Euler equation,

$$E_t\left[m\left(x_t,y_t;x_{t+1},y_{t+1}\right)\left(1+r_i\left(y_{t+1}\right)\right)\right]=1,\quad i=1,\cdots,m, \tag{9.7}$$

where the stochastic discount factor m is

$$m\left(x_t,y_t;x_{t+1},y_{t+1}\right)=\beta^{\theta}\left(\frac{c\left(x_{t+1},y_{t+1}\right)}{c\left(x_t,y_t\right)}\right)^{-\frac{\theta}{\psi}}\left(1+r_M\left(y_{t+1}\right)\right)^{\theta-1},\qquad \theta\equiv\frac{1-\gamma}{1-\frac{1}{\psi}}. \tag{9.8}$$

This stochastic discount factor displays the interesting property that it is affected by the market portfolio return, r_M, at least as soon as $\gamma\neq\frac{1}{\psi}$. In particular, the stochastic discount factor is countercyclical not only due to consumption growth, but (and provided $\theta<1$), due to the market return channel, whereby it gives more weight to states of nature when r_M is low than to states when r_M is high. Moreover, the stochastic discount factor may potentially inherit the excess volatility of market returns in a quite natural fashion.

Note, then, the interesting fixed point problem in this model: market returns affect the stochastic discount factor, which affects market returns in turn! Due to this fixed-point, asset prices predicted by these models are not known in closed-form, except for isolated exceptions. Furthermore, the potential of this model to explain the empirical puzzles needs to be further qualified, as discussed in the next section.

9.2.3 Risk premiums and interest rates

So the Euler equation is,

$$E_t\left(\beta^{\theta}\left(\frac{c_{t+1}}{c_t}\right)^{-\frac{\theta}{\psi}}\left(1+r_{M,t+1}\right)^{\theta-1}\left(1+r_{i,t+1}\right)\right)=1. \tag{9.9}$$

Eq. (9.9) obviously holds for the market portfolio and the risk-free asset. Therefore, by taking logs in Eq. (9.9) for $i=M$, and for the risk-free asset, $i=0$ say, yields the following conditions:

$$0=\ln E_t\left(e^{-\delta\theta-\frac{\theta}{\psi}\ln(c_{t+1}/c_t)+\theta R_{M,t+1}}\right),\quad R_{M,t}=\ln\left(1+r_{M,t}\right), \tag{9.10}$$

where $\delta\equiv-\ln\beta$, and

$$-R_{f,t}=-\ln\left(1+r_{0,t}\right)=\ln E_t\left(e^{-\delta\theta-\frac{\theta}{\psi}\ln(c_{t+1}/c_t)+(\theta-1)R_{M,t+1}}\right). \tag{9.11}$$

Next, suppose that consumption growth, $\ln\left(\frac{c_{t+1}}{c_t}\right)$, and the market portfolio return, $R_{M,t+1}$, are jointly normally distributed. In Appendix 1, we show that the expected excess return on the market portfolio is given by

$$E_t(R_{M,t+1}) - R_{f,t} + \frac{1}{2}\sigma^2_{R_M} = \frac{\theta}{\psi}\sigma_{R_M,c} + (1-\theta)\,\sigma^2_{R_M}, \tag{9.12}$$

where $\sigma^2_{R_M} = var(R_M)$ and $\sigma_{R_M,c} = cov(R_{M,t+1}, \ln(c_{t+1}/c_t))$, and the term $\frac{1}{2}\sigma^2_{R_M}$ in the left hand side is a Jensen's inequality term. Note, Eq. (9.12) is a mixture of the Consumption CAPM (for the part $\frac{\theta}{\psi}\sigma_{R_M,c}$) and the CAPM (for the part $(1-\theta)\,\sigma^2_{R_M}$).

The risk-free rate is

$$R_{f,t} = \delta + \frac{1}{\psi}E_t\left(\ln\frac{c_{t+1}}{c_t}\right) - \frac{1}{2}(1-\theta)\,\sigma^2_{R_M} - \frac{1}{2}\frac{\theta}{\psi^2}\sigma^2_c, \tag{9.13}$$

where $\sigma^2_c = var(\ln(c_{t+1}/c_t))$.

Eqs. (9.12) and (9.13) can be elaborated further. In equilibrium, the asset price and, hence, the asset return, is certainly related to consumption volatility. Precisely, assume that

$$\sigma^2_{R_M} = \sigma^2_c + \sigma^2_*, \quad \sigma_{R_M,c} = \sigma^2_c, \tag{9.14}$$

where σ^2_* is a positive constant that may arise when the asset return is driven by some additional state variable.[1] Under the assumption that the asset return volatility is as in Eq. (9.14), the equity premium in Eq. (9.12) is:

$$E_t(R_{M,t+1}) - R_{f,t} + \frac{1}{2}\sigma^2_{R_M} = \gamma\sigma^2_c + (1-\theta)\,\sigma^2_* = \gamma\sigma^2_c + \frac{\gamma - \frac{1}{\psi}}{1-\frac{1}{\psi}}\sigma^2_*. \tag{9.15}$$

Disentangling risk-aversion from intertemporal substitution does not necessarily lead to a resolution of the equity premium puzzle. To raise the equity premium, we need that $\sigma^2_* > 0$, meaning that additional state variables are required to drive variation in asset returns. At the same time, the volatility of these state variables has the power to affect asset returns only when risk-aversion is distinct from the reciprocal of the EIS. As an example, suppose that σ^2_* does not depend on γ and ψ, and that $\psi > 1$. Then, the equity premium increases with σ^2_* whenever $\gamma > \psi^{-1}$. In other words, these state variables have the potential to inflate the equity premium, but only once they enter the stochastic discount factor through the market return. Section 9.2.5 shall illustrate this feature of the model through a mechanism relying on long-run risks.

Next, we derive the risk-free rate. Assume that $E(\ln(c_{t+1}/c_t)) = g_0 - \frac{1}{2}\sigma^2_c$, where g_0 is the expected consumption growth, a constant. Furthermore, use the assumptions in Eq. (9.14) to obtain that the risk-free rate in Eq. (9.13) is

$$R_f = \delta + \frac{1}{\psi}g_0 - \frac{1}{2}\gamma\left(1+\frac{1}{\psi}\right)\sigma^2_c - \frac{1}{2}\frac{\gamma - \frac{1}{\psi}}{1-\frac{1}{\psi}}\sigma^2_*. \tag{9.16}$$

As we can see, we may increase the level of relative risk-aversion, γ, without substantially affecting the level of the risk-free rate, R_f. This is because the effects of γ on R_f have second-order importance (they multiply variances, which are orders of magnitude less than the expected consumption growth, g_0).

[1] This is for example the case of the Bansal and Yaron (2004) model described below.

9.2.4 Campbell-Shiller approximation

Consider the definition of the return on the market portfolio,

$$R_{M,t+1} = \ln\left(\frac{P_{t+1}+C_{t+1}}{P_t}\right) = \ln\left(\frac{e^{z_{t+1}}+1}{e^{z_t}}\right) + g_{t+1} \equiv f\left(z_{t+1}, z_t\right) + g_{t+1},$$

where P_t is the value of the market portfolio, $g_{t+1} = \ln\frac{C_{t+1}}{C_t}$ is the aggregate dividend growth, and $z_t = \ln\frac{P_t}{C_t}$ is the log of the aggregate price-dividend ratio. A first-order linear approximation of $f\left(z_{t+1}, z_t\right)$ around the "average" level of z leaves,

$$R_{M,t+1} \approx \kappa_0 + \kappa_1 z_{t+1} - z_t + g_{t+1}, \tag{9.17}$$

where $\kappa_0 = \ln\left(\frac{e^{\bar{z}}+1}{e^{\bar{z}}}\right) + \frac{\bar{z}}{e^{\bar{z}}+1}$, $\kappa_1 = \frac{e^{\bar{z}}}{e^{\bar{z}}+1}$ and $\bar{z}$ is the average level of the log price-dividend ratio, such that $\kappa_1 \approx 0.997$, using US data. The approximation in Eq. (9.17) appears for the first time in Campbell and Shiller (1988).

This approximation allows us to isolate the forces that are potentially capable to address the equity premium puzzle. We know that agents must require compensation for risk that goes well beyond that regarding consumption growth. Let us characterize, then, the innovations to the stochastic discount factor predicted by this model. By Eq. (9.8),

$$\ln m_{t,t+1} - E_t\left(\ln m_{t,t+1}\right) = -\gamma\left(g_{t+1} - E_t\left(g_{t+1}\right)\right) - (1-\theta)\kappa_1\left(z_{t+1} - E_t\left(z_{t+1}\right)\right), \tag{9.18}$$

where $g_{t+1} \equiv \ln\frac{C_{t+1}}{C_t}$ and $m_{t,t+1} \equiv m\left(x_t, y_t; x_{t+1}, y_{t+1}\right)$. That is, there can be additional sources of risk the agents require to be compensated for, provided that (i) the price-dividend ratio is random and (ii) agents have a *preference for early resolution of uncertainty* ($\theta < 1$). We now illustrate how this mechanism operates in the context of risks that may be very persistent.

9.2.5 Risks for the long-run

Bansal and Yaron (2004) consider a model in which *persistence* in the expected consumption growth has the potential to explain the equity premium puzzle. To illustrate the main points of this explanation, assume that consumption growth is solution to

$$g_{t+1} = \ln\frac{C_{t+1}}{C_t} = g_0 - \frac{1}{2}\sigma_c^2 + x_t + \epsilon_{t+1}, \quad \epsilon_{t+1} \sim N\left(0, \sigma_c^2\right), \tag{9.19}$$

where x_t is a "small" persistent component in consumption growth, solution to[2]

$$x_{t+1} = \rho x_t + \eta_{t+1}, \quad \eta_{t+1} \sim N\left(0, \sigma_x^2\right). \tag{9.20}$$

To find an approximate solution to the log of the price-dividend ratio, replace the Campbell-Shiller approximation in Eq. (9.17) into the Euler equation (9.10) for the market portfolio,

$$0 = \ln E_t\left(e^{-\delta\theta - \frac{\theta}{\psi}\ln(C_{t+1}/C_t) + \theta(\kappa_0 + \kappa_1 z_{t+1} - z_t + g_{t+1})}\right). \tag{9.21}$$

[2]In this version of the model, the equity premium and volatility are constant. Bansal and Yaron assume that η_t is heteroskedastic to make the model consistent with time-varying statistics.

Conjecture that the log of the price-dividend ratio takes the simple form, $z_t = a_0 + a_1 x_t$, where a_0 and a_1 are two coefficients to be determined. Substituting this guess into Eq. (9.21), and identifying terms, leaves:

$$z_t = a_0 + a_1 x_t, \quad a_1 \equiv \frac{1 - \frac{1}{\psi}}{1 - \kappa_1 \rho}, \tag{9.22}$$

where the constant a_0 is characterized in Appendix 1.

Note that when growth is very persistent, $1 - \kappa_1 \rho \approx 1$, even small changes in the expected dividend growth, x_t, lead to large swings in the price-dividend ratio. A model solved along these lines, where persistent processes would lead to highly volatile prices, was already available in the literature, at least since the discussion of Campbell, Lo and MacKinlay (1997, Chapter 7, p. 265) of a model with persistent expected returns:

> "If this standard deviation is small [i.e. the variability of expected stock returns], it is tempting to conclude that changing expected returns have little influence on stock prices. [...] This conclusion is too hasty: [...] if expected returns vary in a persistent fashion, [prices] can be very variable even when the [expected returns are] not."

The model discussed by Campbell, Lo and MacKinlay is one where expected returns are directly modeled as possibly persistent processes. Instead, the model of this section is one where expected growth is possibly persistent. We shall soon see the implications within such a broader perspective. Note, however, a crucial point. High volatility of the price-dividend ratio does not necessarily lead to a resolution of the equity premium puzzle. For example, in the context of non-expected utility, Eq. (9.15) suggests that in equilibrium, relative risk-aversion and intertemporal elasticity of substitution should play together in the right direction, and the variance σ_* is equally important. For example, if $\gamma = \psi^{-1}$, the price-dividend ratio would not even enter the Euler equation, as previously explained. Therefore, we need to check how this high volatility of the price-dividend ratio translates into a high equity premium.

We use the expression of $R_{M,t+1}$ in Eq. (9.17) and determine σ_*^2 in (9.14), leading to the model prediction regarding the equity premium and the risk-free rate through Eqs. (9.15) and (9.16). We have

$$\begin{aligned} \sigma_{R_M}^2 &= \kappa_1^2 var_t\left(z_{t+1}\right) + \sigma_c^2 \\ &= \underbrace{\kappa_1^2 \left(\frac{1 - \frac{1}{\psi}}{1 - \kappa_1 \rho}\right)^2 \sigma_x^2}_{\equiv \sigma_*^2} + \ \sigma_c^2. \end{aligned}$$

That is, and as anticipated, non-expected utility can lead to a resolution of the equity premium puzzle when asset returns are driven by sources of variation in addition to consumption growth. The previous expression for σ_*^2 shows precisely how long run risks accomplish this: this added volatility stems from large fluctuations in the price-dividend ratio, arising due to the high persistence of the "small" component x_t in Eq. (9.20). An alternative albeit entirely consistent way to explain these facts relies on the expression for the innovations to the stochastic discount factor in this model: by Eq. (9.18),

$$\ln m_{t,t+1} - E_t\left(\ln m_{t,t+1}\right) = -\gamma\left(g_{t+1} - E_t\left(g_{t+1}\right)\right) - \left(1 - \theta\right)\kappa_1 a_1 \left(x_{t+1} - E_t\left(x_{t+1}\right)\right).$$

That is, long-run risks are priced "through a_1," the sensitivity of the price-dividend ratio to changes in x_t (see Eq. (9.22)): even if the innovations to x_t are small, a_1 is large, as explained, and leads to a large compensation for this risk, provided $\theta < 1$.

[Survey briefly developments in the LRR literature]

9.3 Heterogeneous agents and "catching up with the Joneses"

The attractive feature of the Campbell and Cochrane (1999) model of external habit formation, reviewed in Chapter 8, is to have the potential to generate the right properties of asset prices and volatilities, through the channel of a countercyclical price of risk. It does rely on a high risk-averse economy, though. Chan and Kogan (2002) show that a countercyclical price of risk might arise, without assuming the existence of a representative agent with a high risk-aversion. They consider an economy where heterogeneous agents have preferences displaying "catching up with the Joneses" features introduced by Abel (1990, 1999).

In this economy, there is a continuum of agents, indexed by a parameter $\gamma \in [1, \infty)$ appearing in their instantaneous utility,

$$u_\gamma(c, x) = \frac{\left(\frac{c}{x}\right)^{1-\gamma}}{1-\gamma},$$

where c is consumption, and x is the "standard living of others," to be defined below.

The total endowment in the economy, D, follows a geometric Brownian motion,

$$\frac{dD_\tau}{D_\tau} = g_0 d\tau + \sigma_0 dW_\tau. \tag{9.23}$$

By assumption, the standard of living of others, x_τ, is a weighted geometric average of the past realizations of the aggregate endowment D, viz[3]

$$\ln x_\tau = \ln x_0 e^{-\theta\tau} + \theta \int_0^\tau e^{-\theta(\tau-s)} \ln D_s ds, \quad \text{with } \theta > 0.$$

Therefore, x_τ satisfies

$$dx_\tau = \theta s_\tau x_\tau d\tau, \quad \text{where } s_\tau \equiv \ln\left(\frac{D_\tau}{x_\tau}\right). \tag{9.24}$$

By Eqs. (9.23) and (9.24), s_τ is solution to,

$$ds_\tau = \left(g_0 - \frac{1}{2}\sigma_0^2 - \theta s_\tau\right) d\tau + \sigma_0 dW_\tau.$$

This model can be interpreted as one displaying "catching up with the Joneses" features because the utility of each agent is affected by a benchmark, x_τ, which is a weighted average of the past aggregate endowment D_τ, the latter obviously being equal to aggregate consumption (see the constraint in [P1] below). This model is important. We already know (see Chapter 8) that a realistically calibrated economy with habit formation and a representative agent relies on a high risk aversion. Moreover, an economy with "catching up with the Joneses" and a

[3]It is a distributed lags model, in that the weights on $\ln x_0$ and $\ln D_s$ sum up to one. Alternatively, and assuming time starts infinitely far in the past, it is an infinite distributed lags model, in that $\ln x_\tau = \int_{-\infty}^\tau \omega_{s,\tau} \ln D_s ds$, where the weights $\omega_{s,\tau} \equiv \theta e^{-\theta(\tau-s)}$ satisfy $\int_{-\infty}^\tau \omega_{s,\tau} ds = 1$.

representative agent would also rely on a high risk-aversion, as argued below. Chan and Kogan (2002) show that their model, while capturing the spirit of habit formation through "catching up with the Joneses," does not need to rely on a high risk-aversion, once we populate the economy with heterogenous agents, thereby allowing habit formation and "catching up with the Joneses" to perform a role in the explanation of the equity premium puzzle.

In this economy with complete markets, we can determine the asset price, and solve the model by relying on the centralization of competitive equilibrium through Pareto weightings, along lines similar to those in Theorem 2.7 of Chapter 2. As explained in Appendix 2, the equilibrium price process is the same as that in an economy with a representative agent with instantaneous utility,

$$u\left(D,x\right)\equiv\max_{c_{\gamma}}\int_{1}^{\infty}u_{\gamma}\left(c_{\gamma},x\right)f_{\gamma}d\gamma\quad\text{s.t.}\int_{1}^{\infty}c_{\gamma}d\gamma=D, \tag{P1}$$

where ${f_\gamma}^{-1}$ is the marginal utility of income of the agent γ. Appendix 2 provides further details on the derivation of the value function of the program [P1], which is:

$$U\left(s\right)\equiv\int_{1}^{\infty}\frac{1}{1-\gamma}f_{\gamma}^{\frac{1}{\gamma}}\psi\left(s\right)^{\frac{\gamma-1}{\gamma}}d\gamma, \tag{9.25}$$

where ψ is a Lagrange multiplier, a function of the state s, satisfying

$$e^{s}=\int_{1}^{\infty}f_{\gamma}^{\frac{1}{\gamma}}\psi\left(s\right)^{-\frac{1}{\gamma}}d\gamma. \tag{9.26}$$

Finally, the Appendix shows that the unit risk-premium predicted by this model is

$$\lambda(s)=\sigma_{0}\frac{e^{s}}{\int_{1}^{\infty}\frac{1}{\gamma}f_{\gamma}^{\frac{1}{\gamma}}\psi\left(s\right)^{-\frac{1}{\gamma}}d\gamma}. \tag{9.27}$$

To summarize, Eq. (9.26) determines the Lagrange multiplier, $\psi\left(s\right)$, which then feeds $\lambda\left(s\right)$ through Eq. (9.27). Empirically, the Pareto weighting function, f_γ, can be parametrized by a function, which can be calibrated to match selected characteristics of the asset returns and volatility. Note, finally, that this economy collapses to an otherwise identical homogeneous economy, once the social weighting function $f_\gamma=\delta(\gamma-\bar{\gamma})$, the Dirac's mass at $\bar{\gamma}$. In this case, $\lambda\left(s\right)=\sigma_0\bar{\gamma}$, a constant. As anticipated, an economy with a single agent with "catching up with the Joneses" is unlikely to resolve the equity premium puzzle or address other issues such as predictability, because $\sigma_0\bar{\gamma}$ is both small and constant.[4]

A technical, albeit crucial assumption of this model is that the standard of living of others, x_τ, is a process with *bounded* variation (see Eq. (9.24)). This assumption implies that x_τ is not a risk for which the agents require a compensation. In this model, then, it is the agents' heterogeneity that drives variation in the risk-premium $\lambda\left(s\right)$ (see Eq. (9.27)). By calibrating the model to US data, Chan and Kogan find that $\lambda\left(s\right)$ is decreasing and convex in s.[5] The mechanism of the model is an endogenous redistribution of wealth. Note that the less risk-averse agents obviously invest a higher proportion of their wealth into risky assets, compared to the more risk-averse. In the poor states of the world, then, when stock prices decrease, the wealth of the less risk-averse agents lowers more than that of the more risk-averse. The result

[4] See Campbell (2003) for a similar critique.

[5] Their numerical results also revealed that in their model, the *log* of the price-dividend ratio is increasing and *concave* in s. Finally, their lemma 5 (p. 1281) establishes that in a homogeneous economy, the price-dividend ratio is increasing and *convex* in s.

is a reduction in the fraction of wealth held by the less risk-averse individuals in the whole economy. Thus, in bad times, the contribution of these less risk-averse individuals to aggregate risk-aversion decreases and, hence, aggregate risk-aversion increases and so does $\lambda(s)$ in Eq. (9.27).

[Discuss the criticism of Xiouros and Zapatero (2010)]

9.4 Idiosyncratic risk

Aggregate risk is too small to justify the size of the equity premium through a low level of risk-aversion. Do individuals bear some idiosyncratic risk, one that cannot be diversified away by trading in capital markets? And, then, how can this risk affect asset evaluation? Shouldn't this risk be neutral to asset pricing? The answer to these questions is indeed subtle, and relies upon whether idiosyncratic risk affects agents' portfolio choices and, then, the stochastic discounting factor. This section presents and discusses models that activate these channels.

9.4.1 A static model

Mankiw (1986) is the first to point out the asset pricing implications of idiosyncratic risks. In his model, aggregate shocks to consumption do not affect individuals in the same way, ex-post. Ex-ante, individuals know that the business cycle may adversely change—an *aggregate* shock—although they also anticipate that the very same same shock might be particularly severe to only a portion of them—an *idiosyncratic* shock. To illustrate, everyone faces a positive probability of experiencing a job loss during a recession, although then, only a part of the population will actually suffer from a job loss. Alternatively, one thing is to say that a recession will lead to a salary reduction to everyone, and another thing is to say that a number of individuals will not even be affected by the recession but then the others will bear its entire burden.

Idiosyncratic risk might significantly affect agents' portfolio choice and, therefore, rational asset evaluation. Naturally, in the presence of contingent claims able to insure against these shocks, idiosyncratic risk would not matter. But the point is that in reality, these contingent claims do not exist yet, due perhaps to moral hazard or adverse selection reasons. This source of market incompleteness might then potentially explain the aggregate stock market behavior in a way that the model with a standard representative agent cannot.

Mankiw considers the pricing of a risky asset in a two-period model, with the first period budget constraint given by $p\theta + m = w$, and the second period consumption equal to:

$$c = \tilde{w} + R\theta + (w - p\theta)(1 + r) \equiv \tilde{w} + \tilde{X}\theta,$$

where m is the amount to invest in a money market account, r is the safe interest rate on the money market account, normalized to zero, w is the initial endowment, also normalized to zero, p is the price of the risky asset, R is the payoff promised by the risky asset and, finally, $\tilde{X} \equiv R - p$ is the asset "net payoff." That is, we may either endogenize the price p, given the payoff R, or, then, just the net payoff, $\tilde{X}$, as described next. The asset is in zero net supply, and because agents are ex-ante identical, we have that in equilibrium, $\theta = 0$, such that $\tilde{w}$ equals per capita consumption, c.

There are two states of nature for the aggregate economy, which are equally likely. In the good state, the net asset payoff is $\tilde{X} = 1 + \pi$, and per capita consumption is $\tilde{w} = \mu$. In the bad state, the net asset payoff is, instead, $\tilde{X} = -1$, and per capita consumption is also adversely

affected in that, $\tilde{w} = (1-\phi)\,\mu$. The payoff in the good state, π, equals $2E(\tilde{X})$. Therefore, π measures a risk premium; of course it has to be determined in equilibrium.

How does this macroeconomic shock affect asset pricing in the presence of agents who could be heterogeneous ex-post? The crucial feature of the model is the assumption that only a fraction λ of agents will absorbe the macroeconomic risk, in that, literally, a fraction $1-\lambda$ of individuals will not be hit by the aggregate shock, and each of them will still consume μ, for a total of $(1-\lambda)\,\mu$. The residual portion of the population will consume the "residual," $(1-\phi)\,\mu-(1-\lambda)\,\mu=(\lambda-\phi)\,\mu$, such that each agent hit by the shock will consume $\left(1-\frac{\phi}{\lambda}\right)\mu$. The ratio, $\frac{\phi}{\lambda}$, is the per capita fall in consumption for any individual hit by the shock in the bad state of nature. If $\lambda=1$, the aggregate shock hits everyone. The highest concentration of the shock arises when $\lambda=\phi$, i.e. when the fall in consumption is borne by the lowest possible fraction of the population. Table 9.1 summarizes payoffs, per capita consumption and individual consumption in this economy.

	Net Asset Payoff	Per capita consumption	Individual consumption	
Bad state	-1	$(1-\phi)\,\mu$	$\left(1-\frac{\phi}{\lambda}\right)\mu$	(consumed by λ)
			μ	(consumed by $1-\lambda$)
Good state	$1+\pi$	μ	μ	

TABLE 9.1. Aggregate fluctuations and idiosyncratic risk

To summarize, only a selected portion of individuals will bear and "share" a macroeconomic shock, such that the burden of this shock will be particularly heavy should it materialize, especially when λ is small. Since the risky asset performs poorly in bad times, its evaluation should reflect not only business cycle risk, but also the fact that the business cycle might be particularly severe to those who will be hurt by it.

To formalize this reasoning, consider the first order conditions for any expected utility maximizer:

$$\begin{aligned}0 &= E[\tilde{X}u'(\tilde{w}+\tilde{X}\theta)]\\ &= E[\tilde{X}u'(\tilde{w})]\\ &= \tfrac{1}{2}\cdot(-1)\left[u'\left(\left(1-\tfrac{\phi}{\lambda}\right)\mu\right)\lambda+u'(\mu)(1-\lambda)\right]+\tfrac{1}{2}\cdot(1+\pi)\,u'(\mu),\end{aligned}$$

where the second equality follows by the equilibrium condition, $\theta=0$, and u is an utility function satisfying standard regularity conditions. The premium, π, equals:

$$\pi=\lambda\frac{u'\left(\mu\left(1-\frac{\phi}{\lambda}\right)\right)-u'(\mu)}{u'(\mu)}.$$

Mankiw shows that for utility functions leading to prudent behavior, $u'''>0$, the premium π is decreasing in λ: an increase in the concentration of aggregate shocks leads to higher premiums. Moreover, it is easy to see that π can be made arbitrarily large, for λ arbitrarily close to ϕ, once the utility function satisfies the Inada's condition, $\lim_{c\to 0}u'(c)=\infty$, as we have that $\lim_{\lambda\to\phi}\pi=\infty$. For example, in the log-utility case, we have that $\pi=\frac{\lambda\phi}{\lambda-\phi}$.

9.4.2 Self-insurance and persistence of idiosyncratic shocks

The critical assumption underlying Mankiw's model is that once agents are hit by an idiosyncratic shock, the game is over. What happens once we allow the agents to act in a multiperiod horizon? Intuitively, in a dynamic context, agents might implement self-insurance plans, by accumulating financial assets after good shocks and selling or short-selling after bad shocks have occurred. Telmer (1993) and Lucas (1994) show that if idiosyncratic shocks are not persistent, self-insurance is quite effective and asset prices behave substantially the same as they would do in a world without idiosyncratic risk. Therefore, to have asset prices significantly deviate from those arising within a complete market setting, one has to either (i) reduce the extent of risk-sharing, by assuming frictions such as transaction costs, short-selling constraints or in general severe forms of market incompleteness, or (ii) make idiosyncratic shocks persistent. With (i), we just merely eliminate the possibility that agents may implement self-insurance plans through capital market transactions. With (ii), we make idiosyncratic shocks so severe that no capital market transaction might allow agents to insure themselves and achieve portfolio solutions close to the complete market solution; intuitively, once any individual is hit by an idiosyncratic shock, he may short-sell financial assets in the short-run, although then, he cannot persistently do so, given his wealth constraints.

Heaton and Lucas (1996) calibrate a model with idiosyncratic shocks using PSID (Panel Study of Income Dynamics) and NIPA (National Income and Product Accounts) data. They find that idiosyncratic shocks are not quite persistent, and that a large amount of transaction costs is needed to generate sizeable levels of the equity premium. Naturally, idiosyncratic shocks are not always as in the PSID dataset analyzed by Heaton and Lucas long time ago. Models with idiosyncratic risk would actually help think about market behavior in periods such as the "Great Recession"—the recession occurred around 2009, when the persistence of idiosyncratic shocks would arguably be larger than over the "Great Moderation"—the period of low volatility of macroeconomic aggregates, starting after the Monetary experiment in 1982 (e.g., Bernanke, 2004) and presumably ending in 2007.

9.4.3 A model with countercyclical income inequality

Constantinides and Duffie (1996) do actually take the issue of persistence in idiosyncratic risk to the extreme, and consider a model without any transaction costs, but with permanent idiosyncratic risk. They show that in fact, given an asset price process, it is always possible to find a cross-section of idiosyncratic risk processes compatible with the asset price given in advance. We now present this elegant model, which has a quite substantial theoretical importance per se, because of its feature to make so transparent how some state variables affecting consumer choices can be reverse-engineered from the observation of an asset price process.

Central to Constantinides and Duffie analysis is the assumption that each individual i has a consumption equal to $c_{i,t}$ at time t, given by:

$$c_{i,t} = \ell_{i,t} c_t, \quad \ell_{i,t} = \exp\left(\sum_{s=1}^{t}\left(\zeta_{i,s} y_s - \frac{1}{2} y_s^2\right)\right),$$

where $\zeta_{i,t}$ are independent and standard normally distributed, and y_t is a sequence of random variables, interpreted as standard deviations of the cross-sectional distribution of the individual

consumption growth shares, $\ln \frac{\ell_{i,t}}{\ell_{i,t-1}}$. Indeed,

$$\ln\left(\frac{c_{i,t+1}/c_{t+1}}{c_{i,t}/c_t}\right)\Bigg|_{\mathbb{F}_t\cup\{y_{t+1}\}} \sim N\left(-\frac{1}{2}y_{t+1}^2, y_{t+1}^2\right), \tag{9.28}$$

where $\mathbb{F}_t$ is the information set as of time t. Denoting with $\mu(i)$ is the measure of agent i, we have that by Eq. (9.28) and the Law of Large Numbers, the cross-sectional variance of $\ln \frac{\ell_{i,t}}{\ell_{i,t-1}}$ is, $\int (\ln \frac{\ell_{i,t}}{\ell_{i,t-1}} + \frac{1}{2}y_t^2)^2 d\mu(i) = y_t^2 \int \zeta_{i,t}^2 d\mu(i) = y_t^2$.

The meaning of the consumption share $\ell_{i,t}$ is that of an idiosyncratic shock every agent i receives on his consumption share at t. From the perspective of each agent, this shock is uninsurable, in that it is unrelated to the asset returns. Moreover, by construction, the consumption share has a unit root, as $\ln \ell_{i,t} - \ln \ell_{i,t-1} = \zeta_{i,t} y_t - \frac{1}{2}y_t^2$: a change in y_t and/or a shock in $\zeta_{i,t}$ have a permanent effect on the future path of $\ell_{i,t}$.

All agents have a CRRA utility function. We want to make sure this setup is consistent with any given equilibrium asset price process, by requiring two conditions: (i) $\int c_{i,t} d\mu(i) = c_t$, i.e. $\int \ell_{i,t} d\mu(i) = 1$, a condition satisfied by the law of large numbers; (ii) the cross-sectional variances y_t^2 are reverse-engineered so as to be consistent with any stochastic discount factor and, hence, any asset price process given in advance. To achieve (ii), note that by the law of iterated expectations, for any agent i, the value of an asset delivering a payoff equal to $\tilde{X}$ at time $t+1$ is:

$$\begin{aligned}
&E\left[e^{-\rho}\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} E\left(e^{-\gamma\left(\zeta_{i,t+1}y_{t+1}-\frac{1}{2}y_{t+1}^2\right)}\middle|\, \mathbb{F}_t\cup\{y_{t+1}\}\right)\cdot\tilde{X}\middle|\,\mathbb{F}_t\right]\\
&= E\left[e^{-\rho}\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} e^{\frac{1}{2}\gamma(\gamma+1)y_{t+1}^2}\cdot\tilde{X}\middle|\,\mathbb{F}_t\right],
\end{aligned}$$

where ρ is the discount rate and γ is the CRRA coefficient. It is independent of any agent i, such that the stochastic discount factor is:

$$\frac{\xi_{t+1}}{\xi_t} \equiv m_{t+1} \equiv e^{-\rho}\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} e^{\frac{1}{2}\gamma(\gamma+1)y_{t+1}^2}.$$

That is, given an aggregate consumption process, and an arbitrage free asset price process, there exists a cross-section of idiosyncratic risk processes that supports the given price process. As a trivial example, consider the standard Lucas stochastic discount factor, which obtains when $y_t \equiv 0$.

Which properties of the stochastic discounting factor are we looking for? Naturally, we wish to make sure m_t is as countercyclical as ever, which might be the case should the dispersion of the cross-sectional distribution of the log-consumption growth, y_t^2, be countercyclical. However, Lettau (2002) shows that empirically, such a dispersion seems to be not enough, even when multiplied by $\frac{1}{2}\gamma(\gamma+1)$, unless of course, we are willing to assume, again, a high level of risk-aversion. Note that Lettau analyzes a situation that favourably biases his final outcome towards not rejecting the null that idiosyncratic risk matters, as he assumes agents cannot insure themselves at all: once they are hit by an idiosyncratic shock, they just have to consume their income. Constantinides, Donaldson and Mehra (2002) consider an OLG to mitigate the issue of persistence in the idiosyncratic risk process.

Discuss the recent literature.

[In progress]

9.5 Incomplete markets with homogeneous and heterogenous agents

In the previous section, idiosyncratic risk has direct linkages to aggregate consumption. For example, in the seminal paper of Mankiw (1986), idiosyncratic risk is correlated with aggregate shocks, and in Constantinides and Duffie (1996), it is related to time-variation of the cross-sectional distribution of the consumption growth shares.

This section explores the implications of models where market incompleteness does not link to aggregate risk. Instead, market incompleteness is either *given* in the first place (independently of the state of the economy) or is induced by idioyncratic risks that have a more general form (but, still, cannot be hedged while trading in financial markets). Section 9.5.1 reviews a model with agents who are homogeneous ex ante, but heterogenous ex post, due to the realization of idiosyncratic shocks unrelated to aggregate risk. Section 9.5.2 reviews a model with agents who are heterogeneous ex ante, in that they may have unequal access to securities markets.

9.5.1 Idiosyncratic shocks unrelated to aggregate risk

Weil (1992) develops a simple model where idiosyncratic shocks are unrelated to aggregate risk. In this model, there is a continuum of agents who are ex ante identical, in that they have the same utility and the same endowments, a non-random ω in the first period, and a stochastic endowment $\tilde{\omega}$ in the second, independent of dividends, and with possibly different outcomes across them. The agents, then, face the following budgent constraints:

$$\begin{cases} c_1 + S\theta = (S+d)\,\theta_0 + \omega \\ \tilde{c}_2 = \tilde{d}\theta + \tilde{\omega} \end{cases}$$

where θ_0 denotes the endowment of the risky asset, c_1 and $\tilde{c}_2$ are first and second period consumption, and d and $\tilde{d}$ are the dividends in the first and second period. We assume that $\theta_0 = 1$. Because agents are ex ante identical, although ex post heterogenous, they have no incentives to trade. An autarkic equilibrium is therefore one where $\theta = 1$, $c_1 = d + \omega$ and $\tilde{c}_2 = \tilde{d} + \tilde{\omega}$. Markets are incomplete because the endowments $\tilde{\omega}$ cannot be spanned through the risky asset. The asset price satisfies

$$S = \frac{E(u'(\tilde{d}+\tilde{\omega})\tilde{d})}{u'\,(d+\omega)}.$$

Compare this price with that arising in a complete markets economy, where by Pareto optimality, the agents have the same marginal rate of substitution, state by state. In this complete markets economy with agents displaying identical preferences, risk-pooling is the equilibrium outcome, with each agent consuming the same, $\tilde{c}_2 = \tilde{d} + E(\tilde{\omega})$, such that the price in the complete market setting is,

$$S_{\mathrm{c}} = \frac{E(u'(\tilde{d}+E(\tilde{\omega}))\tilde{d})}{u'\,(d+\omega)}.$$

If u' is convex, and as soon as dividend and endowment risks are independent, $E(u'(\tilde{d} + E(\tilde{\omega}))\tilde{d}) \leq E(u'(\tilde{d}+\tilde{\omega})\tilde{d})$, thereby leading to (i) the risk-free rate in the incomplete market case lower than in the complete case, and (ii) $S_{\mathrm{c}} \leq S$.

The interpretation of the condition that u' be convex relates to prudence, similarly as for the condition of the previous section that ensures the equity premium in Mankiw's model is increasing in the concentration of aggregate shocks. In fact, Weil argues that the risk-free rate

and equity premium puzzles might be rationalized by agents facing incomplete markets while engaging in precautionary savings, and restricted by preferences exhibiting *decreasing absolute risk aversion (DARA)* and *decreasing absolute prudence (DAP)*—an utility function u exhibits decreasing absolute prudence if the coefficient of absolute prudence, $-\frac{u'''(x)}{u''(x)}$, is decreasing. The mechanism is quite illuminating. Incomplete markets lead agents to act more prudently than they would if they had to face complete markets and as a result, to increase their demand of both the riskless and the risky assets. Clearly, then, the riskless interest rate decreases although to ensure an increase in the equity premium, we need to ensure that the increase in the price of the risky asset is less than the riskless, which is ensured by DARA and DAP—our agents need then to be less prone to bear dividend risk once they have nontraded labor income.

While the model seems somehow restrictive due to its reliance on DARA and DAP, many standard utility functions satisfy DARA and DAP. Duffie (1992) develops a model with a finite number of agents without exploring its implications. Note that these models are static. It is not clear whether self-insurance would play a role in these models where idiosyncratic and aggregate risk are unrelated. The model in the next section has a completely different mechanism, relying on the assumption some agents would never be able to implement any self-insurance plans, living in a world with an extreme form of market incompleteness, with all the macroeconomic risk being borne by a handful portion of remaining agents.

9.5.2 A two-agents economy

The economy of this section relies on the presence of heterogenous agents. One agent has access to the market for the risk asset, while a second agent has not. The equity premium is the expected excess return the first agent requires to enter the risky asset market, and can be quite large, even in the presence of small risk-aversion, because the agent is being willing to take on the entire aggregate macroeconomic risk.

Basak and Cuoco (1998) consider the following model. An agent does not invest in the stock market, and has logarithmic instantaneous utility, $u_n(c) = \ln c$. From his perspective, markets are incomplete. A second agent, instead, can invest in the stock market, and has instantaneous utility equal to $u_p(c) = (c^{1-\gamma} - 1)/(1-\gamma)$. Both agents are infinitely lived. The competitive equilibrium of this economy cannot be Pareto efficient, and so aggregation results such as those underlying the economy in Section 9.2 cannot obtain. However, Basak and Cuoco show that aggregation still obtains in this economy, once we define social weights in a judicious way.

Let $\hat{c}_{i\tau}$ be the general equilibrium allocation of agent i, $i = p, n$. In equilibrium, $\hat{c}_p + \hat{c}_n = D$, where D is the instantaneous aggregate consumption, taken to be a geometric Brownian motion with parameters g_0 and σ_0,

$$\frac{dD_\tau}{D_\tau} = g_0 d\tau + \sigma_0 dW_\tau. \tag{9.29}$$

Define

$$s_\tau \equiv \frac{\hat{c}_{p\tau}}{D_\tau}, \tag{9.30}$$

which is the consumption share of the market participant.

The first order conditions pertaining to the two agents intertemporal consumption plans are:

$$u_p'(\hat{c}_{p\tau}) = w_p e^{\delta\tau}\xi_\tau, \quad \hat{c}_{n\tau}^{-1} = w_n e^{\delta\tau - \int_0^\tau r_s ds}, \tag{9.31}$$

where w_p, w_n are two constants, ξ is the pricing kernel process, solution to,

$$\frac{d\xi_\tau}{\xi_\tau} = -r_\tau dt - \lambda_\tau \cdot dW_\tau, \tag{9.32}$$

and, finally, r is the short-term rate and λ is the unit risk-premium, which equal

$$r(s) = \delta + \frac{\gamma g_0}{s + \gamma(1-s)} - \frac{1}{2}\frac{1}{s(s+\gamma(1-s))}\sigma_0^2\gamma(1+\gamma), \tag{9.33}$$

and

$$\lambda(s) = \gamma\sigma_0 s^{-1}. \tag{9.34}$$

The expressions for r and λ in Eqs. (9.33)-(9.34) are derived below. Appendix 2 provides a further derivation relying on the existence of a representative agent, as originally put forward by Basak and Cuoco (1998), and explained below.

In this economy, the marginal investor bears the entire macroeconomic risk. The risk premium he requires to invest in the aggregate stock market is large when his consumption share, s_τ, is small. With just a risk aversion of $\gamma = 2$, and a consumption volatility of 1%, this model can explain the equity premium, as the plot of Eq. (9.34) in the Figure 9.1 illustrates. For example, Mankiw and Zeldes (1991) estimate that the share of aggregate consumption held by stock-holders is approximately 30%, which in terms of this model, would translate to an equity premium of more than 6.5%.

Guvenen (2009) makes an interesting extension of the Basak and Cuoco model. He consider two agents in which only the "rich" invests in the stock market, and is such that $\text{EIS}_{\text{rich}} > \text{EIS}_{\text{poor}}$. He shows that for the rich, a low EIS is needed to match the equity premium. However, US data show that the rich have a high EIS, which can not do the equity premium. (Guvenen considers an extension of the model where we can disentangle EIS and CRRA for the rich.)

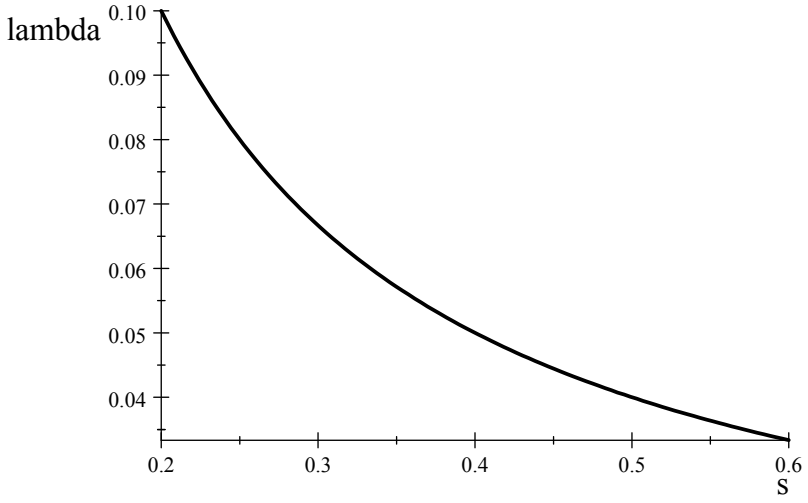


FIGURE 9.1. The equity premium in the Basak and Cuoco (1998) model, for $\gamma = 2$ and $\sigma_0 = 1\%$.

To derive Eqs. (9.33)-(9.34), note that the consumption of the agent not participating in the stock market satisfies, by Eq. (9.31):

$$\frac{d\hat{c}_{n\tau}}{\hat{c}_{n\tau}} = (r_\tau - \delta)\, d\tau. \tag{9.35}$$

Therefore, the consumption of the marginal investor, $\hat{c}_p = D - \hat{c}_n$, satisfies

$$\begin{aligned}\frac{d\hat{c}_{p\tau}}{\hat{c}_{p\tau}} &= \frac{dD_\tau}{D_\tau}\frac{D_\tau}{\hat{c}_{p\tau}} - \frac{d\hat{c}_{n\tau}}{\hat{c}_{n\tau}}\frac{\hat{c}_{n\tau}}{\hat{c}_{p\tau}} \\ &= \frac{1}{s_\tau}\left(g_0 - (r_\tau - \delta)(1 - s_\tau)\right) d\tau + \sigma_0 \frac{1}{s_\tau} dW_\tau,\end{aligned} \tag{9.36}$$

where the second equality follows by the definition of s in Eq. (9.30), and by Eqs. (9.29) and (9.35). Moreover, by the first order conditions of the market participant in Eq. (9.31), and the CRRA assumption for u_p,

$$\gamma d \ln \hat{c}_{p\tau} = -\delta d\tau - d \ln \xi_\tau = -\left(\delta - r_\tau - \frac{1}{2}\lambda_\tau^2\right) d\tau + \lambda_\tau dW_\tau. \tag{9.37}$$

Using the relation, $d \ln c = \frac{dc}{c} - \frac{1}{2}\left(\frac{dc}{c}\right)^2$, then identifying terms in Eq. (9.36) and Eq. (9.37), delivers the two expressions for r and λ in Eqs. (9.33)-(9.34).

How do these results technically relate to aggregation? Basak and Cuoco define the *instantaneous* utility of a representative agent, a social planner, as:

$$U(D, x) \equiv \max_{c_p + c_n = D} \left[u_p(c_p) + x \cdot u_n(c_n)\right], \tag{9.38}$$

where

$$x \equiv \frac{u_p'(\hat{c}_p)}{u_n'(\hat{c}_n)} = u_p'(\hat{c}_p)\hat{c}_n,$$

is a *stochastic social weight* and, once again, $\hat{c}_p$ and $\hat{c}_n$ are the private allocations, satisfying the first order conditions in Eqs. (9.31). By the definition of ξ, and Eqs. (9.31), x_τ is solution to,

$$dx_\tau = -x_\tau \lambda_\tau dW_\tau. \tag{9.39}$$

Then, the equilibrium in this economy is supported by a fictitious representative agent with utility $U(D, x)$. Intuitively, the social planner "allocations" satisfy, by construction,

$$\frac{u_p'(c_{p\tau}^*)}{u_n'(c_{n\tau}^*)} = \frac{u_p'(\hat{c}_{p\tau})}{u_n'(\hat{c}_{n\tau})} = x_\tau,$$

where the starred variables denote social planner's "allocations." In other words, Basak and Cuoco approach is to find a stochastic social weight process x_τ such that the first order conditions of the representative agent leads to the market allocations. The utility in Eq. (9.38) can then be used to compute the short-term rate and risk premium, and lead precisely to Eqs. (9.33)-(9.34), as shown in Appendix 2.

9.6 Disagreement and learning

This is the first of two sections dealing with economies in which agents have a limited knowledge of the fundamentals, which they attempt to improve by processing publicly available signals. In this section, we assume agents have different opinions: while they all learn from signals, their learning pattern might differ for a variety of reasons—for example, agents might suffer from different cognitive biases regarding the real content of the signals they have access to.

We analyze the impact of disagreement on asset prices in two models. In both, the mechanism leading to disagreement is the agents' overconfidence about their ability to process information they consider as "their own." In the first model, developed by Scheinkman and Xiong (2002), bubbles arise as a result of differences in opinion amongst risk-neutral agents in the presence of short-sale constraints. In the second, agents' disagreement is cast within a more standard general equilibrium with complete markets, risk-averse agents and no short-sale constraints; a special case of this model is the two-person model of Dumas, Kurshev and Uppal (2009), which we discuss in some detail.

To deal with the models in this section, we need a preliminary survey of tools, which allow us to think about how agents process information conveyed by multiple signals. These tools generalize the scope of those introduced in Chapter 8 (Section 8.5.4), and are in Section 9.6.1. In Section 9.6.2, we present details regarding the model of Scheinkman and Xiong, and in Section 9.6.3, we analyze how differences in opinions are dealt with in general equilibrium in the absence of frictions. Section 9.6.3 also deals with issues of aggregation in markets with heterogeneous beliefs, providing a general framework to address a number of issues, from the analysis of survival of irrational traders, to themes such as the equity premium puzzle or excess volatility in irrational markets.

Note that the vast majority of the explanations in this section have a "behavioral slant." We are about to assume that agents have psychological biases, in that they make systematic mistakes whilst assessing the probability distribution of the fundamentals, by emphasizing aspects of the markets such as correlations or informativeness of signals, which are less pronounced than in the real markets they operate in. In Section 9.7, we move to an alternative approach, relying on Knightian uncertainty: while agents still have a poor knowledge of the complex environment in which they take decisions, they behave rationally, by explicitly acknowledging their ignorance and acting while fearing the errors possibly arising from it.

9.6.1 Learning with multiple signals

We state filtering results on Bayesian learning that generalize those in Chapter 8 while taking into account the possibility that agents might update beliefs relying on multiple signals.

Consider the following result, a special case of Theorem 12.7 in Liptser and Shiryaev (2001; page 36). Suppose that some unobservable process θ_t is solution to

$$d\theta_t = (a_0 + a_1\theta_t)\, dt + b dZ_t, \tag{9.40}$$

where a_0 and a_1 are scalar constants, b is k-dimensional vector of constants, and Z_t is a k-dimensional standard Brownian motion. Assume that a ℓ-dimensional vector of signals $\mathcal{S}_t$ on θ_t are observed, solution to

$$d\mathcal{S}_t = A\theta_t dt + B dZ_t, \tag{9.41}$$

where A and B are a ℓ-dimensional vector and a $\ell \times k$ matrix of constants. Define

$$\hat{\theta}_t \equiv E\left(\theta_t | \mathbb{F}_t\right),$$

where $\mathbb{F}_t$ is the information set available at time t. Then, $\hat{\theta}_t$ is a diffusion process with time varying but deterministic volatility.

Assuming enough time has elapsed to have made the variance of $\hat{\theta}_t$ stationary, we have

$$d\hat{\theta}_t = (a_0 + a_1\hat{\theta}_t)dt + \left(bB^\top + \alpha_* A^\top\right)(BB^\top)^{-1}(d\mathcal{S}_t - A\hat{\theta}_t dt), \tag{9.42}$$

where

$$\alpha_* : 0 = 2a_1\alpha_* + bb^\top - \alpha_*^2 A^\top (BB^\top)^{-1}A. \tag{9.43}$$

Note that the Brownian motions driving the unobservable θ_t and the signals $\mathcal{S}_t$ are (potentially) the same, such that the interpretation of $bB^\top$ into the variance term of Eq. (9.42) is that of the instantaneous covariance between θ_t and $\mathcal{S}_t$,

$$cov\left(d\theta_t d\mathcal{S}_t\right) = bB^\top dt. \tag{9.44}$$

The next sections rely on these filtering results and evaluate how agents update their beliefs in light of new information in a number of models.

9.6.2 Overconfidence and bubbles

Scheinkman and Xiong (2002) consider a market in which the *cumulative* dividend process D_t^c of an asset satisfies

$$dD_t^c = D_t dt + \sigma_d dZ_t^d, \tag{9.45}$$

where Z_t^d is a Brownian motion, σ_d is a constant, and D_t is the expected instantaneous dividend. Note that in all the models dealt with so far, cumulative dividend is locally deterministic, i.e. $\sigma_d = 0$ in Eq. (9.45). So this model differs, and relies on the additional assumption that D_t is not observed—only D_t^c is—, although known to be a mean-reverting process,

$$dD_t = \lambda\left(\bar{D} - D_t\right)dt + \sigma_f dZ_t^f, \tag{9.46}$$

where Z_t^f is a standard Brownian motion, and λ and σ_f are two positive constants.

There are two sets of risk-neutral agents, A and B, who observe signals on D_t, which satisfy:

$$ds_t^i = D_t dt + \sigma_s dZ_t^i, \quad i \in \{A, B\}. \tag{9.47}$$

However, agent A thinks of s^A as being "his own signal" and believes its volatility is $\frac{\sigma_s}{\phi}$, for some $\phi \geq 1$, not σ_s, and agent B thinks the same regarding "his own signal" s^B. This overconfidence is the source of a behavioral bias similar to that considered in previous work by Kyle and Wang (1997) and Odean (1998): agents make systematic mistakes about the asset fundamentals while processing information, by assigning higher weight to a signal they arbitrarily perceive as their own.

These assumptions are those appearing in the working paper version of Scheinkman and Xiong (2002). In the published version, the assumption is that agents perceive signals to be correlated with the fundamentals even if they are not. These two alternative assumptions give rise to the same conclusions. We maintain the assumptions in this section because we deal with the latter assumption in Section 9.6.3, while discussing another model, thereby providing the reader with an additional exercise about learning in an overconfidence context.[6] We now apply the filtering results in the previous section and analyze how overconfidence leads to disagreement formally. We then turn to the determination of an equilibrium, and explain the mechanism leading to bubbles.

[6] Alternatively, we may think of agent A (resp. agent B) as having access to more precise information about signal A (signal B) than agent B (agent A). However, in this case, agents would not disagree once information is aggregated.

9.6.2.1 The inference process

We provide a step-by-step derivation of the inference process. In terms of the framework of Section 9.6.1, the expected instantaneous dividend D_t in Eq. (9.46) is the unobservable θ_t, and the signals available to each agent, $\mathcal{S}_t$, are the cumulative dividend D_t^c in Eq. (9.45) and the two signals s_t^i in Eq. (9.47), where the latter are subject to behavioral biases as explained.

First, consider an agent who observes the cumulative dividend D_t^c in Eq. (9.45), such that the Brownian motion in Section 9.6.1 is $Z_t = [Z_t^d \; Z_t^f]$ here, with the covariance term in Eq. (9.44) being then equal to zero, $cov\,(dD_t dD_t^c) = 0$. Then, by Eq. (9.42), the expected instantaneous dividend at t given all available information, $\hat{D}_t \equiv E\,(D_t|\,\mathbb{F}_t)$, satisfies:

$$d\hat{D}_t = \lambda(\bar{D} - \hat{D}_t)dt + \alpha_* \sigma_d^{-2}(dD_t^c - \hat{D}_t dt), \tag{9.48}$$

where α_* is the positive root of $0 = \sigma_d^{-2}\alpha_*^2 + 2\lambda\alpha_* - \sigma_f^2$. This is entirely consistent with the early models of learning reviewed in Chapter 8.

Next, assume that agents also observe the signals s_t^i in Eq. (9.47) but have no overconfidence. Then, below, we shall argue that all agents update priors through new information as follows,

$$d\hat{D}_t = \lambda(\bar{D} - \hat{D}_t)dt + \alpha_* \left(\frac{1}{\sigma_d} dW_t^d + \frac{1}{\sigma_s} dW_t^A + \frac{1}{\sigma_s} dW_t^B\right), \tag{9.49}$$

where W_t^d is a Brownian motion under the information set $\mathbb{F}_t$ available to the agents, and is defined as:

$$dW_t^d \equiv \frac{1}{\sigma_d}(dD_t^c - \hat{D}_t dt).$$

The two Brownian motions W_t^i are defined similarly, and will be dealt with later (see Eqs. (9.53) below). Note that Eq. (9.49) collapses to Eq. (9.48), once we assume that agents have no access to the two signals s_t^i in Eq. (9.47) or, equivalently, that these signals are uninformative, i.e. $\sigma_s \to \infty$.

Finally, consider the case in which agents have overconfidence. Agent A, to start with, thinks of the signal s^B as being generated by Eq. (9.47), but regarding s_t^A, he assumes that,

$$ds_t^A = D_t dt + \frac{\sigma_s}{\phi} dZ_t^A. \tag{9.50}$$

In terms of the filtering problem in Section 9.6.1, the information set of this agent includes realizations of $\mathcal{S}_\tau = [D_\tau^c \; s_\tau^A \; s_\tau^B]$ up to time $\tau = t$, where s_τ^A is solution to Eq. (9.50) and s_τ^B is solution to Eq. (9.47), such that b and B in Eq. (9.40) and (9.41) are,

$$b = [\sigma_f \;\; 0 \;\; 0 \;\; 0], \quad B = \begin{bmatrix} 0 & \sigma_d & 0 & 0 \\ 0 & 0 & \frac{\sigma_s}{\phi} & 0 \\ 0 & 0 & 0 & \sigma_s \end{bmatrix},$$

and then, by straightforward calculations, and Eq. (9.42), the dividend expected by agent A, $\hat{D}_t^A \equiv E\left(D_t|\,\mathbb{F}_t^A\right)$, satisfies

$$d\hat{D}_t^A = \lambda(\bar{D} - \hat{D}_t^A)dt + \alpha_* \left(\frac{1}{\sigma_d} dW_{At}^d + \frac{\phi}{\sigma_s} dW_{At}^A + \frac{1}{\sigma_s} dW_{At}^B\right), \tag{9.51}$$

where the Brownian motions will be defined in a moment, and α_* is the positive root of, $0 = (\frac{1}{\sigma_d^2} + \frac{\phi^2}{\sigma_s^2} + \frac{1}{\sigma_s^2})\alpha_*^2 + 2\lambda\alpha_* - \sigma_f^2$.

Likewise, the dividend expected by agent B, $\hat{D}_t^B \equiv E\left(D_t|\,\mathbb{F}_t^B\right)$, satisfies

$$d\hat{D}_t^B = \lambda(\bar{D} - \hat{D}_t^B)dt + \alpha_* \left(\frac{1}{\sigma_d} dW_{Bt}^d + \frac{1}{\sigma_s} dW_{Bt}^A + \frac{\phi}{\sigma_s} dW_{Bt}^B\right). \tag{9.52}$$

Finally, the Brownian motions in Eqs. (9.51)-(9.52) satisfy,

$$dW_{it}^d = \frac{1}{\sigma_d}(dD_t^c - \hat{D}_t^i dt), \quad dW_{it}^j = \frac{\phi_i}{\sigma_s}(ds_t^j - \hat{D}_t^i dt), \quad i,j \in \{A,B\}, \tag{9.53}$$

where,

$$\phi_i = \begin{cases} 1, & \text{if } i \neq j \\ \phi \geq 1, & \text{if } i = j \end{cases}$$

Naturally, the two standard Brownian motions W_t^A and W_t^B in Eq. (9.49) are a special case of those in Eq. (9.53), namely for $\phi \equiv 1$.

Let us define the differences in beliefs as,

$$g_t \equiv g_t^A \equiv \hat{D}_t^B - \hat{D}_t^A. \tag{9.54}$$

That is, a negative value of g_t is interpreted as a state in which agent A is more "optimistic" than B.

We wish to express the dynamics of g_t^A under the probability space of agent A, and the dynamics of g_t^B under the probability space of agent B. Let us consider the terms of disagreement between the two types of agents. As regards the dynamics of cumulative dividends, we have, by Eq. (9.53), that:

$$dW_{Bt}^d = dW_{At}^d - \frac{1}{\sigma_d} g_t dt,$$

and concerning the signals,

$$dW_{Bt}^A = \frac{1}{\phi} dW_{At}^A - \frac{g_t}{\sigma_s} dt, \quad \frac{1}{\phi} dW_{Bt}^B = dW_{At}^B - \frac{g_t}{\sigma_s} dt,$$

such that, using Eqs. (9.51)-(9.52),

$$dg_t = -\rho g_t dt + \sigma_g dW_{At}^g, \tag{9.55}$$

for some positive constant ρ and σ_g, and a Brownian motion W_{At}^g.[7] Differences in beliefs are mean-reverting to zero. Intuitively, both agents believe that the long term value of the instantaneous dividend is $\bar{D}$. Moreover, agents in group A think that future dividends will reflect their own beliefs, not the beliefs mistakenly held by agents in group B. Hence, they expect the difference in opinions to shrink over time. Naturally, agents in group B make a symmetrically opposite reasoning.

[7] The constants are $\rho \equiv \lambda + (\frac{1+\phi^2}{\sigma_s^2} + \frac{1}{\sigma_d^2})\alpha_*$ and $\sigma_g \equiv (\phi^2 - 1)\sqrt{1+\phi^{-2}}\sigma_s^{-1}\alpha_*$, and the Brownian motion satisfies: $\sqrt{1+\phi^{-2}} dW_{At}^g \equiv (dW_{At}^B - \frac{1}{\phi} dW_{At}^A)$.

9.6.2.2 Bubbles as re-sale American options

The distinctive mark of the model is that trading occurs between agents due to difference in beliefs, and leads to bubbles arising as a result of short-selling constraints, consistent with previous insights of Harrison and Kreps (1979). Intuitively, at any point in time, the asset is held by the relatively more optimistic agent, whose asset evaluation is higher than his own assessment of the fundamentals, expecting as he is to sell the asset to some future relatively more optimistic agent. Prices then deviate from any agent's fundamental evaluation because short-selling constraints bias the price towards the more optimistic agents.

When is the optimal time to sell? It is a real option problem, of the kind introduced in Chapter 5. Let $S_{A,t}$ denote the asset price at t that agents in group A are willing to pay, such that,

$$S_{A,t} = \sup_{\tau \geq 0} E_t^A \left(\int_t^{t+\tau} e^{-r(s-t)} dD_s^c + e^{-r\tau} \left(S_{B,t+\tau} - c \right) \right), \tag{9.56}$$

where E_t^A denotes the time t conditional expectation of agents A, D_t^c is the cumulative dividend process in Eq. (9.45), r is the constant interest rate, and c is a transaction cost, to be discussed below.

In words, the price agents A are willing to pay at t reflects the dividends paid out over the holding period and the price agents B are willing to pay at time $t + \tau$, net of the transaction cost. The conjecture is that the solution $S_{A,t}$ contains a bubble component,

$$S_{A,t} = \frac{\bar{D}}{r} + \frac{\hat{D}_t^A - \bar{D}}{r + \lambda} + \mathcal{B}\left(g_t\right), \tag{9.57}$$

The first two terms on the R.H.S. of Eq. (9.57) amount to the payoff of the asset over its life expected by agents in group A, $E_t^A(\int_t^\infty e^{-r(s-t)} dD_s^c)$—the fundamentals. According to the conjecture, agents A are willing to pay more than the fundamentals, with $\mathcal{B}\left(g_t\right)$ representing a bubble, a function of g_t, the difference in beliefs in Eq. (9.54). The bubble arises exactly because agents in group A bid up the price to their own evaluations and agents in group B cannot sell short.

By replacing Eq. (9.57) into Eq. (9.56), and using the definition of g_t in Eq. (9.54) leaves:

$$\mathcal{B}\left(g_t\right) = \sup_{\tau \geq 0} E_t^A \left(e^{-r\tau} \left(\frac{g_{t+\tau}}{r + \lambda} + \mathcal{B}\left(-g_{t+\tau}\right) - c \right) \right). \tag{9.58}$$

The bubble is a re-sale option: it arises because the current asset holders have the option to re-sell it in the future at a price higher than their own evaluation. Eq. (9.58) shows that its value is determined through an optimal stopping time to sell, similarly as with a perpetual American option (see Chapter 5), with the complication that the "strike" is endogeneous, reflecting the value of the bubble as perceived by the future buyer at the optimal stopping time.

By results in Chapter 5 (see Section 5.6), one can conjecture that there exists a region of inaction, where the asset owner holds the asset, and a threshold g^*, such that the optimal stopping time in Eq. (9.58) is $\tau^* = \inf\left(\tau : g_{t+\tau} \geq g_*\right)$. In particular, in the continuation region, where the agent does not sell, the (discounted) bubble is a martingale, such that by Itô's lemma and Eq. (9.55),

$$\mathcal{L}\left[\mathcal{B}\right] - r\mathcal{B} = 0, \quad g \leq g_*, \tag{9.59}$$

where $\mathcal{L}\left[\mathcal{B}\right] = \frac{1}{2}\sigma_g^2 \mathcal{B}'' - \rho g \mathcal{B}'$ is the infinitesimal generator for g in (9.55). In the stopping region, where $\mathcal{L}\left[\mathcal{B}\right] - r\mathcal{B} \leq 0$,

$$\mathcal{B}\left(g\right) = \frac{g}{r + \lambda} + \mathcal{B}\left(-g\right) - c, \quad g \geq g_*. \tag{9.60}$$

Finally, a solution to this optimal stopping time satisfies the usual smooth-pasting conditions: (i) the function solution to Eq. (9.59) for $g \leq g_*$ is the same as the function satisfying Eq. (9.60) at $g = g_*$; and (ii) the derivatives of these two functions are the same at $g = g_*$. Scheinkman and Xiong (2002) show that there is a solution to this problem, and that $g_* \geq 0$, and that $g_* = 0$ only when the transaction cost $c = 0$. To illustrate, agent A needs to become sufficiently pessimistic ($g_t \geq g^* > 0$) to be able to justify the transaction cost while selling the asset. Intuitively, Brownian motions hit zero infinitely many times, such that agents will trade very often, driven by their ever switching differences in opinions. The presence of a transaction cost mitigate the occurrence of a such a trading frenzy.

9.6.3 General equilibrium without frictions

We illustrate how to incorporate differences in beliefs in general equilibrium in a framework with complete markets. We begin with an introductory model, formulated by Kogan, Ross, Wang and Westerfield (2006), which actually addresses an old and important question brought about at least since Friedman (1953): how long would irrational investors survive in financial markets?

The technical issue arising while addressing this question links to how the agents' priors are aggregated in equilibrium: irrational agents are those who systematically believe in a model deviating from the truth, and never wish to learn, thereby holding "different beliefs" than those of the rational agents. Section 9.6.3.1 studies such a model and Section 9.6.3.2 generalizes it to a context of learning, in which there are no agents who are more or less rational than others, along the lines of Scheinkman and Xiong (2002), but within a framework of frictionless markets without short-sales constraints. The "sentiment" model of Dumas, Kurshev and Uppal (2009) is a special a case of this framework, and vividly illustrates the main predictions arising therefrom. It will be discussed in Section 9.6.3.3.

9.6.3.1 Difference in beliefs and extinction of irrational traders

Kogan, Ross, Wang and Westerfield (2006) (KRWW, in the sequel) analyze the market impact of a drastic source of differences in beliefs, arising due to the irrationality of some agents. We briefly review this model both because of its outstanding economic importance and the aggregation techniques needed to solve it. The ensuing aggregation results are generalized in the following subsection devoted to the analysis of disagreement and learning in general equilibrium with multiple agents and frictionless markets.

Consider an economy in which there is a riskless asset in zero net supply, and one risky asset that entitles to an instantaneous dividend flow. Dividends follow a geometric Brownian motion under the physical probability P, with parameters μ and σ,

$$\frac{dD_t}{D_t} = \mu dt + \sigma dW_t, \tag{9.61}$$

with obvious notation. Rational agents correctly believe dividends are as in Eq. (9.61). However, irrational agents think that the drift of the dividend process differs from that in Eq. (9.61),

$$\frac{dD_t}{D_t} = \left(\mu + \upsilon\sigma^2\right) dt + \sigma dW_{2t}, \tag{9.62}$$

where υ is a constant and W_{2t} is a second Brownian motion, to be defined in a moment. Eq. (9.62) formalizes the idea that irrational agents understand the market under "their own"

probability P^2 (say), on which $dW_{2t} = dW_t - \upsilon\sigma dt$ is a Brownian motion increment by Girsanov's theorem. If $\upsilon > 0$, the irrational agent is "optimistic," in that he views dividends grow more than justified by the true probability laws.

The change of probability from the benchmark rational agent to the irrational is given by the Radon-Nikodym derivative,

$$\eta_T \equiv \left.\frac{dP^2}{dP}\right|_{\mathbb{F}_T},$$

where $\mathbb{F}_T$ is the information set available to all agents, and the density process η_t is solution to

$$\frac{d\eta_t}{\eta_t} = \upsilon\sigma dW_t. \tag{9.63}$$

In this model, irrational agents do not attempt to learn from the history of dividends and their cognitive bias is not eliminated as a result.

KRWW assume that all agents consume at some time T, that they have CRRA equal to γ, and that they have the same endowments.[8] Therefore, the rational and irrational investors maximize

$$E\left(\frac{c_{1T}^{1-\gamma}}{1-\gamma}\right) \quad \text{and} \quad E^2\left(\frac{c_{2T}^{1-\gamma}}{1-\gamma}\right) = E\left(\eta_T \frac{c_{2T}^{1-\gamma}}{1-\gamma}\right), \tag{9.64}$$

respectively, where E^2 denotes the expectation under P^2 taken by the irrational agent, and the equality follows by a change of probability. Note that because η_T is strictly positive, the irrational agent becomes more and more aggressive as he "disagrees" more with the rational, where "disagreement" is captured by η_T.

Because markets are complete, consumption allocations can be determined by solving a central planner program (just as explained in the static context of Chapter 2), whose value is:

$$U_T = \max_{(c_i)_{i=1,2}: c_{1T}+c_{2T}=D_T} \left[f_1 \frac{c_{1T}^{1-\gamma}}{1-\gamma} + f_2 \eta_T \frac{c_{2T}^{1-\gamma}}{1-\gamma}\right], \tag{8.P1}$$

where f_i denotes the reciprocal of the marginal utility of income for agent i. The solution is

$$c_{1T} = \frac{1}{1+(b\eta_T)^{1/\gamma}} D_T, \quad c_{2T} = \frac{(b\eta_T)^{1/\gamma}}{1+(b\eta_T)^{1/\gamma}} D_T, \quad b \equiv \frac{f_2}{f_1}. \tag{9.65}$$

As anticipated, the irrational agent receives an allocation that increases with his disagreement: it is *as if* his utility function in (9.64) had a random weight, η_T, arising through how much he will disagree regarding the dividend path leading to D_T at the consumption date T.

We need to determine the ratio of marginal utilities of income, b. Replace the consumption allocations in (9.65) into the central planner program [8.P1], leaving:

$$U(D_T, \eta_T) \equiv U_T = \left(f_1^{1/\gamma} + (\eta_T f_2)^{1/\gamma}\right)^{\gamma} \frac{D_T^{1-\gamma}}{1-\gamma},$$

such that, denoting partial derivatives with subscripts, the pricing kernel is given by

$$\xi_T = \frac{U_1(D_T, \eta_T)}{E_t\left[U_1(D_T, \eta_T)\right]} = \frac{(1+(b\eta_T)^{1/\gamma})^{\gamma} D_T^{-\gamma}}{E_t\left[(1+(b\eta_T)^{1/\gamma})^{\gamma} D_T^{-\gamma}\right]}. \tag{9.66}$$

[8] Work on extinction by Cvitanić and Malamud (2011) relaxes the assumption that rational and irrational agents have the same mass.

Note that because consumption only takes place at terminal date, T, the denominator of ξ_T contains the expectation of the representative agent's marginal utility, not its value at zero as it is instead the standard case (Chapter 8, Section 8.5.2). As shown in Chapter 5 (see Appendix 2), ξ_T is the pricing kernel that determines the equilibrium prices using the money market account as the numéraire. KRWW show that (see Appendix 2 for a few steps underlying this derivation),

$$b = e^{-(1-\gamma)\upsilon\sigma^2 T}. \tag{9.67}$$

Eqs. (9.66) and (9.67) can now be used to dermine the asset price and the relative consumption share. Regarding the asset price process (expressed in terms of the money market account numéraire), we have that, $S_t = E_t(\xi_T D_T)$. When agents have logarithmic preferences, $\gamma = 1$, the expression got S_t simplifies to

$$S_t = \frac{1+\eta_t}{1+\eta_t \cdot e^{-\upsilon\sigma^2(T-t)}} S_t^*, \qquad S_t^* = e^{\mu(T-t)} D_t, \tag{9.68}$$

where S_t^* denotes the price in an economy that is only populated by rational agents, i.e. $\upsilon = 0$. That is, the presence of "optimistic" agents ($\upsilon > 0$) inflates asset prices beyond rational evaluations.

The density process can, thus, be interpreted as "sentiment," leading as it does to the following "euphoria." Suppose dividends are hit by a positive shock; then, the asset price goes up because both S_t^* and η_t go up in Eq. (9.68). That is, prices are driven by both rational evaluations and sentiment. The term "sentiment" has been used in a context with irrational agents and learning by Dumas, Kurshev and Uppal (2009) (see Section 9.6.3.3).

This property can be generalized to the case of any CRRA coefficient. Indeed, given the expression for ξ_τ in Eq. (9.66), we can rely on results in Chapter 8 (Section 8.5.2), and determine the unit risk-premium λ_τ in this economy. It is:

$$\lambda_t = \sigma\gamma + \frac{(b\eta_t)^{1/\gamma}}{1+(b\eta_t)^{1/\gamma}} \frac{g}{\sigma}, \quad g \equiv -\upsilon\sigma^2, \tag{9.69}$$

and is less than that in the rational economy, due to the permanent difference difference in beliefs, $E_t(\frac{dD_t}{D_t}) - E_t^2(\frac{dD_t}{D_t}) = g < 0$. All in all, the aggregate risk-premium in this economy is lower, reflecting the optimistic view of the irrational agents: asset prices (expressed in terms of the money market numéraire) are higher in this economy than in the purely rational. Note, also, that the aggregate risk-appetite is time-varying and driven by market sentiment, η_t. In good times (i.e., after a positive dividend shock), sentiment increases, such that λ_t lowers in Eq. (9.69), leading to "speculative enthusiasm" and higher asset evaluations.

Would irrational traders ever "disappear" in the long term? Define the relative consumption share of the rational agent against the irrational, which by Eq. (9.65) and then (9.63) and (9.67) is,

$$\frac{c_{2T}}{c_{1T}} = (b\eta_T)^{1/\gamma} = e^{-\frac{1}{\gamma}\left(1-\gamma+\frac{1}{2}\upsilon\right)\upsilon\sigma^2 T + \frac{1}{\gamma}\upsilon\sigma W_T}.$$

KRWW define *relative extinction* of irrational traders occurs whenever, $\lim_{T\to\infty} \frac{c_{2T}}{c_{1T}} = 0$ almost surely. By the Strong Law of Large Numbers for Brownian motions (Karatzas and Shreve, 1991), we have that given two constants a_0 and a_1,

$$\lim_{T\to\infty} e^{a_0 T + a_1 W_T} = \begin{cases} 0, & a_0 < 0 \\ \infty, & a_0 > 0 \end{cases}$$

irrespective of the sign of a_1.

In the log-utility case, $\gamma = 1$, the irrational trader does not survive. For $\gamma > 1$, relative extinction occurs whenever the irrational agent is either pessimistic, $\upsilon < 0$, or "too" optimistic, $\upsilon > 2(\gamma - 1)$. However, the irrational trader would survive as soon as he is "moderately optimistic," $0 < \upsilon < 2(\gamma - 1)$, in which case the rational agent would not survive. The interpretation for the rational trader extinction is that the irrational trader has a price impact that makes low dividend states cheaper than the good, such that the rational trader accumulates more wealth in bad states than in the (relatively more likely) better states.

9.6.3.2 Differences in beliefs in markets with multiple agents

We consider an economy in which n agents hold different beliefs regarding the fundamentals of the economy, generalizing the two-person setting originally formulated by Detemple and Murthy (1994), Zapatero (1998), and others such as Basak (2000, 2005), Berrada (2006), or Buraschi and Jiltsov (2006). For example, some agent, the "benchmark," may think that output has expected growth $\hat{f}_t$ and constant volatility σ_0,

$$\frac{dD_t}{D_t} = \hat{f}_t dt + \sigma_0 dW_t. \tag{9.70}$$

We define the agents with these beliefs as the benchmark, in that any other agent's beliefs are about to be gauged against theirs, as explained below. Accordingly, and for a given T, consider the Radon-Nikodym derivative of the probability P^i of agent i against the benchmark $i = 1$, and its density process, η_{it} say, which by Girsanov's theorem satisfies

$$\frac{d\eta_{it}}{\eta_{it}} = -g_{it}\frac{1}{\sigma_0}dW_t, \quad g_{it} \equiv \hat{f}_t - \hat{f}_{it}. \tag{9.71}$$

For each agent i, the dynamics of output is,

$$\frac{dD_t}{D_t} = \hat{f}_{it} dt + \sigma_0 dW_{it}, \tag{9.72}$$

where $dW_{it} = dW_t + g_{it}\frac{1}{\sigma_0}dt$ is a Brownian motion under P^i. This example generalizes the model in the previous section, in which $\hat{f}_t$ and $\hat{f}_{it}$ are constant, as summarized by Eq. (9.63).

We refrain from specifying additional examples and mechanisms leading to disagreement. For now, and until the next subsection, we focus on the determination of the equilibrium in this economy. Without loss of generality, define the first agent as the benchmark. We assume that markets are complete and that there is symmetric albeit incomplete information, and denote with $\mathbb{F}_t$ the information set available to all agents. The Radon-Nikodym derivatives,

$$\eta_{it} \equiv \left.\frac{dP^i}{dP^1}\right|_{\mathbb{F}_t}, \quad i \in \{2, \cdots, n\}, \tag{9.73}$$

formalize the notion of disagreement between agent i and the benchmark at t—a notion made more precise below.

Each agent maximizes his intertemporal utility, subject to the budget constraint,

$$\max_{(c_{it})_{t\geq 0}} E^i\left(\int_0^\infty e^{-\rho t} u^i(c_{it})\, dt\right), \quad \text{s.t.} \quad E^i\left(\int_0^\infty \xi_{it} c_{it} dt\right) = V_{0i}, \tag{8.P2}$$

where E^i denotes expectation under P^i, V_{0i} is the initial wealth available to agent i, and ξ_{it} is the "private" state-price process, or pricing kernel, of agent i, a concept we elaborate on in a moment. Remaining notation is straightforward.

We know since from Part I of the lectures that a state-price process links to the evaluation of a consumption unit at some future pre-specified state of nature. Heuristically, it is,

$$\xi_{it} \equiv \left.\frac{dQ}{dP^i}\right|_{\mathbb{F}_t} e^{-\int_0^t r_u du}, \tag{9.74}$$

where Q denotes, as usual, the risk-neutral probability, which links to the value of Arrow-Debreu securities, as we know. Note that ξ_{it} in Eq. (9.74) can vary across agents due to difference in opinions, although then the agents need to agree on the price of the assets they observe—they do, simply, have different perspectives regarding the future developments of the assets price.

To formalize the idea that agents disagree against a same benchmark, note that

$$\xi_{1t} = \eta_{it}\xi_{it},$$

and cast the program [8.P2] under the probability space of the benchmark agent, such that each agent i (note, i) maximizes his intertemporal utility subject to his budget constraint,

$$\max_{(c_{it})_{t\geq 0}} E^1\left(\int_0^\infty e^{-\rho t}\eta_{it}u^i\left(c_{it}\right)dt\right), \quad \text{s.t.} \quad E^1\left(\int_0^\infty \xi_{1t}c_{it}dt\right) = V_{0i}. \tag{8.P3}$$

The program [8.P3] is a standard complete markets problem in which each agent has an instantaneous utility function equal to $\eta_{it}u^i\left(c_{it}\right)$ and a commonly agreed state-price process ξ_{1t}, that of the benchmark agent. All agents do agree on the current price of all assets, although they now act more aggressively on their consumption as their divergence of opinion against the first agent widens, as summarized by an "utility distorsion factor" η_{it}. The program [8.P3] is the intermediate consumption, infinite horizon extension to the program of maximizing terminal utility in (9.64).

Because markets are complete, we can study the asset pricing implications of this economy by centralizing it as usual. Consider a representative agent with instantaneous utility equal to,

$$U_t = \max_{(c_i)_{i=1}^n} \sum_{i=1}^n \eta_{it}u^i\left(c_{it}\right)f_i, \quad \text{s.t.} \quad \sum_{i=1}^n c_{it} = D_t, \tag{8.P4}$$

where D_t denotes output at t, and f_i^{-1} is the marginal utility of *income* of agent i. We simplify the presentation, and to focus on the salient aspects of disagreement, we set $u^i \equiv u$ for all i. The first order conditions of the program [8.P4] are,

$$\eta_{it}u'\left(c_{it}\right)f_i = \psi_t,$$

where ψ_t is a Lagrange multiplier. Assuming agent have CRRA equal to γ, we can solve the previous equation for the consumption allocation to agent i, $c_{it} = \left(\frac{\psi_t}{\eta_{it}}\frac{1}{f_i}\right)^{-1/\gamma}$, plug it into the constraint of the program [8.P4], thereby determining the Lagrange multiplier, $\psi_t = \left(D_t^{-1}\sum_{i=1}^n\left(\eta_{it}f_i\right)^{1/\gamma}\right)^\gamma$ and, ultimately, the sharing rule regarding consumption allocation,

$$c_{it} = \frac{\left(\eta_{it}f_i\right)^{1/\gamma}}{\sum_{i=1}^n\left(\eta_{it}f_i\right)^{1/\gamma}}D_t. \tag{9.75}$$

Intuitively, and as anticipated, the more aggressive agents are, compared to the benchmark (i.e., the higher η_{it}), the higher their consumption allocation. Eq. (9.75) is, naturally, reminiscent of the consumption allocation in Eq. (9.65) applying to the KRWW model. However, disagreement, η_{it}, is now allowed to take a more general form than in (9.63): it is time-varying, possibly reflecting agents' learning in (9.71).

The value function for the central planner is obtained by replacing Eq. (9.75) into the maximand of the program [8.P4],

$$U\left(D_t,(\eta_i)_{i=1}^n\right)\equiv U_t=\left(\sum_{i=1}^n\left(\eta_{it}f_i\right)^{1/\gamma}\right)^{\gamma}\frac{D_t^{1-\gamma}}{1-\gamma}. \tag{9.76}$$

By the usual arguments (see Section 8.5.2 in Chapter 8), the pricing kernel in this economy is given by $\xi_{1t}\equiv\xi_t$, where

$$\frac{\xi_t}{\xi_0}=e^{-\rho t}\frac{U_C\left(D_t,(\eta_{it})_{i=1}^n\right)}{U_C\left(D_0,(\eta_{i0})_{i=1}^n\right)}, \tag{9.77}$$

where subscripts to U denote a partial derivative with respect to the first argument.

Note that to price any asset, we still need to determine the marginal utility of income for each agent i, f_i^{-1}. One approach could be to search for a cross-sectional distribution of Pareto weights that best fits selected moments of the price distribution, as Chan and Kogan (2002) did in their model surveyed in Section 9.3. Note that the cross-section of Pareto weights depends not only on wealth distribution, but also on beliefs'. Indeed, by results in Chapter 5 (see Section 5.5.2), we have that the reciprocal of the marginal utility of income is,

$$\frac{1}{f_i}=V_{0i}^{-\gamma}\left(\int_0^{\infty}e^{-\frac{\rho}{\gamma}t}E^1\left(\eta_{it}^{\frac{1}{\gamma}}\xi_{1t}^{1-\frac{1}{\gamma}}\right)dt\right)^{\gamma}.$$

The asset market equilibrium can now be studied by determining the risk-premiums, λ_t say, and interest rates as implied by the pricing kernel ξ_i for $i=1$ in Eq. (9.74),

$$\xi_{1t}\equiv\left.\frac{dQ}{dP^1}\right|_{\mathbb{F}_t}e^{-\int_0^t r_u du}\equiv\eta_t^Q e^{-\int_0^t r_u du},$$

where,

$$\frac{d\eta_t^Q}{\eta_t^Q}=-\lambda_t dW_t. \tag{9.78}$$

Finally, Eq. (9.77) shows that agents' disagreement, as summarized by the cross-section $(\eta_{it})_{i=1}^n$, is a potential source of increased volatility of the pricing kernel and, then, resolution of the equity premium and other puzzles. The model of the next section discusses a special case of this framework that seems to illustrate this role.

9.6.3.3 Two-person equilibrium

Dumas, Kurshev and Uppal (2009) (DKU, in the sequel) consider a model in which agents disagree on the fundamentals due to overconfidence, similarly as in the model of Scheinkman and Xiong (2002) discussed in Section 9.6.2. This model is a special case of that in the previous section: there are two agents ($n=2$) with the same CRRA coefficient γ, with one of them holding different beliefs due to overconfidence, as explained below.

The key assumptions of the model are that the expected growth of output D_t, f_t say, is unobserved, and that the only available information is the observation of D_t and one additional

signal that is totally uninformative about the state of the economy. However, an overconfidence investor believes this signal is correlated with the economic fundamentals, whence a difference in beliefs between him and a second, rational investor, the benchmark.

In detail, output is solution to

$$\begin{cases} \dfrac{dD_t}{D_t} = f_t dt + \sigma_0 dZ_t^o \\ df_t = \zeta(\bar{f} - f_t)dt + \sigma_f dZ_t^f \end{cases} \tag{9.79}$$

where σ_0, ζ, $\bar{f}$ and σ_f are constant, and Z_t^0 and Z_t^f are standard Brownian motions. Investors also observe an uninformative signal, solution to

$$ds_t = \sigma_s dZ_t^s.$$

where Z_t^s is another standard Brownian motion. However, the overconfident investor believes that this signal is correlated with the fundamentals, in that

$$s_t = s_t^* : ds_t^* = \sigma_s \left(\sqrt{1-\rho^2} dZ_t^s + \rho dZ_t^f \right).$$

These assumptions substantially match those in the published version of Scheinkman and Xiong (2002), as mentioned in Section 9.6.2.

Let us solve for the inference problem of the irrational investor, the problem of the rational being a special case, notably for $\rho = 0$. In terms of the learning problem of Section 9.6.1, we have that the vector Brownian motion is $Z_t = [Z_t^o \; Z_t^s \; Z_t^f]$, such that b and B in Eq. (9.40) and (9.41) are[9]

$$b = [0 \;\; 0 \;\; \sigma_f], \quad B = \begin{bmatrix} \sigma_0 & 0 & 0 \\ 0 & \sigma_s\sqrt{1-\rho^2} & \sigma_s\rho \end{bmatrix}, \tag{9.80}$$

and the inferred output growth $f_t^* \equiv E(f_t | \mathbb{F}_t^*)$ is

$$df_t^* = \zeta(\bar{f} - f_t^*)dt + \frac{\alpha_B}{\sigma_0} dW_t^* + \rho\sigma_f dW_{st}, \tag{9.81}$$

where $\mathbb{F}_t^*$ denotes the information set available to the irrational agent. Instead, the rationally expected output growth is $\hat{f}_t \equiv E(f_t | \mathbb{F}_t)$ and satisfies

$$d\hat{f}_t = \zeta(\bar{f} - \hat{f}_t)dt + \frac{\alpha_A}{\sigma_0} dW_t, \tag{9.82}$$

where $\mathbb{F}_t$ is the information set available to the rational agent. Remaining notation is as follows: α_A and α_B are solutions to Eq. (9.43), with $a_0 = \zeta\bar{f}$, $a_1 = -\zeta$, $A = [1 \; 0]$, and the matrix B is as in Eqs. (9.80), where α_A is determined by setting $\rho = 0$. Below, we shall define the two Brownian motions in Eqs. (9.81) and (9.82).

We take the rational agent to be the benchmark. In terms of the model in the previous subsection, he thinks output is as in Eq. (9.70), where $\hat{f}_t$ is his expected dividend growth, solution to Eq. (9.82)—and by construction, the two Brownians in Eqs. (9.70) and (9.82) coincide. The second overconfident investor disagrees, and thinks that the expected dividend growth is f_t^*, solution to Eq. (9.81).

[9]Note that the in Section 8.6.1, the dynamics of the signals, $\mathcal{S}_t$, is expressed in basis point terms, as opposed to the dynamics of output in Eq. (9.79). However, the inference yields the same result after suitable definition of the Brownian motions.

As in the previous subsections, the density process summarizing the difference in opinions between the irrational and the benchmark is the process η_t solution to Eq. (9.71) with $g_{it} \equiv g_t \equiv \hat{f}_t - f_t^*$. Thus, and consistent with similar conventions in previous sections, $g_t < 0$ is indicative of states in which the irrational agent is more optimistic than the rational agent. Note that from the perspective of the irrational agent, output is solution to Eq. (9.72), with $\hat{f}_{it} \equiv f_t^*$, and for the Brownian motion $W_{it} \equiv W_t^*$ in Eq. (9.81), such that $dW_t^* = dW_t + g_t \frac{1}{\sigma_0} dt$. Finally, the Brownian motion W_{st} is simply defined as $\frac{ds_t^*}{\sigma_s}$, and is therefore uncorrelated with W_t. The dynamics of disagreement, g_t, are

$$dg_t = -\psi g_t dt + \sigma_g dW_t - \rho \sigma_f dW_{st},$$

where $\psi \equiv \zeta + \alpha_B/\sigma_0^2$ and $\sigma_g \equiv (\alpha_A - \alpha_B)/\sigma_0$.

DKU term the density process η_t "sentiment" to emphasize it arises through overconfidence over the agents' information processing of publicly available signals. Because markets are complete, the asset pricing implications of this economy are obtained by relying on the instantaneous utility of a representative agent, which by Eq. (9.76), is

$$U_t = \frac{D_t^{1-\gamma}}{1-\gamma} \left(f_1^{1/\gamma} + (\eta_t f_2)^{1/\gamma} \right)^{\gamma},$$

where f_i denotes the reciprocal of the marginal utility of agent i, as usual.

By results in Chapter 8 (Section 8.5.2), we can determine interest rates and risk-premium and, then, the pricing measure, as indicated by Eq. (9.78). We can, equivalently, rely on the solution of the central planner. By Eq. (9.77), we have that the pricing kernel satisfies

$$\frac{\xi_t}{\xi_0} = e^{-\rho t} \frac{D_t^{-\gamma}}{D_0^{-\gamma}} \frac{\left(f_1^{1/\gamma} + (\eta_t f_2)^{1/\gamma} \right)^{\gamma}}{\left(f_1^{1/\gamma} + (\eta_0 f_2)^{1/\gamma} \right)^{\gamma}}, \tag{9.83}$$

such that the risk premium is

$$\lambda(\eta_t, g_t) = \sigma_0 \gamma + \frac{(f_2 \eta_t)^{1/\gamma}}{f_1^{1/\gamma} + (f_2 \eta_t)^{1/\gamma}} \frac{g_t}{\sigma_0}. \tag{9.84}$$

It is increasing in disagreement g_t, and increasing (resp. decreasing) in sentiment η_t if $g_t > 0$ (resp. $g_t < 0$). Similarly, one can show that the short-term rate at time t, r_t, is a function of $\hat{f}_t$, η_t and g_t, i.e. $r_t = r(\eta_t, g_t, \hat{f}_t)$ for some function $r(\cdot)$.

The risk-premium in Eq. (9.84), $\lambda(\eta_t, g_t)$, generalizes the risk-premium in the Kogan, Ross, Wang and Westerfield (2006) market (see Eq. (9.69) in Section 9.6.3.1), in that the risk-premium is now time-varying due to both sentiment and differences in opinion (differences in opinion never change in the model of Section 9.6.3.1). When the irrational agent is optimistic, $g_t < 0$, an increase in sentiment lowers the risk-premium to a level below that in the Lucas model. However, note that in times in which the irrational agent is pessimistic, $g_t > 0$, a higher sentiment η_t would exacerbate the equity premium.

Given the previous characterization of the risk-premium and the interest rate, it follows that the price-dividend ratio is driven by the three state variables η_t, g_t and $\hat{f}_t$, and is solution to a partial differential equation that has one of the usual forms seen in Chapter 8 for similar

problems. DKU actually note that a closed-form solution exists under the parameter restriction that γ is an integer; they note that by the binomial formula,

$$\left(f_1^{1/\gamma} + (\eta_t f_2)^{1/\gamma}\right)^{\gamma} = f_1 \sum_{k=0}^{\gamma} \binom{\gamma}{k} \left(\frac{f_2}{f_1}\eta_t\right)^{\frac{k}{\gamma}}.$$

Relying on this formula and the expression for the pricing kernel in Eq. (9.83), one finds that the price-dividend ratio satisfies

$$\begin{aligned}\int_0^{\infty} E\left(\frac{\xi_t}{\xi_0}\frac{D_t}{D_0}\right) dt &= \int_0^{\infty} e^{-\rho t} \sum_{k=0}^{\gamma} \binom{\gamma}{k} \frac{\left(\frac{f_2}{f_1}\eta_0\right)^{k/\gamma}}{\left(1+\left(\frac{f_2}{f_1}\eta_0\right)^{1/\gamma}\right)^{\gamma}} E\left(\left(\frac{D_t}{D_0}\right)^{-\gamma}\left(\frac{\eta_t}{\eta_0}\right)^{k/\gamma}\right) dt \\ &\equiv p(\eta_0, g_0, \hat{f}_0),\end{aligned}$$

where the dependence of p on g_0 and $\hat{f}_0$ arises through the expectation inside the integral, which DKU calculate in closed form.

Asset prices are, then, driven by three state variables: (i) sentiment, η_t, (ii) difference in opinions, g_t, and (iii) the correctly specified expected output growth, $\hat{f}_t$. The model generates plausible values of both volatility and the equity premium. The main determinant of the equity premium is the rational agent's risk-aversion to the sources of risk introduced by the irrational, summarized by sentiment and difference in opinions, η_t and g_t, as illustrated by the risk-premium in Eq. (9.84). Expected returns and volatility are, then, high, compared to an economy with rational agents (i.e. with $\rho = 0$), with rational investors increasing the proportion of their wealth in equity only when their evaluation of the expected fundamentals, $\hat{f}_t$, increases by a considerable amount. DKU also find that the time needed for the extinction of the irrational investors is quite high.

9.7 Coping with Knigthian uncertainty

9.7.1 Prelude

9.7.1.1 Knightian uncertainty

This section still studies asset prices in economies where agents have limited knowledge of the statistical laws for the fundamentals. The new element is that we assume that agents give up thinking of having a single model to decipher the signals they receive. Rather, they formulate *multiple priors* underlying the laws of the fundamentals, and act while being averse to the "uncertainty" inherent their own priors.

Note the difference between this approach and that in the previous section. In the previous section, limited knowledge of the fundamentals had led us to consider economies in which agents disagree on the "right" model. Naturally, disagreement is not logical necessity given limited knowledge; however, it is a natural assumption in this context, as exemplified by the overconfidence bias models dealt with in Section 9.6.2.[10] Still, the previous models with disagreement rely on the assumption that agents have a unique prior with which they interpret

[10]Differences in beliefs do not necessarily arise through overconfidence. Cujean (2013) develops a two-agent model in which expected dividend growth is unknown, with one agent thinking it is a continuous process, and another, thinking it is a discrete Markov chain instead.

the complex world where they live. Instead, in this section, limited knowledge leads agents to be skeptical about their own ability to process complicated pieces of information. The agents acknowledge that many explanations are possible regarding how the economy works. We survey models that allow for this line of reasoning but, to simplify, consider only one representative agent.

The context in which agents operate in this section has come to be dubbed as *Knightian uncertainty* (Keynes, 1921; Knight, 1921), or *ambiguity*, that is, uncertainty that cannot be quantified probabilistically. In fact, one may argue that, since the dawn of financial economics, the entire building blocks of our models have been relying on the assumption that everything could be quantified probabilistically; the leading examples in these early developments are trivially easy to detect in the precursory mean-variance model of Markowitz (1952) and in the subsequent work leading to the CAPM (see Chapter 1).

But introducing Knightian uncertainty in economics is not a trivial task. We would need to know how to model rational decision makers who face uncertainty. Standard decision theory would not be helpful while thinking of uncertainty. We now proceed to a short survey of the literature on ambiguity, while insisting on the key references in decision theory on which models can be built. In subsequent sections, we show how these decision-theoretic foundations can be relied upon to build up models that could be used to address the typical issues arising in financial economics.

9.7.1.2 Survey notes regarding theoretic aspects of Knightian uncertainty

From the economic-theoretic standpoint, the important issue regards how to model aversion to uncertainty.[11] One of the first approaches to emerge relies on the so-called "capacities" as explained in more detail below. This approach goes back to at least Schmeidler (1982, published in 1989). The idea underlying capacities is to make rigorous use of non-additive measures to formalize the concept of "loss in probability" as an attitude towards uncertainty in the context of decision theory.

Dow and Werlang (1992) are the first to analyze the implications of this theory in the context of portfolio selection. First, they show, agents do not trade when prices are not favourable enough. Second, and within their simple introductory example, they illustrate very clearly a fundamental albeit somehow technical result known since at least Schmeidler (1986) and further elaborated by Gilboa and Schmeidler (1989) in a context of decision theory: once capacities are convex, the agent's behaviour is the same as that of an agent who has a max-min criterion. Max-min criteria are best described as decision rules the agents implement while believing nature will draw worst-case scenario events. Agents, then, take "robust" decisions, in that their choices will lead to outcomes that are the ideal ones in bad times.

Max-min criterions lead to an analytically convenient framework utilized in both finance and macroeconomics, as explained in this section. The approach to max-min in situations of uncertainty was originally advocated by Wald (1950), Ellsberg (1961) and Rawls (1971), and axiomatized by Gilboa and Schmeidler (1989). The max-min criterion of choice has been extended to "smooth" formulations that allow to disentangle a cognitive notion of uncertainty from the attitude towards it.

Provide references in decision theory and on work in macroeconomics and finance.

[In progress]

[11] Gilboa and Marinacci (2011) provide a survey of Knightian uncertainty in economics and decision theory.

9.7.1.3 Plan of the section

[In progress]

9.7.2 Uncertainty aversion and Ellsberg paradox

The Ellsberg paradox (Ellsberg, 1961) describes situations in which agents prefer to take considerable amounts of risk rather than to engage into situations plagued with ambiguity. To illustrate this paradox, consider the following example relying on a two-period market, zero interest rates, and a risk-neutral Robinson Crusoe agent.

We assume that there are three Arrow-Debreu securities that pay off in three mutually exclusive states of the world, states A, B and C (see Table 9.2 below). We initially assume that the probability of state A is $\frac{1}{3}$ and that of state B is $\frac{1}{4}$, such that the securities values, $\phi_\cdot$ say, are

$$\phi_{\mathrm{A}} = \frac{1}{3}, \quad \phi_{\mathrm{B}} = \frac{1}{4}, \quad \phi_{\mathrm{C}} = \frac{5}{12}, \tag{9.85}$$

with straightforward notation. Obviously, then, the value of a security paying off in states A or C is higher than that paying off in states B or C,

$$\phi_{\mathrm{A}\cup\mathrm{C}} = \frac{9}{12}, \quad \phi_{\mathrm{B}\cup\mathrm{C}} = \frac{8}{12}. \tag{9.86}$$

That is, the ranking of two portfolios is preserved once we include the *same* additional assets in each of these portfolios, provided the additional assets pay in states of the world in which the initial portfolios do not.

Arrow-Debreu securities	states A	B	C
A	1	0	0
B	0	1	0
C	0	0	1
A ∪ C	1	0	1
B ∪ C	0	1	1

TABLE 9.2. Arrow-Debreu securities in an uncertain market.

Next, suppose that the probability of state B, q say, is unknown. It is a situation of *uncertainty*, not *risk*. We have:

$$\phi_{\mathrm{A}} = \frac{1}{3}, \quad \phi_{\mathrm{B}} = q, \quad \phi_{\mathrm{A}\cup\mathrm{C}} = 1 - q, \quad \phi_{\mathrm{B}\cup\mathrm{C}} = \frac{2}{3}.$$

How would we expect Mr Crusoe to rank these assets? We are actually stuck without making additional assumptions regarding his attitude vis-à-vis uncertainty. We may proceed as follow. We assume that Mr Crusoe is risk-neutral but at the same time so averse to uncertainty that if hypothetically asked to go long on these assets, he would evaluate them at worst-case scenario. To formalize this uncertainty, we may assume that Mr Crusoe may conceive a "band" within which the unknown probability q lies, $q \in (\underline{q}, \bar{q})$. The wider this band, the more averse he is to uncertainty regarding q. Note that $\underline{q}$ and $\bar{q}$ can be interpreted both in terms of *uncertainty*

aversion and *cognitive* terms, that is, in terms of the "extent of ignorance" about q. We address this issue later; for now, we interpret this band as indicating Mr Crusoe's aversion to uncertainty.

How does asset evaluation reflect uncertainty aversion? Assuming that Mr Crusoe evaluates the asset at worst-case scenarios,

$$\hat{\phi}_{\mathrm{B}} \equiv \min_{q\in(\underline{q},\bar{q})} q = \underline{q} \quad \text{and} \quad \hat{\phi}_{\mathrm{A}\cup\mathrm{C}} \equiv \min_{q\in(\underline{q},\bar{q})} (1-q) = 1-\bar{q}.$$

An interesting phenomenon results. The ranking summarized by Eqs. (9.85)-(9.86) may break down. In particular, if Mr Crusoe sets his band such that $\underline{q} = \frac{1}{3} - \epsilon_1$ and $\bar{q} = \frac{1}{3} + \epsilon_2$, for two positive numbers ϵ_1 and ϵ_2 small enough, then,

$$\phi_{\mathrm{A}} > \hat{\phi}_{\mathrm{B}} \quad \text{and} \quad \hat{\phi}_{\mathrm{A}\cup\mathrm{C}} < \phi_{\mathrm{B}\cup\mathrm{C}}. \tag{9.87}$$

That is, uncertainty aversion might actually undermine a "sure thing" principle, a concept we shall return to in a moment.

The previous example is consistent with experimental evidence initially provided by Ellsberg (1961), who showed that individuals tend to avoid situations where it is difficult to describe events probabilistically. Klibanoff, Marinacci and Mukerji (2005) explain this context by relying on the following "lottery counterpart" to the Arrow-Debreu security example in Table 9.2.

	states		
Lotteries	A	B	C
f	1	0	0
g	0	1	0
f'	1	0	1
g'	0	1	1

TABLE 9.3. Lotteries in an uncertain environment.

Savage's axiom P2, known as the *sure thing principle*, would tell us that in this context and for any decision maker,

$$f \succ g \Longrightarrow f' \succ g'.$$

Yet assume that the probability of state B is unknown, similarly as in the example of Table 9.2. Experimental evidence is consistent with the hypothesis that in this case, decision makers would prefer the risky lottery g' (paying off £1 with probability $\frac{2}{3}$) rather than the lottery f' (paying off £1 with unknown probability), even if their preferences would lead them to choose f over g. In other words, aversion to uncertainty entails the following counterpart to the inequalities in (9.87),

$$f \succ g \quad \text{and} \quad f' \prec g'.$$

The previous examples reveal that there are new elements in asset evaluation in the presence of aversion to Knightian uncertainty. We now proceed with a few key models that provide new predictions in this context.

[In progress]

9.7.3 Portfolio selection and market participation

9.7.3.1 A static model

Dow and Werlang (1992) model aversion to uncertainty relying on the notion of *capacities*. Heuristically, capacities are non-additive measures, in that they do not sum up to one from the perspective of an uncertainty-averse decision maker. For example, an investor may be unaware of the distribution of the asset discounted payoff, yet he may consider that the asset discounted payoff is high (H) with probability no lower than π, or low (L) with probability no lower than π', with these "probabilities" being such that $\pi + \pi' < 1$. In other words, the (unknown) probability that the asset payoff is H, π_o say, satisfies $\pi \leq \pi_o \leq 1 - \pi'$. We now illustrate how the fact that $\pi + \pi' < 1$ reflects the investor's aversion to uncertainty. We shall explain that while the investor does not know the true distribution of the asset payoff, he evaluates the asset by assigning low chances of good outcomes.

Suppose a risk-neutral agent is contemplating buying the asset. The worst-case scenario for him is that the asset payoff is L; however, there are chances this worst-case scenario could "improve" by an amount equal to $H - L$. That is, this improvement occurs with probability at least equal to π. The "minimum" expected improvement is, thus, $\pi (H - L)$, such that the "minimum" expected return from being long the asset is $\pi_b \equiv (L + \pi (H - L)) - p$, where p denotes the asset price, as usual. The expected improvement and expected returns are "minimum" because, by assumption, the true probability is taken to satisfy $\pi_o \geq \pi$. Now, consider a risk-neutral agent who calculates expected returns at these minimum levels; he will buy the asset if $\pi_b > 0$, i.e., whenever the price satisfies

$$p < \underline{p} \equiv L + \pi (H - L) . \tag{9.88}$$

Similarly, consider a risk-neutral agent who contemplates selling the asset. His worst-case scenario is that the asset is actually good, in which case his payoff is $-H$ and the "minimum" expected improvement is $\pi' (H - L)$, such that the expected return from being short the asset is $\pi_s \equiv -(H + \pi' (L - H)) + p$. Similarly as in the buy-case, the expected improvement is the minimum one because by assumption, the true probability that the asset is L is $1 - \pi_o \geq \pi'$. Our agent would now sell the asset when $\pi_s > 0$, i.e., when the price satisfies

$$p > \bar{p} \equiv H - \pi' (H - L) . \tag{9.89}$$

Note how aversion to uncertainty operates. Conditions (9.88) and (9.89) tell us that the lower π and π', the more averse to uncertainty you are. Aversion to uncertainty leads the agent to presume that $\pi + \pi' < 1$. Let us elaborate. By (9.88), the agent buys when $p < \underline{p}$ and, by (9.89), the agent sells when $p > \bar{p}$, where $\underline{p} \leq \bar{p}$, with an equality holding only when $\pi + \pi' = 1$. That is, and unless the agent is uncertainty neutral (i.e., $\pi + \pi' = 1$), the agent will not participate in the market when the price $p \in (\underline{p}, \bar{p})$. When $p \in (\underline{p}, \bar{p})$, indeed, the price is too high to "break-even" while deciding to buy, and too low to break-even while deciding to sell. Dow and Werlang (1992, p. 200) define the quantity $1 - \pi - \pi'$ as "the amount of probability lost by the presence of uncertainty aversion."

9.7.3.2 Worst-case scenario interpretation

We can re-interpret the previous behavior in terms of decisions made under worst-case scenarios. That is, we may assume, now, that the agent relies on a set of addititive priors, i.e., on fully specified probability distributions. The previous model can then be interpreted as one in which the agent picks up the worst-case distribution according to the acts he makes (buy or sell).

Let λ be the unknown probability that the asset payoff is good, such that the expected profits can be written as

$$\pi_\lambda \equiv (\lambda H + (1-\lambda) L - p) x,$$

where x denotes the position in the asset, with $|x| = 1$ because we are assuming that the agent can only buy or sell one unit of the asset. Note that λ "sums up to one" now.

This model can be made consistent with the previous capacity framework. In the previous model, the asset is worth H with a probability *at least* π, meaning that the asset is worth L with probability *at most* $1-\pi$. Moreover, in the previous model, the asset is worth L with probability *at least* π', meaning that the asset is worth H with probability *at most* $1-\pi'$. Therefore, when the uncertainty averse agent buys, he does so while thinking that the probability $\lambda \in (\pi, 1-\pi')$ that the asset payoff is H is as small as possible. Similarly, when he sells, he does so while thinking that the probability $1-\lambda \in (\pi', 1-\pi)$ that the asset payoff is L is also as small as possible. That is, when the agent buys, the profits he expects are:

$$\min_{\lambda \in (\pi, 1-\pi')} \pi_\lambda = L + \pi (H - L) - p = \underline{p} - p,$$

and when he sells,

$$\min_{1-\lambda \in (\pi', 1-\pi)} (-\pi_\lambda) = p - H + \pi' (H - L) = p - \bar{p}.$$

Thus, this model collapses to the previous capacity model, as claimed. it also provides additional rationale regarding the probability bands underlying the Arrow-Debreu asset evaluation in Section 9.7.2.

A technical detour: Provide definitions/relations amongst capacities, convex measures, cores, etc.

[In progress]

9.7.3.3 Maxmin preferences

The previous model can be extended to one in which agents are risk averse. Suppose that an asset pays off a dividend $f \sim N\left(\mu, \sigma_f^2\right)$, where the mean μ is unknown albeit presumed to belong to a certain band, $\mu \in [\underline{\mu}, \bar{\mu}]$, for two constants $\underline{\mu}$ and $\bar{\mu}$. We shall come back to the interpretation of this band below. Suppose that a safe asset is elastically supplied at an interest rate equal to zero, and denote as usual the price of the risky asset with p. A risk-averse agent (with CARA equal to τ) chooses to invest $x \cdot p$ dollars in the risky asset while solving the following optimization problem:

$$\hat{x} = \arg\max_x \min_{\mu \in [\underline{\mu}, \bar{\mu}]} E_\mu \left(-e^{-\tau W_x}\right), \tag{9.90}$$

where $W_x = (f - p) x$, and E_m denotes the expectation taken under the probability law that results when $\mu = m$.

Knightian uncertainty arises in this model due to lack of knowledge regarding the dividend distribution: while the agent knows f is normally distributed, he is unaware of its exact location, μ. The presumption that $\mu \in [\underline{\mu}, \bar{\mu}]$ might indicate his aversion to this uncertainty: the wider he presumes this band is, the more uncertainty averse he is. One alternative interpretation of the band is that this very same band plays a merely cognitive role, in that the agent *only knows* that there are two bounding constants to the truth, $\underline{\mu}$ and $\bar{\mu}$. In this model, it is actually impossible to disentangle the cognitive notion of uncertainty from the agent attitude towards

it. There are more general models of ambiguity in which these two notions can be separated, surveyed in the next subsection.

The optimization problem in (9.90) is solved as follows. First, the agent solves for the inner problem; that is, he takes a portfolio choice x as given and, then, figures out what a malevolent Nature would pick up for him given his portfolio choice. For example, if the agent buys the asset, $x > 0$, it easy to see that the solution for the inner problem is $\hat{\mu} \equiv \underline{\mu}$. Analogously, if $x < 0$, then, the inner solution is $\hat{\mu} \equiv \bar{\mu}$. In other words, the agent's aversion to uncertainty leads him to presume that Nature will pick up a bad asset for him when he buys ($\hat{\mu} = \underline{\mu}$), and a good one when he sells ($\hat{\mu} = \bar{\mu}$).

Given the solution to the inner problem, $\hat{\mu}$, the agent proceeds with solving for his portfolio choice. The first order conditions lead to

$$\hat{x} = \frac{E_{\hat{\mu}}(f) - p}{\tau \sigma_f^2}, \quad \hat{\mu} \in \left\{\underline{\mu}, \bar{\mu}\right\}.$$

We have that $\hat{x} > 0 \iff E_{\underline{\mu}}(f) > p$ and $\hat{x} < 0 \iff E_{\bar{\mu}}(f) < p$, such that the solution to the overall optimization problem in (9.90) is

$$\hat{x} = \begin{cases} \dfrac{E_{\underline{\mu}}(f) - p}{\tau \sigma_f^2}, & \text{for } p < E_{\underline{\mu}}(f) \quad \text{(buy region)} \\ 0, \quad \text{for } p \in \left[E_{\underline{\mu}}(f), E_{\bar{\mu}}(f)\right] & \text{(non-participation region)} \\ \dfrac{E_{\bar{\mu}}(f) - p}{\tau \sigma_f^2}, & \text{for } p > E_{\bar{\mu}}(f) \quad \text{(sell region)} \end{cases} \tag{9.91}$$

The interpretation of the agent's behavior is similar to that regarding Dow and Werlang model. The agent participates in the market if the price is sufficiently favourable to him, compared to the worst-case scenario: he buys (sells) when the price is lower (higher) than his own most pessimistic (optimistic) expectation of the asset payoff. In case the price is not favourable enough, the agent does not participate. Mele and Sangiorgi (2015) first provide the solution to a portfolio choice that generalizes $\hat{x}$ in (9.91) to markets with asymmetric information.

It is interesting to examine the equilibrium implications of the model. Suppose that the asset is in positive supply, $\bar{x}$, say. In equilibrium, $\hat{x} = \bar{x}$. Then, the equilibrium price is

$$p = E_{\underline{\mu}}(f) - \tau \sigma_f^2 \bar{x}.$$

That is, the equilibrium price reflects the most pessimistic evaluation of the dividend. This property propagates to dynamics models, as explained below.

9.7.3.4 Smooth ambiguity aversion

Klibanoff, Marinacci and Mukerji (2005) (KMM, in the sequel) introduce a model of ambiguity in which the cognitive notion of uncertainty can be disentangled from the attitude towards it. We illustrate the main features of this model while extending the previous market, as follows. Conditionally on μ, the asset payoff f is still normally distributed, but uncertainty aversion is modeled assuming that there exists an increasing and concave function $h : \mathbb{R} \to \mathbb{R}$, such that a decision maker prefers the portfolio holdings x' to x if and only if

$$E_{\mathcal{M}}[h(E_{\mu}(-e^{-\tau W_{x'}}))] \geq E_{\mathcal{M}}\left[h\left(E_{\mu}\left(-e^{-\tau W_{x}}\right)\right)\right], \tag{9.92}$$

where $\mathcal{M}$ denotes the set of priors on μ. For example, Mele and Sangiorgi (2015) assume that $\mu \sim N\left(\mu_0, \sigma_\mu^2\right)$ in their model with asymmetric information, an assumption made in this section too.

Concavity of h is crucial to the definition of ambiguity aversion. Let us explain. Lack of knowledge of μ is the source of uncertainty in the model, and concavity of h implies that a decision maker dislikes mean-preserving spreads in expected utility values that arise due to μ: he is thus ambiguity averse, unless h is linear, in which case he is ambiguity neutral. Therefore, in this model, Knightian uncertainty arises because μ is unknown, with σ_μ^2 measuring how acute uncertainty is. What makes the model new compared to one with only "second-order uncertainty" (i.e., the presence of a stochastic mean in the asset payoff) is the aversion to this second-order uncertainty, arising through the concavity of h.

Regarding the functional function for h, one may consider $h(u) = -\frac{1}{\sigma}\left(e^{-\sigma u} - 1\right)$, where the parameter σ measures "absolute ambiguity aversion," with the model collapsing to one with an ambiguity neutral agent only when $\sigma = 0$. Based on KMM (Proposition 3), one can show that maxmin expected utility obtains for σ large. This section follows Mele and Sangiorgi, in that h is taken to lead to "constant relative ambiguity aversion," i.e., $h(u) = -(-u)^\alpha$, for some constant $\alpha \geq 1$. Thus, the model collapses to one with an ambiguity neutral agent only if $\alpha = 1$; for otherwise, the higher α, the more averse to ambiguity the agent is, independent of the extent of parameter uncertainty about μ, which is summarized by σ_μ^2, as explained.

The agent, then, solves the following optimization problem:

$$\hat{x} = \arg\max_x E_{\mathcal{M}}\left[h\left(E_\mu\left(-e^{-\tau W_x}\right)\right)\right], \tag{9.93}$$

where Appendix 3 shows that

$$E_{\mathcal{M}}\left[h\left(E_\mu\left(-e^{-\tau W_x}\right)\right)\right] = -\exp\left(-\alpha\tau\left(\mu_0 - p\right)x + \frac{1}{2}\alpha\tau^2\left(\alpha \text{var}\left(f\right) + \left(1-\alpha\right)\text{var}\left(f|\,\mu\right)\right)x^2\right). \tag{9.94}$$

Thus, the program of this ambiguity-averse agent resembles that of a mean-variance agent, although the "variance term" is replaced by a convex combination of the unconditional variance $\text{var}(f) = \sigma_f^2 + \sigma_\mu^2$ and the conditional variance $\text{var}(f|\,\mu) = \sigma_f^2$, such that $\text{V} \equiv \alpha\text{var}(f) + (1-\alpha)\,\text{var}(f|\,\mu) = \sigma_f^2 + \alpha\sigma_\mu^2$. If the agent were ambiguity neutral, $\alpha = 1$, $\text{V} = \sigma_f^2 + \sigma_\mu^2$, such that the problem would be indistinguishable from a standard mean-variance program with an increased variance: in other words, second-order uncertainty would not matter. Instead, second-order uncertainty matters when the agent is ambiguity averse, in which case the solution in (9.93) is

$$\hat{x} = \frac{\mu_0 - p}{\tau\left(\sigma_f^2 + \alpha\sigma_\mu^2\right)}.$$

The more averse to ambiguity the agent is, the higher α, and the less aggressive his portfolio holding will be for a given uncertainty level σ_μ^2. The equilibrium implications of the model are straightforward. With the asset in positive supply, $\bar{x}$, the equilibrium price is $p = \mu_0 - \tau\left(\sigma_f^2 + \alpha\sigma_\mu^2\right)\bar{x}$, with an "uncertainty premium" component, being clearly related to both uncertainty, σ_μ^2, and uncertainty aversion, α.

9.7.4 A model of multiple likelihoods

Leippold, Trojani and Vanini (2008) (LTV, in the sequel) develop a model in which the drift of dividend growth is unobservable, similarly as in the learning models of Chapter 8 (see Section 8.5.4) and the previous Section 9.6. Their point of departure from these models is the

assumption that agents do not know the statistical laws governing the unobserved dividend growth—in previous models, expected dividend growth is unobserved, although the underlying laws of movement are unknown. This lack of knowledge adds complexity and leads to interesting predictions regarding the equity premium.

In the model, agents are unsure about the unobservable drift of dividend growth, yet they formulate conjectures regarding the existence of a "cloud" of models containing the truth. In detail, this cloud is placed around a reference model for the dynamics of aggregate dividends,

$$\frac{dD_t}{D_t} = \theta dt + \sigma_0 dZ_t, \tag{9.95}$$

where σ_0 is a constant, θ is the unobservable expected dividend growth, and Z_t is a standard Brownian motion. Eq. (9.95), the reference model, is referred to as by LTV as a "rough approximation to reality," for reasons explained below.[12]

Eq. (9.95) is an approximation to reality, as mentioned. That is, the agents acknowledge that they are dealing with a possibly misspecified model, which leads them to consider *multiple likelihoods* governing the statistical laws of the fundamentals. That is, even while they take Eq. (9.95) as a benchmark, they do not trust this benchmark. They actually assume that the true, unknown, model is contained in a "neighborhood" of this benchmark, i.e., it belongs to the following family of models:

$$\frac{dD_t}{D_t} = \left(\theta + h\left(\theta\right)\sigma_0\right) dt + \sigma_0 dZ_{ht}, \tag{9.96}$$

for another Brownian motion Z_{ht}. In this family of models, the "distorsion function" h generates deviations from the benchmark. It captures the idea that agents face Knightian uncertainty, in that they have limited knowledge regarding whether their reference model is correctly specified. Moreover, the agents assume that h satisfies

$$h^2\left(\theta\right) \leq \eta\left(\theta\right), \quad \text{for all } \theta, \tag{9.97}$$

and for some known function η. In words, Eq. (9.96) contains model specifications that are statistically close to the reference model (9.95): they are so close that it is actually difficult to distinguish them statistically. Naturally, in the absence of ambiguity, one has that $\eta \equiv 0$, such that this model would collapse to those seen in Chapter 8.

Ambiguity, as described until now, is a source of second-order uncertainty, i.e., one that adds uncertainty (the agents' acknowledgement of dealing with model misspecification) to an already unobserved process (the drift of dividend growth), which they may wish to learn about. The crucial point is how the agents behave vis-à-vis this added uncertainty. The behavioral assumption is that they *fear* model misspecification, and choose consumption and portfolio policies (c_t, π_t) that maximize their worst-case scenario welfare, i.e., their lifetime utility arising when a malevolent Nature chooses the worst possible model in (9.96) for them:

$$\max_{(c_t,\pi_t)_{t\in(0,\infty)}} \inf_h E^h \left[\int_0^\infty e^{-\rho t} u\left(c_t\right) dt\right], \tag{9.98}$$

subject to the usual wealth constraints, and where E^h is the expectation taken under the statistical laws in Eq. (9.96) arising with a distortion function equals to h, ρ is the instantaneous

[12] LTV assume that the agents also observe additional signals on θ, which are not considered in this section to simplify the presentation of the model.

discounting rate, and $u(c)$ is the instantaneous utility with CRRA equal to γ. This formulation of the agent's problem has the same flavor as the max-min problem in the static market of Section 9.7.3.

Regarding the unobserved drift in the reference model, θ, we may assume that θ is a continuous process or that it is a Markov chain on set of finite values, as explained in Chapter 8 (Section 8.5.4). One simplifying assumption is that θ takes a finite number of values, similarly as in the introductory example of learning of Chapter 8 (see Section 8.5.4.2).[13] Moreover, we assume that θ could only take two values, $\underline{\theta}$ and $\bar{\theta}$, and denote the family of drifts of the dividend growth with $\theta^h = \theta + h(\theta)\sigma_0$. As for the agents' inference process, let the conditional probability $\pi_t \equiv \Pr(\theta^h = \bar{\theta}^h \Big| \mathbb{F}_t)$, where $\mathbb{F}_t$ denotes the information set availaible to agents at time t. Then, by results in Chapter 8, we have that

$$d\pi_t = \frac{\bar{\theta}^h - \underline{\theta}^h}{\sigma_0}\pi_t(1-\pi_t)\,dW_{ht}, \quad \pi_0 \equiv p,$$

where $dW_{ht} = \sigma_0^{-1}\left(\frac{dD_t}{D_t} - E^h\left(\theta^h \middle| (D_s)_{s\leq t}\right)dt\right)$.

Note that, by results reviewed in Chapter 8 (see Section 8.5.4), π_t will converge to one or to zero according to whether $\theta^h = \bar{\theta}^h$ or $\theta^h = \underline{\theta}^h$. That is, in this market, the agent will eventually *learn about the model quality*, so to speak (i.e., high or low growth), although he will always be uncertain about the exact parameter values of the model.

An equilibrium is one in which (i) a representative agent solves the optimization problem (9.98) and (ii) his optimal consumption equals dividends, $c_t = D_t$ for all t. To solve the optimization problem, and in analogy with the static maxmin model of Section 9.7.3, the agent first determines the infimum in (9.98), while taking as given the consumption and portfolio choices in the outer optimization problem. Given the thusly determined functions, the agent solves for the outer portfolio policies while imposing the equilibrium condition that $c_t = D_t$. Given this equilibrium condition, we have that the infimum is attained with

$$\hat{h}(\cdot) \equiv \arg\inf_h E^h\left[\int_0^\infty e^{-\rho t}u(D_t)\,dt\right].$$

Let $\Xi(\theta) \equiv \left\{\theta^h : (9.97) \text{ holds}\right\}$ and assume that $\Xi(\underline{\theta}) \cap \Xi\left(\bar{\theta}\right) = \emptyset$. Note that because the drift of D_t in (9.96) is increasing in $h(\cdot)$, then, by a comparison theorem (e.g., Karatzas and Shreve (1991, p. 291-295)), the previous expectation is increasing in $h(\cdot)$, such that[14]

$$\hat{h}(\theta) = -\sqrt{\eta(\theta)}. \tag{9.99}$$

In other words, asset prices are now evaluated *as if* the aggregate dividends dynamics were fully observed, but with "left-tilted" bounds to growth,

$$\begin{cases} \dfrac{dD_t}{D_t} = \hat{g}_t dt + \sigma_0 dW_{ht} \\ d\hat{g}_t = \frac{1}{\sigma_0}(\hat{g}_t - \underline{\theta}^{\hat{h}})(\bar{\theta}^{\hat{h}} - \hat{g}_t)dW_{ht} \end{cases} \tag{9.100}$$

[13] Note, then, that with this assumption, the model does not generalize Veronesi (2000) model's of learning, in which the unobserved drift is a Markov chain (see Section 8.5 in Chapter 8). However, it captures the salient features brought up by Knightian uncertainty.

[14] Precisely, note that the value function in (9.98) that has to be extremized is $E^h\left[\int_0^\infty e^{-\rho t}u(D_t)\,dt\right] = \sum_{j=1}^2 \int_0^\infty \pi_i e^{-\rho t}E^h\left(u(D_t)|\,\theta_i\right)dt$, where $\theta_1 \equiv \underline{\theta}$, $\theta_2 \equiv \bar{\theta}$, $\pi_1 \equiv \pi$, $\pi_2 \equiv 1-\pi_2$. The expectation in the integral is increasing in the drift of D by a comparison theorem.

where $\hat{g}_t \equiv E^{\hat{h}}(\theta^{\hat{h}} \Big| (D_s)_{s\leq t}) = \pi_t \bar{\theta}^{\hat{h}} + (1-\pi_t)\,\underline{\theta}^{\hat{h}}$, such that, by results in Chapter 8 (Section 8.5.4), the equity premium, $\mathcal{E}_t^h$, the short-term rate, r_t, and volatility, Vol_t, are given by

$$\mathcal{E}_t^h = r_t + \gamma\sigma_0 \cdot \text{Vol}_t, \quad r_t = \rho + \gamma\hat{g}_t - \frac{1}{2}\gamma\left(\gamma+1\right)\sigma_0^2, \quad \text{Vol}_t = \sigma_0 + \frac{p'\left(\hat{g}_t\right)}{p\left(\hat{g}_t\right)} v\left(\hat{g}_t\right), \tag{9.101}$$

where $v\left(\cdot\right)$ denotes the diffusion coefficient of $\hat{g}_t$ in (9.100) and p is the price-dividend ratio.[15] By results given in Chapter 8 (see Section 8.4), the price-dividend ratio is an affine function of the dividend growth expected under the agent's worst-case scenario probability, $\hat{g}_t$. Appendix 3 shows that

$$p\left(\hat{g}_t\right) = \underbrace{\frac{\hat{g}_t - \underline{\theta}^{\hat{h}}}{\bar{\theta}^{\hat{h}} - \underline{\theta}^{\hat{h}}}}_{=\pi_t} \frac{1}{\rho - (1-\gamma)\left(\bar{\theta}^{\hat{h}} - \frac{1}{2}\gamma\sigma_0^2\right)} + \underbrace{\frac{\bar{\theta}^{\hat{h}} - \hat{g}_t}{\bar{\theta}^{\hat{h}} - \underline{\theta}^{\hat{h}}}}_{=1-\pi_t} \frac{1}{\rho - (1-\gamma)\left(\underline{\theta}^{\hat{h}} - \frac{1}{2}\gamma\sigma_0^2\right)}. \tag{9.102}$$

The price-dividend ratio is a weighted average (weighted with the posterior π_t) of the discounted lifetime expected dividends conditional upon the up ($\bar{\theta}$) and down ($\underline{\theta}$) states of the world, and under the worst-case drifts. Provided $\gamma < 1$, the price-dividend ratio is decreasing in the degree of ambiguity aversion, $\eta\left(\cdot\right)$.[16]

Note that the equity premium in (9.101) originates from the perspective of the worst-case probability in the family of models included in (9.96). To determine the equity premium under the reference model (9.95), note that $\hat{g}_t = g_t + \mathcal{A}_t\sigma_0$, where $g_t \equiv E(\theta|\left(D_s\right)_{s\leq t}) = \pi_t\bar{\theta} + (1-\pi_t)\,\underline{\theta}$ is the growth estimate in the absence of any distortions, and

$$\mathcal{A}_t \equiv \pi_t\hat{h}(\bar{\theta}) + (1-\pi_t)\,\hat{h}(\underline{\theta}),$$

is the average size of ambiguity aversion. Therefore, by Girsanov's theorem (see Chapter 5), we can define a probability P_o with Radon-Nikodym derivative against the worst-case probability P^h (i.e., the probability under which $(D, \hat{g})$ are solution to (9.100)),

$$\left.\frac{dP_o}{dP^h}\right|_{\mathbb{F}_t} = e^{-\frac{1}{2}\int_0^t \mathcal{A}_u^2 du - \int_0^t \mathcal{A}_t dW_{\hat{h}u}},$$

such that $W_t = W_{\hat{h}t} + \int_0^t \mathcal{A}_u du$ is a Brownian motion under P^r. Then, under P^h, we have that

$$\begin{cases} \dfrac{dD_t}{D_t} = \underbrace{(\hat{g}_t - \mathcal{A}_t\sigma_0)}_{\equiv g_t} dt + \sigma_0 dW_t \\ d\hat{g}_t = -\mathcal{A}_t v(\hat{g}_t)dt + v(\hat{g}_t)dW_t \end{cases}$$

Therefore, the equity premium under the reference model (9.95) is

$$\mathcal{E}_t = \mathcal{E}_t^h - \text{Vol}_t \cdot \mathcal{A}_t, \tag{9.103}$$

[15] Note that the probability we are using while defining these statistics is that under the distorted beliefs: the agent prices the asset through $\hat{h}\left(\theta\right)$, as explained, and the statistics in (9.101) all originate from the pricing function $p\left(\cdot\right)$ given below (see Eq. (9.102). The expression for the equity premium under the reference model is given in Eq. (9.103) below.

[16] The assumption that $\gamma < 1$ is standard. The sensitivity of the short-term rate to growth is the agent's CRRA, γ (see (9.101)). Therefore, if $\gamma > 1$, growth changes would make discount rates change even more, such that the price-dividend ratio would be decreasing in growth and, hence, increasing in ambiguity aversion. Non-expected utility would fix this seemingly counterintuitive issue while allowing the elasticity of intertemporal substitution to be disentangled from relative risk aversion (see Section 9.2).

where $\mathcal{E}_t^h$ is the worst-case scenario equity premium in (9.100). Note that, by Eq. (9.99), $\mathcal{A}_t < 0$, such that ambiguity aversion contributes positively to the thusly defined equity premium.[17]

The rationale behind the definition of $\mathcal{E}_t$ in (9.103) relies on the following *interpretation* of the reference model. The ambiguity averse agent prices the asset at the worst-case scenario, yet Nature draws aggregate dividends according to Eqs. (9.95). Naturally, the agent is unaware of this statistical law, and places a band around θ, a band that could be symmetric or asymmetric, reflecting the modeling assumption that the agent only knows that for each regime (up or down), the expected dividend growth belongs to a certain band.[18]

It is easy to see that if the band is symmetric, with, say, $\eta\left(\underline{\theta}\right) = \eta\left(\bar{\theta}\right) \equiv \hat{\eta}$, the model predicts that the price-dividend ratio is the same as that in an economy without ambiguity, but a higher time preference rate (see Eq. (9A.43) in Appendix 3). However, this isomorphism breaks down once ambiguity aversion is state-dependent, $\eta\left(\underline{\theta}\right) \neq \eta\left(\bar{\theta}\right)$. Furthermore, even with homogeneous ambiguity aversion, the assumptions underlying the model leads to a sizeable equity premium under the reference probability, $\mathcal{E}_t - r_t$ in (9.103), a premium that cannot arise in an economy without ambiguity.

The left panel of Figure 9.2 plots the equity premium predicted by the model, with parameters fixed at the values given in the legend of Figure 9.2. Note that the $\mathcal{E}_t^h$ component is small, whereas the ambiguity premium $\text{Vol}_t \cdot \mathcal{A}_t$ is able to boost the premium at the levels we need to rationalize the empirical evidence reviewed in Chapter 8. Note, also, from the right panel of Figure 9.2., that return volatility, Vol_t, has the right order of magnitude in the calibrated model. Finally, the price-dividend ratio, not shown, ranges from about 25 to 38.

Note that the equity premium is inverse-U shaped against $\hat{g}_t$, a property arising mainly because return volatility is inverse-U shaped. The origins of this property are clarified in Chapter 8 (Section 8.5.4). Models with multiple states (such as those considered by LTV) and mean-reverting behavior (ensured while assuming θ is a Markov chain) are natural candidates to make volatility and equity premiums visit their descending parts more often than their ascending parts, thereby leading to a countercyclical behavior.

[17] Eq. (9.103) follows by Eq. (8.26) in Chapter 8, and by the change in probability from P^h to P_o.

[18] Precisely, the band is $[\underline{\theta} - \sqrt{\eta\left(\underline{\theta}\right)}\sigma_0, \underline{\theta} + \sqrt{\eta\left(\underline{\theta}\right)}\sigma_0]$ in the down regime, $\theta = \underline{\theta}$, and $[\bar{\theta} - \sqrt{\eta\left(\bar{\theta}\right)}\sigma_0, \bar{\theta} + \sqrt{\eta\left(\bar{\theta}\right)}\sigma_0]$ in the up regime, $\theta = \bar{\theta}$.

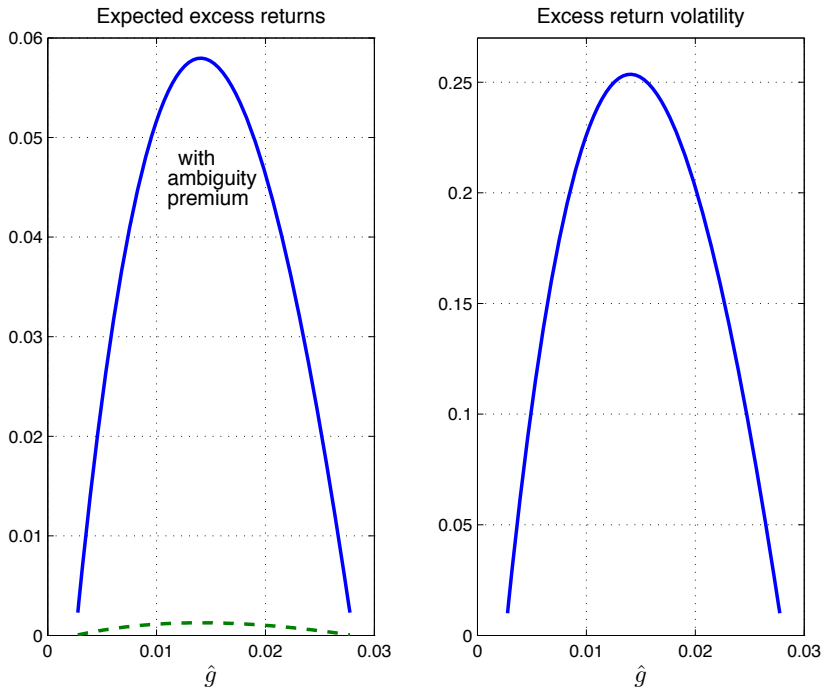


FIGURE 9.2. Left panel: The solid line depicts the equity premium under the reference probability, $\mathcal{E}_t - r_t$ in Eq. (9.103), as a function of the expected growth, $\hat{g}_t$; the dashed line is the equity premium under the worst-case probability, $\mathcal{E}_t^h - r_t$ in Eq. (9.101). Right panel: return volatility, Vol_t in Eq. (9.101). In both panels, parameter values are $\underline{\theta} = 0.005$, $\bar{\theta} = 0.03$, $\sigma_0 = 0.01$, $\gamma = \frac{1}{2}$, $\rho = 0.04$, and, finally, $\eta(\underline{\theta}) = \eta(\bar{\theta}) = \hat{\eta} = 0.05$.

9.8 Production

Consider an economy with one representative firm producing one single good, as in Section 3.4.1.2 of Chapter 3, and paying off a dividend $D(K_t, I_t)$ in each period t, expressed as a function of capital K_t and investment I_t, with partial with respect to capital K_t equal to $D_K(K_t, I_t)$:

$$\begin{aligned} D(K_t, I_t) &\equiv \tilde{y}(K_t, N(K_t)) - w_t N(K_t) - p_t I_t - \phi\left(\frac{I_t}{K_t}\right) K_t \\ D_K(K_t, I_t) &\equiv \tilde{y}_K(K_t, N(K_t)) - \frac{\partial}{\partial K}\left(\phi\left(\frac{I_t}{K_t}\right) K_t\right) \end{aligned}$$

Remember, Tobin's marginal q and average q are the same, by Theorem 3.2, meaning that the stock market value of the firm, $V(K_t)$, coincides with the value of installed capital, $V(K_t) = q_t K_{t+1}$, where q_t collapses to Tobin's q, once we fix the price of uninstalled capital to one, $p_t \equiv 1$, which is the case as soon as the firm produces uninstalled capital, simply. A few calculations

allow us to define equity returns in this economy. First, we note that:

$$\begin{aligned} V(K_t) &= q_t K_{t+1} \\ &= E_t\left[m_{t+1}\left(D_K\left(K_{t+1}, I_{t+1}\right) K_{t+1} + (1-\delta) q_{t+1} K_{t+1}\right)\right] \\ &= E_t\left[m_{t+1}\left(D_K\left(K_{t+1}, I_{t+1}\right) K_{t+1} + q_{t+1}\left(K_{t+2} - I_{t+1}\right)\right)\right] \\ &= E_t\left[m_{t+1}\left(D_K\left(K_{t+1}, I_{t+1}\right) K_{t+1} - q_{t+1} I_{t+1} + V\left(K_{t+1}\right)\right)\right] \\ &= E_t\left[m_{t+1}\left(D\left(K_{t+1}, I_{t+1}\right) + V\left(K_{t+1}\right)\right)\right], \end{aligned}$$

where the second line follows by the q theory, as developed in Chapter 3, the third and fourth lines by the law of capital accumulation, and the expression for $V(K_{t+1})$, the fifth line by the condition $q_{t+1} = -D_I(K_{t+1}, I_{t+1})$, and the homogeneity of the function D. Therefore, equity returns are:

$$\begin{aligned} \tilde{R}_{t+1} - 1 &\equiv \frac{D\left(K_{t+1}, I_{t+1}\right) + V\left(K_{t+1}\right) - V\left(K_t\right)}{V\left(K_t\right)} \\ &= \frac{D_K\left(K_{t+1}, I_{t+1}\right) K_{t+1} - q_{t+1} I_{t+1} + V\left(K_{t+1}\right) - V\left(K_t\right)}{V\left(K_t\right)}. \end{aligned}$$

In the absence of adjustment costs, $\phi \equiv 0$, Tobin's q collapses to one, $q_t = 1$, such that the capital gains, $V(K_{t+1})/V(K_t) - 1 = -\delta + I_{t+1}/K_{t+1}$, bringing equity returns to:

$$\tilde{R}_{t+1} - 1 = \tilde{y}_K\left(K_t, N\left(K_t\right)\right) - \delta.$$

To match the volatility of equity returns, a model without adjustment costs would require a counterfactually large volatility of the marginal product of capital. Therefore, not only are adjustment costs needed to rationalize the existence of time-varying market-to-book ratios. Adjustment costs would have the potential to boost return volatility. But then, the equity premium puzzle can only be exacerbated in a setting without adjustment costs. Note, indeed, that by the usual representation of the equity premium in Section 7.5 of Chapter 7,

$$E_t\left(r_{t+1}^e\right) = -\mathrm{corr}_t(m_{t+1}, r_{t+1}^e) \cdot \frac{\mathrm{Std}_t\left(m_{t+1}\right)}{E_t\left(m_{t+1}\right)} \cdot \mathrm{Std}_t\left(r_{t+1}^e\right),$$

where r^e denotes the equity return in excess of the risk-free rate. Unless the excess returns predicted by the model co-vary substantially, and negatively, with the stochastic discount factor, the equity premium can only be small, when $\mathrm{Std}_t\left(r_{t+1}^e\right)$ is very small. One route to inflate the equity premium might seem to be one where risk-aversion is increased. However, in equilibrium, equity returns obviously relate to consumption, and in models with production, consumption smoothing may make the equity premium puzzle worsen: as originally pointed out by Rouwenhorst (1995), if consumption is endogenous, it becomes smoother as risk-aversion increases, thereby making $\mathrm{Std}_t\left(r_{t+1}^e\right)$ smaller, in equilibrium.

The main issue with the neoclassical model is that capital supply is perfectly elastic, such that the price of capital and, hence, capital gains, are roughly constant, consistently with the previous arguments. As Jermann (1998) and Boldrin, Christiano and Fisher (2001) note, we need to introduce some sort of hindrance to the adjustment of capital supply to shocks. For example, Jermann (1998), assumes the presence of adjustment costs. Instead, Boldrin, Christiano and Fisher assume, among other things, that investment decisions can be thought to be determined prior to the realization of the shocks. Both Jermann (1998) and Boldrin, Christiano and Fisher (2001) consider economies with habit persistence anyway, which allows them to generate variability in the demand for capital and, hence, boost price volatility.

[In progress]

9.9 Government spending and asset prices

[In progress]

9.10 Leverage and volatility

Is firms' leverage responsible for a sustained stock volatility? Can leverage explain countercyclical stock volatility? We already know, from the previous chapter, that ex-post stock returns are high in good times, whence stock volatility is negatively related to ex-post returns. According to the *leverage effect* hypothesis, this negative relation arises because after a negative shock hits a stock price, the debt/equity ratio increases and a s result, the firm becomes riskier, leading to an increase in the stock volatility. Empirically, it is often argued, the leverage effect is too weak. Most of the contributions to these issues are empirical (e.g., Black, 1976; Christie, 1982; Schwert, 1989a,b; Nelson, 1991). Naturally, another possibility is that stock volatility and returns are negatively related for reasons unrelated to the leverage effect. For example, stock volatility can be countercyclical because agents' preferences and beliefs, combined with macroeconomic conditions, lead precisely to this property, as in the models discussed in Chapter 7, and in the previous sections.

In this section, we explore an additional explanation, namely that countercyclical volatility arises as a result of a combined effect of the properties of the previous models, and leverage. A difficulty is that in many empirical studies, tests of the leverage effect hypothesis are performed without regard to a well specified economic model. Gallmeyer, Aydemir, Hollifield (2007) show that the reasoning underlying this hypothesis can be made rigorous. They formulate a general equilibrium model with levered firms, which they realistically calibrate, to disentangle leverage effects from "real" effects such as habit formation. They make use of a stochastic discount factor known to price assets fairly well, and conclude that leverage effects do indeed have little effects in general equilibrium. This section develops a variant of their model, which has the mere merit to admit a closed-form solution.

9.10.1 Primitives

We consider an endowment economy, and denote endowment at time t with δ_t, assuming it is a Geometric Brownian motion with parameters g_0 and σ_0. The reason we denote output with δ, rather than the usual D, is that we assume a representative firm issues debt, denoted with D, such that the value of the firm is, by Modigliani-Miller (see Chapter 3), $V_t \equiv E_t + D_t$, where E_t is the value of equity at time t. Let T denote debt maturity. The payoffs of the firm are such that $\delta_t = \delta_{Et} + \delta_{Dt}$, with obvious notation. We assume a representative agent has habit formation preferences, and to obtain closed-form solutions for the asset prices, we make reference to the Menzly, Santos and Veronesi (2004) economy in Section 8.5.4 of Chapter 8. We denote the equilibrium surplus consumption ratio with $s_t = \frac{\delta_t - x_t}{\delta_t}$, where, as explained extensively in Chapter 8, s is solution to,

$$d\left(\frac{1}{s_t}\right) = \beta\left(\frac{1}{\bar{s}} - \frac{1}{s_t}\right)dt - \alpha\left(\frac{1}{s_t} - \frac{1}{\upsilon}\right)\sigma_0 dW_t,$$

and β, $\bar{s}$, α and υ are parameters. In this economy, Sharpe ratios are countercyclical, being equal to $\lambda(s) = \sigma_0\left(1 + \alpha\left(1 - \frac{s}{\upsilon}\right)\right)$, as mentioned in Section 8.5.4. We assume debt services are $\delta_D = q\delta$, for some $q \in (0,1)$, and set the benchmark for debt maturity to $T = 10$ years.

We now show that a calibration of the model leads to the following results: (i) the price of debt is procyclical; (ii) return volatility is countercyclical; (iii) the leverage ratio is countercyclical; and finally, (iv) the contribution of leverage to equity returns volatility is quantitatively limited.

9.10.2 Equity volatility: a decomposition formula

From Chapter 8, we know that the price-dividend ratio for the aggregate consumption claim is $p(s) \equiv a + bs$, for two constants a and b, which we shall give below again. It is easy to show that the debt value is,

$$\frac{D_t}{\delta_t} = q\left(a_{T-t} + b_{T-t}s_t\right), \quad a_\tau = \frac{1 - e^{-(\rho+\beta)\tau}}{\rho+\beta}, \quad b_\tau = \frac{\beta\left(1 - e^{-\rho\tau}\right) + \rho\left(e^{-(\rho+\beta)\tau} - e^{-\rho\tau}\right)}{\rho\left(\rho+\beta\right)\bar{s}},$$

where $a = \lim_{T\to\infty} a_\tau$ and $b = \lim_{T\to\infty} b_\tau$. Equity volatility is,

$$\text{Vol}\left(\frac{dE}{E}\right) = \sigma_V + \left(\sigma_V - \sigma_D\right)\frac{D}{E}, \quad \sigma_V = \sigma_0 + \frac{b}{a+bs}\text{Vol}\left(ds\right), \; \sigma_D = \sigma_0 + \frac{b_{T-t}}{a_{T-t}+b_{T-t}s}\text{Vol}\left(ds\right),$$

where $\text{Vol}\left(ds\right) = \alpha\sigma_0 s\left(1 - \frac{1}{v}s\right)$, such that

$$\begin{aligned}\underbrace{\text{Vol}\left(\frac{dE_t}{E_t}\right)}_{\approx 0.15} &= \underbrace{\sigma_0}_{=0.01} + \underbrace{\frac{b}{a+bs_t}}_{=\text{ endog. P/D fluct.}\approx 26.31} \cdot \underbrace{\text{Vol}\left(ds_t\right)}_{\approx\, 5\cdot 10^{-3}} \\ &+ \underbrace{\left(\frac{b}{a+bs_t} - \frac{b_T}{a_T+b_Ts_t}\right)}_{=\text{ leverage multiplier}\approx 11.08} \cdot \underbrace{\text{Vol}\left(ds_t\right)}_{\approx\, 5\cdot 10^{-3}} \cdot \underbrace{\frac{D_t}{E_t}}_{\approx\, 0.24},\end{aligned} \tag{9.104}$$

where we have indicated the approximate average values taken by the variables of interest, and obtained by calibrating the model with the values of Table 9.2 on Section 9.9 below. Note, also, that the leverage ratio, $\frac{D_t}{E_t}$, is endogenous and equal to,

$$\frac{D_t}{E_t} = \frac{\left(a_{T-t} + b_{T-t}s_t\right)q}{a + bs_t - \left(a_{T-t} + b_{T-t}s_t\right)q}.$$

In other words, we only "see" what happens to $\frac{D_t}{E_t}$ and $\text{Vol}\left(\frac{dE_t}{E_t}\right)$, as the surplus s_t changes. As the numerical values in Eq. (9.104) show, much of the action in this model derives from the large swings in the price-dividend ratio, $\frac{p'(s_t)}{p(s_t)} = \frac{b}{a+bs_t}$. What is the statistical relation between the leverage ratio $\frac{D}{E}$ and return volatility that the model predicts? Figure 9.3 depicts values of the leverage ratio and volatility consistent with the model. Note that Figure 9.3. does not depict a causal relation, as leverage and equity volatility are both driven by the same state

variable, the surplus consumption ratio.[19]

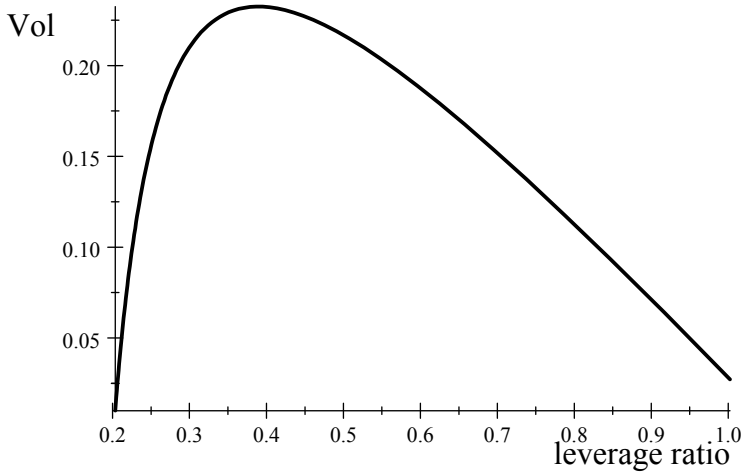


FIGURE 9.3. Leverage and equity volatility: a "naked" eye view.

9.10.3 Bankruptcy

The previous model has no role for bankruptcy, which plays an obviously fundamental role, as developed in great detail in in Chapter 14. Let us consider bankruptcy in a simple setting. Consider a two date economy, and suppose that the value of the firm in one year is, $\tilde{V}$, which equals $V_{\text{bad}} <$ Nominal debt, with probability p, and $V_{\text{good}} >$ Nominal debt, with probability $1-p$. We assume risk-neutrality. and that are no bankruptcy costs. Let $\tilde{R} = \frac{\tilde{S}_1 - S_0}{S_0}$ be the equity return, where $\tilde{S}_1$ is the equity value at the second period. Then, we have that $\text{vol}(\tilde{R}) = \sqrt{\frac{p}{1-p}}$. For example, if $p = 2\%$, then $\text{vol}(\tilde{R}) = 14\%$!

9.11 Multiple trees and the cross-section of asset returns

Menzly, Santos and Veronesi (2004), Cochrane, Longstaff and Santa-Clara (2008), Pavlova and Rigobon (2008), Martin (2011).

9.12 The term-structure of interest rates

What are the term-structure implications of the main paradigms considered so far? Consider the habit formation model introduced by Campbell and Cochrane (1999). While Campbell and Cochrane consider an economy where interest rates are constant, in their working paper they allow the short-term rate to be time-varying, and as explained in Appendix 5 of Chapter 8, set equal to:

$$r(s) = \delta + \gamma\left(g_0 - \frac{1}{2}\sigma_0^2\right) - \frac{1}{2}\left(\gamma(1-\phi) - b\right) + b\left(\bar{s} - \ln s\right), \tag{9.105}$$

[19] Does debt maturity lead to a greater contribution of leverage to volatility? It is not the case. Given the model parameter values, the effects of debt maturity on leverage can be shown to be quite limited.

where s is the surplus consumption ratio, b is a constant, and all the remaining parameters are as in Section 8.5.2 of Chapter 8. Wachter (2006) analyzes the term-structure implications of this model in detail, both real and nominal, within an environment with time-varying expected inflation.

Note, the constant b does not depend on anything relating to the agents' preferences. Its mere role is to make interest rates time-varying. How to ensure that Eq. (9.105) is consistent with optimizing behavior? As explained in Chapter 8, the short-term rate depends on the sensitivity of habit to consumption shocks, a function of s, $l(s)$, through an effect due to precautionary savings: the higher this sensitivity, the higher the volatility of habit and, hence, the propensity to save, which drives interest rates down. This sensitivity $l(s)$ is "free," in that it is not restricted by the theory—Campbell and Cochrane simply guide us with heuristic considerations leading to it. One of these considerations is that the short-term rate also relates to habit, due to intertemporal substitution effects, and negatively, due to mean-reversion. Campbell and Cochrane choose $l(s)$ such that intertemporal substitution effects exactly offset precautionary savings, thereby making the short-term rate constant or, at most, affine in the log surplus consumption ratio, as in Eq. (9.105). Naturally, the sensitivity, $l(s)$, is a function of b, once this reverse engineering has unfold, as shown in Appendix 5 of Chapter 8.

The question arises as to which sign we should expect from the parameter b, empirically. Are real interest rates countercyclical? They *are*. It is somehow puzzling, from the perspective of the basic production economies analyzed in Chapter 3, where real interest rates are procyclical, being positively related to the marginal product of capital and, hence, to productivity shocks. However, economies with habit formation might be capable of generating countercyclical real rates, due to intertemporal substitution effects—It is the case, for example, for the models with frictions in the adjustment of capital supply to shocks of Boldrin, Christiano and Fisher (2001). In endowment economies and habit formation, countercyclical real rates are, then, quite likely to arise. Consider, for example, the Menzly, Santos and Veronesi (2004) model of external habit formation presented in Section 8.5.4. we remind that this model predicts that the short-term rate is:

$$r(s_t) = \delta + g_0 - \sigma_0^2 + \beta\left(1 - \frac{s_t}{\bar{s}}\right) - \sigma_0^2\alpha\left(1 - \frac{s_t}{\upsilon}\right). \tag{9.106}$$

Figure 9.4 depicts the short-term rate as a function of s, obtained using the parameter values in Table 9.2, which are similar to those used by Menzly, Santos and Veronesi.

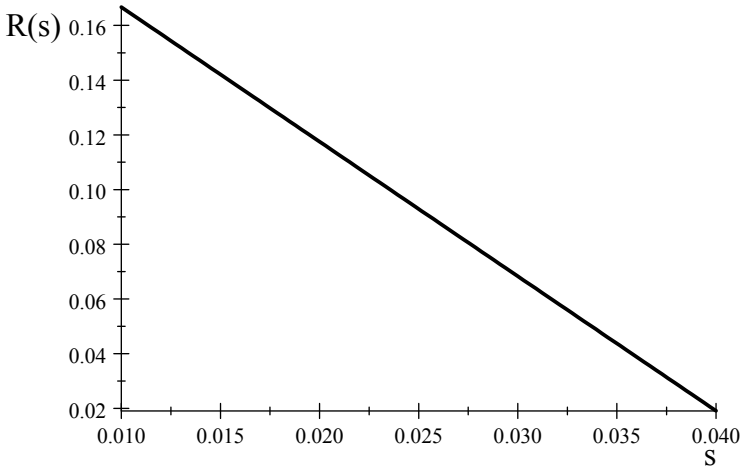


FIGURE 9.4. The short-term rate predicted by Menzly, Santos and Veronesi (2004) model of external habit formation, with parameter values as in Table 9.2.

g_0	σ_0	δ	β	$\bar{s}$	α	υ	q
0.03	0.01	0.04	0.15	0.03	40	0.05	0.60

TABLE 9.2. Parameter values utilized for the Menzly, Santos and Veronesi (2004) model of external habit formation.

The fourth term of Eq. (9.106) reflects intertemporal substitution effects, and is the dominating term, leading to countercyclical interest rates, due to the mean reversion in the surplus consumption ratio, and similarly as in the Campbell and Cochrane model, as explained in Section 8.5.2 of Chapter 8. Finally, the catching-up model of Chan and Kogan (2002) reviewed in Section 9.3 leads to the same prediction: real interest rates are countercyclical.

[In progress]

9.13 Prices, quantities and the separation hypothesis

One compelling lesson from this chapter is that to address the asset pricing puzzles brought about by the neoclassical model, we would need a substantial re-vamp of the standard paradigms underlying dynamic macroeconomic theory—namely, a re-vamp of the basic version of the real business cycle theory reviewed in Chapter 3. For example, we would need to introduce adjustment costs, habit formation, or restricted stock market participation. How is it, then, that while attempting to explain quantity dynamics, macroeconomists would simply ignore the advances made by financial economists? Tallarini (2000) considers a different possibility, a real business cycles model in which a representative agent has non-expected utility, as in Section 9.2.

9.13.1 A closed-form expression for non-expected utility

To illustrate these old and yet important ideas, let us determine the stochastic discounting factor in a model with non-expected utility and without production; later, we shall describe the main implications of a production-based economy. In Appendix 1 (see Eq. (9A.5)), we explain that with non-expected utility, an alternative representation of m_{t+1} to Eq. (9.8) is:

$$m_{t+1} = \beta \left(\frac{c_{t+1}}{c_t} \right)^{\rho-1} \left(\frac{E_t \left(V_{t+1} \right)}{V_{t+1}} \right)^{\frac{\rho}{1-\gamma}-1}, \quad V_{t+1} = \frac{v_{t+1}^{1-\gamma}}{1-\gamma}, \tag{9.107}$$

where v_t denotes the certainty equivalent continuation utility expressed in consumption units, and is solution to Eq. (9.3). Next, define $z_t = (1-\beta)^{\frac{1}{\rho}} v_t$, which is obviously ordinally equivalent to v_t and satisfies

$$z_t = \left((1-\beta) c_t^{\rho} + \beta \left(E_t(z_{t+1}^{1-\gamma}) \right)^{\frac{\rho}{1-\gamma}} \right)^{\frac{1}{\rho}}.$$

As the intertemporal elasticity of substitution gets close to one, $\rho \to 0$, $z_t = c_t^{1-\beta} (E_t(z_{t+1}^{1-\gamma}))^{\frac{\beta}{1-\gamma}}$. Define, then, $U_t : z_t \equiv e^{(1-\beta)U_t}$, i.e.,

$$U_t = \ln c_t + \frac{\beta}{\sigma} \ln E_t(e^{\sigma U_{t+1}}), \quad \sigma \equiv (1-\gamma)(1-\beta). \tag{9.108}$$

Next, assume that log consumption is a random walk with drift,

$$g_{t+1} = \ln \frac{c_{t+1}}{c_t} = \mu_g - \frac{1}{2}\sigma_c^2 + \epsilon_{t+1}, \quad \epsilon_{t+1} \sim N\left(0, \sigma_c^2\right). \tag{9.109}$$

for some constant μ_g. Note that Eq. (9.109) is a special case of the dynamics of consumption in the Bansal and Yaron (2004) model in Section 9.2—it does not include the small persistent component x_t of Eq. (9.19). Guessing that $U_t = a + b \ln c_t$, and solving for the undetermined coefficients a and b using Eq. (9.108), delivers $a = \frac{2\beta\mu_g(1-\beta)-\beta(1-\beta-\sigma)\sigma_c^2}{2(1-\beta)^3}$ and $b = \frac{1}{1-\beta}$, such that the stochastic discounting factor in Eq. (9.107) for $\rho = 0$ is

$$m_{t+1} = \beta \frac{c_t}{c_{t+1}} \frac{V_{t+1}}{E_t\left(V_{t+1}\right)} = \beta \frac{c_t}{c_{t+1}} \frac{e^{\sigma U_{t+1}}}{E_t\left(e^{\sigma U_{t+1}}\right)} = \beta e^{-\mu_g + \frac{1}{2}(2-\gamma)\gamma\sigma_c^2} e^{-\gamma\epsilon_{t+1}}.$$

We have,

$$E_t\left(m_{t+1}\right) = \beta e^{-\mu_g + \gamma\sigma_c^2}, \quad \sqrt{var_t\left(m_{t+1}\right)} = E_t\left(m_{t+1}\right)\sqrt{e^{\gamma^2\sigma_c^2} - 1}. \tag{9.110}$$

By increasing γ, the volatility of the stochastic discounting factor increases although, then, $E_t\left(m_{t+1}\right)$ remains substantially flat. Once again, this property is what makes non-expected utility address the interest rate puzzle (see Section 9.2). As also discussed in Section 9.2, non-expected utility does not necessarily imply a resolution of the equity premium puzzle: Eqs. (9.110) clearly indicate that large values of γ are needed to inflate the equity premium to empirically plausible levels. Alternatively, we need additional sources of variation in the primitives of the economy—for example, after Tallarini, long-run risks are proposed by Bansal and Yaron (2004).

9.13.2 Preferences for robustness

Hansen and Sargent (2008, page 317) argue that this model can be understood one leading to a resolution of the equity premium puzzle even without assuming long-run risks, by simply re-interpreting the parameter $-\sigma^{-1}$ in Eq. (9.108) as a multiplier for an ambiguity averse individual with multiplier preferences.

[In progress]

9.13.3 Irrelevance

The main point of Tallarini is an irrelevance result: we can understand the dynamics of asset markets independently of those of real economic aggregates, due to the linearity of preferences resulting from the assumption that intertemporal elasticity of substitution is one, $\rho = 0$. To extend the previous model to one with real aggregates, Tallarini considers the following extension to Eq. (9.108),

$$U_t = \ln c_t + \varphi \ln L_t + \frac{\beta}{\sigma'} \ln E_t\left(e^{\sigma' U_{t+1}}\right), \tag{9.111}$$

where the parameter $\varphi > 0$ now loads leisure, L_t, and is such that $\sigma' = \sigma/\left(1+\varphi\right)$. Tallarini does not consider adjustment costs, and yet his model can explain the equity premium, through a simple increase in the risk aversion parameter σ', while maintaining intertemporal substitution constant, i.e. by keeping on assuming log consumption in the right hand side of Eq. (9.111). Interestingly, raising risk-aversion does not affect the quantity dynamics macroeconomists are interested in, only intertemporal substitution might affect it. Naturally, there are many other

dimensions we should consider, to conclude on any model's prediction about asset prices. For example, Tallarini's assumption of no adjustment costs implies Tobin's q is one. Moreover, welfare calculations such as those in Lucas (19??) are likely to change, as Alvarez and Jermann (20??) demonstrate.

[In progress]

9.14 Endogenous risk and the financial accelerator doctrine

9.14.1 Procyclicality

In the models reviewed so far in this chapter, asset prices move in response to changes in the fundamentals: it is *as if* asset prices were "derivatives" written on a few key variables summarizing the state of the economy. Can capital market movements lead to feedbacks to the real economic activity? The Great Recession likely has its origins in the mortgage led credit crisis erupted in 2007 (see Chapter 14). Adverse developments in credit markets led financial intermediaries to losses and a diminished propensity to lend and, ultimately, a "credit crunch." In turn, the credit crunch exacerbated business cycle conditions, over a spiral.

These historical episodes are instances of "endogenous risk," or "procyclicality," by which a shock leads to adverse developments, which in turn reinforce the initial shock, creating a trend. Specifically, a turmoil in capital markets may lead to adverse economic developments, which reinforce the initial financial turmoil, over a spiral. Even a relatively small shock, such as the initial subprime mortgage losses in 2007, may quite adversely affect the business cycle. This thesis is known as "financial accelerator" hypothesis. Understanding asset prices within such a general perspective would allow us to explain economic trends in contexts as critical as those underlying the Great Recession. But defining the details of these mechanisms leads to alternative explanations.

Fisher (1933) was one of the earliest proponent of a theory of procyclicality. His explanations aimed at clarifying the origins of the Great Depression based on a debt-deflation spiral. In an economy with highly levered firms, such as that of the US during the 1930s, a negative productivity shock leads to bankruptcy of a fraction of these firms, which generates less investments and, hence, depresses aggregate demand and creates deflation. Debt (and its nominal rigidity) is the element that makes Fisher's explanations novel compared to neoclassical predictions. In particular, deflation boosts the real value of debt borne by firms, increasing the firms' burden and leading a higher fraction of these firms to default. This debt-deflation spiral results in a deterioration of the balance sheet of financial intermediaries, because banks obviously bleed money as their borrowers default. The result is a default contagion, one traveling from firms to financial intermediaries. Eventually, financial intermediation shrinks and the vicious feedback loop may go through a consistently long period.[20]

[20]This explanation of the Great Depression was challenged by Friedman and Schwartz (1963), who proposed a "monetary view" instead. The monetary view holds that the prolonged recession and the banking crises in the 1930s link to non-accommodating monetary policies. Friedman and Schwartz find a statistical relation between monetary policy and real macroeconomic aggregates in the U.S. (from the Civil War through 1960): an expansionary monetary policy is associated with an expansion of real aggregates. They find that this linkage is particularly strong in the 1930s, and suggest causality from monetary policy to real economic developments. According to Friedman and Schwartz, the only role banks may have possibly played in explaining the crisis was their contribution to the shrinkage in money supply through a lower money multiplier (see Section 14.6.4 for a definition of the money multiplier). In particular, the deflation of the early 1930s would not allow for any adjustment (a higher output following a higher *real* money supply), because real money supply did not substantially change due to a shrinkage in the nominal monetary aggretates determined by a lower multiplier (due to the banking crises). Friedman and Schwartz considered that the FED should have targeted a much higher monetary base to mitigate this issue. The monetary view was subsequently challenged by the "credit view" as well

This section surveys models that have this financial accelerator flavor, but focus on an amplification mechanism of a different nature. In bad times, the cost of external funds increases, typically due to agency problems. That is, due to asymmetric information, financial intermediaries agree on lending plans based on the collateral made available by borrowers. In bad times, then, financial intermediaries reduce their fundings as collateral values are also reduced in bad times. The ensuing lending shrinkage contributes to further depress real economic activity and, hence, assets and collateral values, feeding a vicious feedback loop.[21]

We begin with variants of the Kiyotaki and Moore (1997) model, in which small and temporary shocks generate large and persistent swings in asset prices and the business cycle. The mechanism is similar to the credit multiplier channel in Bernanke and Gertler (1989), although it does not rely on the presence of financial intermediaries. Bernanke, Gertler and Gilchrist (1999) provide an early survey of models with more emphasis on macroeconomics, presenting a unified view of how agency problems make funding opportunities depend on firms' collateral. Brunnermeier, Eisenbach and Sannikov (2012) survey more recent developments.

[In progress]

9.14.2 Premise: Asymmetric information, asset prices and business cycles

Explain succinctly the reasons we are interested in models with asymmetric information to explain how asset prices can extert feedback effects into business cycles. Furthermore, explain that the next chapter will be entirely devoted to study asset prices in markets with asymmetric information with many details.

[In progress]

9.14.3 Credit cycles

Kiyotaki and Moore (1997) provide a framework to think about business cycle fluctuations originating from and, in turn, influencing asset prices movements. Persistence and amplifications of shocks are the key predictions of this model. In fact, *persistence* of shocks can be understood without reference to asset prices: it could simply arise through the firms' financial constraints. Kiyotaki (1998) develops a model along these lines, which is analyzed in this section. The next section explains how introducing an asset allows a shock to be *amplified*.

There are two types of agents: (i) "productive" or "farmers" and (ii) "unproductive" or "gatherers." Farmers invest x_t into a linear technology, obtaining output y_{t+1},

$$y_{t+1} = \alpha x_t. \tag{9.112}$$

Gatherers have access to a less productive technology, in that when investing $x_{g,t}$, their output, $y_{g,t+1}$, satisfies

$$y_{g,t+1} = \gamma x_{g,t}, \tag{9.113}$$

where

$$1 < \gamma < \alpha. \tag{9.114}$$

as further developments described below as a part of financial accelerator mechanisms. Additional details regarding this fascinating part of the history of economic thought fall somehow beyond the scope of these lectures.

[21] Borio, Furfine and Lowe (2001) explain that the agents' misperception of risk might constitute an additional amplification mechanism. For example, the credit/GDP ratio might be procyclical because financial intermediaries under-estimate risk in good times, and over-estimate risk in bad, thereby lending too much in good times and too less in bad.

Productivity is random: in each period, farmers may become gatherers, and gatherers may become farmers. We shall develop more details regarding this assumption below. Intuitively, switching types is required to make sure both types survive.

All agents maximize the same intertemporal utility of consumption subject to their budget constraint. For example, the farmers' plan is

$$\max_{(c_{t+\tau})_{\tau=0}^{\infty}} E_t\left(\sum_{\tau=0}^{\infty}\beta^{\tau}\ln c_{t+\tau}\right), \quad \text{s.t.} \quad c_t + x_t = y_t + \frac{b_{t+1}}{R_t} - b_t, \tag{9.115}$$

where b_t denotes debt outstanding at time t and R_t is the gross interest rate on debt for the period from t to $t+1$. The budget constraint says that at time-t, consumption, c_t, and investments, x_t, are financed by output, y_t, and new issuance of debt, worth $\frac{b_{t+1}}{R_t}$, after repaying current debt b_t.

We are about to describe a mechanism by which gatherers lend to farmers as the latter "know how" to produce efficiently (i.e., they can make better money than while just investing at the short-term rate). This mechanism works due to a friction in the market for lending: while farmers are efficient, they might also default on their debt, but then, gatherers would only provide collateralized (that is, a limited amount of) loans. Thus, in the resulting economy, the gross interest rate $R_t = \gamma < \alpha$ can be sustained because farmers can only access to a limited amount of loans (due to the collateral constraint friction).[22]

Without such friction, the gross interest rate $R_t = \alpha$. Indeed, if $R_t > \alpha$, the farmers would never produce because the return on investment would be less than the real interest rate; and if $R_t < \alpha$, the farmers would be willing to borrow an infinite amount, which cannot be an equilibrium either. Note, then, that by Eq. (9.114), the interest rate $R_t = \alpha$ would discourage the gatherers from investing. In fact, gatherers are out of this economy once we assume their initial wealth is zero: because they do not produce, they have no resources to survive.

Let us analyze the details of this frictionless case, first. Define the farmers' net worth as $a_t \equiv y_t - b_t$. The flow-of-funds accounting in Eq. (9.115), the equilibrium condition $R_t = R \equiv \alpha$ and Eq. (9.112) yield

$$a_{t+1} = (a_t - c_t)\, R. \tag{9.116}$$

The farmers' plan is then as in (9.115), subject to the flow-of-funds constraint in Eq. (9.116). The solution to this program is well-known (see, e.g., Section 3.2 in Chapter 3).[23] Optimal consumption is $c_t = (1-\beta)\, a_t$. Moreover, gatherers do not participate in this economy. Therefore, in equilibrium, $b_t = 0$ for all t, such that by the constraint in (9.115), $x_t = \beta a_t$. Aggregate investments are $X_t \equiv \beta W_t$, where W_t denotes the aggregate net-worth, which equals aggregate output, Y_{t+1}, because aggregate borrowing is zero. To sum up,

$$W_{t+1} = Y_{t+1} = \alpha X_t = \alpha\beta W_t = \alpha\beta Y_t. \tag{9.117}$$

The growth of the economy is constant and equal to $\alpha\beta$. It is a very simple outcome, arising within a representative agent economy.[24] Note that this economy has farmers only because the

[22] A formal proof that $R_t = \gamma$ in this economy relies on the first order conditions of the gatherers problem (the equivalent to (9.115)) with respect to x_t and b_{t+1} and based on the production function in (9.113).

[23] Below, we shall assume that productivity shocks are a Markov chain in that farmers and gatherers may switch types. The value function of (9.115) is then $V(a_t) \equiv u(c_t) + \beta E_t(V(a_{t+1}))$, where E_t denotes the expectation taken at time t, and so the results in Section 3.2 of Chapter 3 can still be utilized in the context of this model.

[24] Moreover, Eq. (9.117) shows that classical results (reviewed in Chapter 3) hold, namely, that in steady state ($Y_{t+1} = Y_t$), the interest rate equals to the subjective discount rate $R = \alpha = \beta^{-1}$.

interest rate is too high to allow "penniless" gatherers to survive. Gatherers could survive in an economy in which the interest rate is lower than their marginal productivity, i.e., $R \leq \gamma$. However, we know this outcome cannot arise in absence of any frictions: farmers would find it optimal to borrow an infinite amount of funds and try to exploit this opportunity, which cannot be an equilibrium.

An economy with both farmers and gatherers needs to display some sort of hindrance in the borrowing opportunites opened to farmers, for example, a limit to the amount of borrowing. Kiyotaki and Moore (1997) propose the following framework, also used by Kiyotaki (1998). First, only *those who have initiated the investment* have the skills to produce the full output. Without those skills, one may only achieve to a fraction θ of the full output (even while having access to the initial producers' technology). Moreover, each producer can declare bankruptcy at any time after he starts produce. Anticipating this possibility of default, the creditors will not lend more than it can be obtained without the farmers' skills,

$$b_{t+1} \leq \theta y_{t+1}. \tag{9.118}$$

That is, should any farmer walk away, a gatherer could then expropriate his technology and produce the output fraction θy_{t+1}, such that we can interpret this fraction as the "collateral" value of the farmers' investments. In other words, creditors are "collaterizing" their loans.[25]

We search for an equilibrium in which $R = \gamma$ around the stationary state, and (9.118) holds as an equality. Below, we confirm that such an equilibrium exists, by just verifying that the amount of the gatherers' investments is bounded and bounded away from zero.

Note that (9.118) is the friction that prevents farmers from borrowing in an infinite amount even while the interest rate is less than their productivity. That is, farmers have access to "good" financing conditions but they cannot borrow an infinite amount as they are borrowing-constrained by (9.118). They will then borrow as much as they could, i.e.,

$$b_{t+1} = \theta y_{t+1} = \theta \alpha x_t. \tag{9.119}$$

The previous constraint imposes an upper limit to the farmers' investment, x_t. Using the budget constraint in (9.115), the equilibrium condition $R = \gamma$, and Eq. (9.119), it is

$$x_t = \frac{y_t - b_t - c_t}{1 - \frac{\theta\alpha}{\gamma}}. \tag{9.120}$$

Investments x_t are just mechanically derived from the budget constraints. The numerator in (9.120) equals the farmers' savings. It is the down-payment that is used to invest, x_t, over and above the amount borrowed, $\frac{b_{t+1}}{R}$. Furthermore, to ensure this investment plan is part of the equilibrium, it must be that $\theta\alpha < \gamma$: the collateralized return (i.e., the percentage of loanable funds to the investment size, $\frac{b_{t+1}}{x_t}$) cannot exceed the interest rate ($R = \gamma$) to ensure that the current borrowing is less than x_t. The flow-of-funds accounting is now

$$a_{t+1} \equiv y_{t+1} - b_{t+1} = (1-\theta)\,\alpha x_t = \frac{(1-\theta)\,\alpha}{1 - \frac{\theta\alpha}{\gamma}}\,(a_t - c_t)\,. \tag{9.121}$$

The program of the farmers in this economy is to maximize the intertemporal utility in (9.115) subject to (9.121). We solve for consumption c_t, which then determines x_t in (9.120).

[25] An alternative interpretation of the parameter θ is the fraction of skilled producers in contexts where lenders do not observe the lender's type.

The solution for consumption follows by the usual argument: for a log-utility maximizer, $c_t = (1-\beta)\, a_t$. Replacing into (9.120), and aggregating over all farmers leaves

$$\frac{X_t}{W_t} = \beta \frac{s_t}{\bar{s}}, \quad s_t \equiv \frac{A_t}{W_t}, \quad \bar{s} \equiv 1 - \frac{\theta\alpha}{\gamma}, \tag{9.122}$$

where capitalized letters denote aggregate variables.

Under the conjecture that $R = \gamma$, gatherers are indifferent as to whether to invest or to lend. However, we can determine their investments from equilibrium conditions. Indeed, note that their aggregate savings are equal to $\beta A_{g,t}$, with obvious notation. Hence, in equilibrium, aggregate savings satisfy $S_t \equiv \beta\,(A_t + A_{g,t}) \equiv \beta W_t$, where $W_t = Y_t + Y_{g,t}$ because aggregate borrowing is obviously zero. Therefore, the investment made by the gatherers can be determined in equilibrium, by equating aggregate savings to aggregate investments, viz

$$X_{g,t} = S_t - X_t = \beta W_t \left(1 - \frac{s_t}{\bar{s}}\right), \tag{9.123}$$

such that aggregate output satisfies: $W_{t+1} = Y_{t+1} + Y_{g,t+1} = \alpha X_t + \gamma X_{g,t}$, or,

$$\frac{W_{t+1}}{W_t} = \beta \left(\gamma + (\alpha - \gamma)\, \frac{s_t}{\bar{s}}\right). \tag{9.124}$$

The dynamics of growth are summarized by the state variable s_t defined in (9.123), which is the farmers' aggregate share of net-worth. Growth is increasing in s_t: the larger the relative size of productive agents, the higher the aggregate productivity. Note, also, that growth is lower than that in the unconstrained economy, $\alpha\beta$ (see Eq. (9.117)) if and only if $s_t < \bar{s}$. There are indeed conditions under which $s_t < \bar{s}$ in the neighborhood of the steady state. Furthermore, note that as soon as $s_t < \bar{s}$, the gatherers's investment $X_{g,t}$ is strictly positive and bounded by (9.123), confirming that $R = \gamma$ is an equilibrium. Finally, the economy collapses to that with a representative agent once that $\gamma = \alpha$.

Finally, we determine aggregate output. Output dynamics depends on how farmers and gatherers switch their types. Assume that there is a mass $n : 1$ of farmers to gatherers, and that productivity is a Markov chain, as in the following table

	U_{t+1}	P_{t+1}
U_t	$1 - n\delta$	$n\delta$
P_t	δ	$1 - \delta$

It is easy to verify that given the previous transition density, the mass $n : 1$ is constant over time.[26] Moreover, assume that productivity is persistent, in that $\Pr(U_{t+1}|\, U_t) > \Pr(P_{t+1}|\, U_t)$ and $\Pr(P_{t+1}|\, P_t) > \Pr(U_{t+1}|\, P_t)$, i.e.,

$$1 - \delta > n\delta. \tag{9.125}$$

That is, $\Pr(U_{t+1}|\, U_t) > \Pr(U_{t+1}|\, P_t)$. In words, a gatherer at time $t+1$ is more likely to have been a gatherer than a farmer at time t.

[26] Formally, a stationary density in this simple model is a 2×1 vector π satisfying $\pi^\top \mathbf{1}_2 = 1$, which is the solution to $\pi = \Pi\pi$, where Π denotes the transition density of the Markov chain. For example, the first condition is $\pi_1 = \Pr(U_{t+1}|\, U_t)\,\pi_1 + \Pr(U_{t+1}|\, P_t)\,\pi_2$, where π_1 denotes the stationary fraction of gatherers. The solution is $\pi_1 = \frac{1}{n+1}$, whence, $\frac{\pi_2}{\pi_1} = n$.

Let A_{t+1} be the aggregate net worth of the farmers. Then, the farmers' aggregate new worth at $t+1$ *after* the productivity shock, A_{t+1}, includes that of the time-t farmers who are still farmers (and have a mass equal to $1-\delta$) and that of the time-t gatherers who become farmers (and have mass equal to $n\delta$), viz

$$A_{t+1} = (1-\delta)(Y_{t+1} - B_{t+1}) + n\delta(Y_{g,t+1} - B_{g,t+1}), \quad B_{g,t+1} = -B_{t+1}, \tag{9.126}$$

where capitalized letters denote aggregate variables, and the second equality is the equilibrium requirement in the lending market.

Requiring that agents can swich their types ($\delta > 0$) ensures that a stationary state steady exists, which is compatible with an equilibrium in which both types survive. Intuitively, if $\delta = 0$, the farmers' net-worth would dominate, driven by the farmers' higher productivity. Moreover, persistence in productivity shocks ($1 - \delta > n\delta$) is needed to ensure that a stationary steady state exists.

To corroborate these claims, we calculate the net worth of both farmers and gatherers, and replace them into the farmers' law of new-worth accumulation, Eq. (9.126). By using the optimal consumption $c_t = (1-\beta)\, a_t$ into (9.121) and aggregating, yields the farmers' aggregate net-worth *before* the productivity shock,

$$Y_{t+1} - B_{t+1} = \bar{\alpha}\beta A_t, \quad \bar{\alpha} \equiv \frac{(1-\theta)\,\alpha}{1-\frac{\theta\alpha}{\gamma}}. \tag{9.127}$$

The gatherers' aggregate net-worth is obtained by plugging Eq. (9.122) into the farmers' aggregate debt obtained through Eq. (9.119), and plugging $X_{g,t}$ from Eq. (9.123) into $Y_{g,t+1} = \gamma X_{g,t}$, leaving

$$Y_{g,t+1} + B_{t+1} = \gamma\beta(W_t - A_t). \tag{9.128}$$

Finally, we plug Eqs. (9.127)-(9.128) into Eq. (9.126) and use the expression for $\frac{W_{t+1}}{W_t}$ in Eq. (9.124), obtaining

$$s_{t+1} = \Phi(s_t) \equiv \frac{(1-\delta)\,\bar{\alpha}s_t + n\delta\gamma(1-s_t)}{\gamma + (\alpha-\gamma)\frac{s_t}{\bar{s}}}. \tag{9.129}$$

An equilibrium is a sequence $(s_t)_{t=0}^{\infty}$, for a given initial condition s_0. Note that $\Phi(0) = n\delta$ and $\Phi(1) = 1-\delta$. Therefore, (9.125) guarantees that a fixed point s^* exists, and is such that $s^* = \Phi(s^*)$, with $s^* = 1$ when $\delta = 0$. Moreover, under conditions, there is a unique $s^* < \bar{s}$. So productivity switches are needed to prevent this economy from reaching the trivial steady state with gatherers' extinction. Moreover, persistence of these switches needs not to be too large.

Suppose we are at the steady state, $s^* < \bar{s}$, and that there is an *aggregate* but *temporary* productivity shock at t, in that both α and γ lower by some Δ. After this shock, the farmers' aggregate net worth is, by Eq. (9.126):

$$A_{t+1} = (1-\Delta)((1-\delta)\,\alpha X_t + n\delta\gamma X_{g,t}) - (1-\delta-n\delta)\,B_{t+1}.$$

That is, A_{t+1} obviously goes down, although proportionately more than aggregate output, $W_{t+1} = Y_{t+1} + Y_{g,t+1}$, due to leverage.[27] Therefore, after the aggregate shock, $s_{t+1} < s^*$. It will now take some time for the farmers' aggregate share of net-worth s_t to converge towards its steady state s^*, through Eq. (9.129). During the "recovery process" (i.e., while s_t catches up

[27] That is, after the aggregate shock, $s_{t+1} = \frac{(1-\Delta)Y_{t+1} - (1-\Delta)B_{t+1} - \Delta B_{t+1}}{(1-\Delta)Y_{t+1} + (1-\Delta)Y_{g,t+1}} < \frac{Y_{t+1} - B_{t+1}}{Y_{t+1} + Y_{g,t+1}} = s^*$.

to s^*) output growth $\frac{W_{t+1}}{W_t}$ is obviously lower than that in the steady state (see Eq. (9.124)), and does only gradually converge to it. Note that this propagation mechanism is solely due to credit constraints: a temporary shock in an unconstrained economy will only have a temporary impact, one period, after which growth will immediately stabilize to $\alpha\beta$.

9.14.4 Amplification

The model in the previous section illustrates how propagation mechanisms operate in an economy with borrowing constraints. However, the main idea underlying the financial accelerator doctrine is that of "contagion," that is, spillover of capital markets shocks into the real sphere of the economy. In the examples of the previous section, the farmers' borrowing capacity can depend on the value of some collateral, say "land." In bad times, when the land value drops, the borrowing capacity decreases, which makes land value decrease even more. Land is a parody for assets.

We explain how this channel works by relying on the Kiyotaki and Moore (1997) model, which actually is a model simpler than that in the previous section.

[In progress]

9.14.5 Amplification with a banking sector

Bernanke (1983), Bernanke and Gertler (1999) and early reviews of Bernanke and Gertler (1999).

[In progress]

9.14.5.1 Premise: credit rationing

Credit rationing is perhaps one of the most clear instances that illustrates adverse selection in financial economics.[28] Interest rates are typically taken as "exogenous" in neo-classical macro-finance models, in that any loan rate required by a bank does not affect the riskiness of the project to be undertaken by a borrower. But in the presence of asymmetric information, an adverse selection problem may arise: after an increase in the required loan rate, a bank may well receive riskier loan applications to the extent that the bank's expected profits could actually fall. Below, we shall briefly explain that this property may imply credit rationing.

Stiglitz and Weiss (1981) formalize these ideas in a risk-neutral environment. They consider a market with a number of potential borrowers indexed by a parameter θ, where each of them has a project with the same expected gross return $\bar{R}$, i.e., $E_{\theta_i}(R) \equiv \int_0^\infty Rf(R;\theta_i)\,dR = \bar{R}$ for each borrower i, and $f(R;\theta)$ denotes the density function of R for project θ. Critically, projects are ordered in a way that a greater θ implies a higher risk in the sense of mean preserving spreads (see Appendix 5 to Chapter 1). Equally critically, there is asymmetric information between any borrower and the bank, in that the latter is unaware of the riskiness of the specific project that is being asked to be financed.

Loans are made at a nominal value B, rate r, and are backed by the borrower's collateral C, such that the borrower is insolvent whenever the project return does not cover the firm's liabilities, i.e., $C + R < B(1+r)$. We assume that $C < B(1+r)$ for otherwise the firm would be anticipated not to be in need of financing in the first place. In other words, the profits of the firm,

$$\Pi \equiv \max\{R - B(1+r), -C\}, \tag{9.130}$$

[28] Part I (Chapter 4) contains introductory descriptions of adverse selection problems in financial markets.

are increasing and convex in R. Thus, the expected profits $E_\theta(\Pi)$ are increasing in θ, because by construction, an increase in θ is a mean-preserving spread. It then follows that there is a "break-even" level of risk, $\hat{\theta}$ say, in that the expected profits are zero in correspondence of such a $\hat{\theta}$, i.e., $E_{\hat{\theta}}(\Pi) = 0$. In other words, a firm is willing to borrow only when it carries a project that is sufficiently risky: firms for which $\theta < \hat{\theta}$ will not apply for a loan, and firms for which $\theta \geq \hat{\theta}$ will apply for a loan. Note that these outcomes arise because of convexity of the profits, Π: some investors are willing to take more risks when allowed to default.

These properties lead to adverse selection: the riskiness of the borrowers applying for a loan increases after an increase in the required loan rate, r. Intuitively, any firm's profits lower after r increases, such that, then, a higher risk is needed to raise expected profits so as to break-even. In other words, $\hat{\theta}$ is an increasing function of r, $\hat{\theta}'(r) > 0$ say. The argument is very famous and is made in Appendix 4 to this chapter.

Next, consider the bank profits:

$$\Pi_b \equiv \min\{R + C, B(1 + r)\}, \tag{9.131}$$

which are increasing and concave in R, such that the expected profits $E_\theta(\Pi_b)$ are decreasing in θ. What are the effects of an increase in r on the banks' expected profits? A first, mechanical effect is that the banks' expected profits increase. But there is a second effect, due to adverse selection: by increasing r, the pool of applicants becomes riskier because $\hat{\theta}'(r) > 0$, which depresses the bank's expected profits. In other words, due to limited liability, higher risk borrowers have higher expected returns and, thus, are willing to pay higher interest rates. These two effects can take turns in dominating the overall bank's profitability such that $E_\theta(\Pi_b)$, considered as a function of r, can actually achieve a local maximum.

The equilibrium in the credit market is determined as follows. First, the bank determines the "screening schedule," $\hat{\theta}(r)$ and, then, based on this, calculates the posterior distribution of its profits. This posterior distribution is the distribution of Π_b conditional on the adverse selection outcome that only the riskiest potential borrowers would apply for a loan, $\theta \geq \hat{\theta}(r)$. Naturally, the bank is unable to understand who is the riskiest in the pool of applicants and utilizes the distribution of the types while determining the profits posterior distribution. Based on this posterior distribution, the bank determines the interest rate r^* to charge, which in turns determines the overall demand for loans, $L^d(r^*) \equiv B\int_{\hat{\theta}(r^*)}^{\infty}\gamma(\theta)\,d\theta$, where $\gamma(\theta)$ is the distribution of the potential borrowers.

Note that $L^d(r)$ is decreasing in r but there is no guarantee that the loanable funds supplied by the bank at the optimal r^* coincide with $L^d(r^*)$. Appendix 4 develops a parametric example to illustrate this type of equilibrium credit rationing.

9.14.5.2 Collateral again, and business cycles

Models with costly state verification

[In progress]

9.14.6 Additional literature

Adrian and Shin (2011); Brunnermeier and Sannikov (2013); Danielsson, Zigrand and Shin (2011); Geanakoplos (2010); Gertler and Kiyotaki (2011); He and Krishnamurthy (2012): Hugonnier and Prieto (2012); Shin (2010).

[In progress]

9.15 Appendix 1: Non-expected utility

9.15.1 Detailed derivation of optimality conditions and selected relations

Derivation of Eq. (9.5). We have,

$$\begin{aligned} x_{t+1} &= \sum\nolimits_i \left(P_{it+1} + D_{it+1}\right)\theta_{it+1} \\ &= \sum\nolimits_i \left(P_{it+1} + D_{it+1} - P_{it}\right)\theta_{it+1} + \sum\nolimits_i P_{it}\theta_{it+1} \\ &= \left(1 + \sum\nolimits_i \frac{P_{it+1} + D_{it+1} - P_{it}}{P_{it}} \frac{P_{it}\theta_{it+1}}{\sum_i P_{it}\theta_{it+1}}\right) \sum\nolimits_i P_{it}\theta_{it+1} \\ &= \left(1 + \sum\nolimits_i r_{it+1}\omega_{it}\right)\left(x_t - c_t\right) \end{aligned}$$

where the last line follows by the standard budget constraint $c_t + \sum P_{it}\theta_{it+1} = x_t$, the definition of r_{it+1} and the definition of ω_{it} given in the main text. ■

Optimality. Define

$$\mathcal{W}\left(c, \nu\right) = \frac{1}{1-\gamma}\left(c^{\rho} + \beta\left(\left(1-\gamma\right)\nu\right)^{\frac{\rho}{1-\gamma}}\right)^{\frac{1-\gamma}{\rho}},$$

and consider Eq. (9.6) in the main text,

$$V\left(x_t, y_t\right) = \max_{c_t, \omega_t} \mathcal{W}\left(c_t, E_t\left(V\left(x_{t+1}, y_{t+1}\right)\right)\right).$$

The first order condition for c_t yields

$$\mathcal{W}_1\left(c_t, E_t\left(V\left(x_{t+1}, y_{t+1}\right)\right)\right) = \mathcal{W}_2\left(c_t, E_t\left(V\left(x_{t+1}, y_{t+1}\right)\right)\right) E_t\left[V_1\left(x_{t+1}, y_{t+1}\right)\left(1 + r_M\left(y_{t+1}\right)\right)\right], \quad \text{(9A.1)}$$

where subscripts denote partial derivatives. Thus, optimal consumption is some function of the state $c\left(x, y\right)$ such that,

$$x_{t+1} = \left(x_t - c\left(x_t, y_t\right)\right)\left(1 + r_M\left(y_{t+1}\right)\right).$$

By differentiating the value function with respect to x_t,

$$\begin{aligned} V_1\left(x_t, y_t\right) = {} & \mathcal{W}_1\left(c\left(x_t, y_t\right), E_t\left(V\left(x_{t+1}, y_{t+1}\right)\right)\right) c_1\left(x_t, y_t\right) \\ & + \mathcal{W}_2\left(c\left(x_t, y_t\right), E_t\left(V\left(x_{t+1}, y_{t+1}\right)\right)\right) E_t\left[V_1\left(x_{t+1}, y_{t+1}\right)\left(1 + r_M\left(y_{t+1}\right)\right)\right]\left(1 - c_1\left(x_t, y_t\right)\right), \end{aligned}$$

where subscripts denote partial derivatives. By replacing Eq. (9A.1) into the previous equation leaves the envelope condition for the dynamic programming problem,

$$V_1\left(x_t, y_t\right) = \mathcal{W}_1\left(c\left(x_t, y_t\right), E_t\left(V\left(x_{t+1}, y_{t+1}\right)\right)\right). \quad \text{(9A.2)}$$

By replacing Eq. (9A.2) back into Eq. (9A.1), and rearranging terms,

$$E_t\left[\frac{\mathcal{W}_2\left(c\left(x_t, y_t\right), \nu\left(x_t, y_t\right)\right)}{\mathcal{W}_1\left(c\left(x_t, y_t\right), \nu\left(x_t, y_t\right)\right)} \mathcal{W}_1\left(c\left(x_{t+1}, y_{t+1}\right), \nu\left(x_{t+1}, y_{t+1}\right)\right)\left(1 + r_M\left(y_{t+1}\right)\right)\right] = 1.$$

where we have set $\nu\left(x_t, y_t\right) \equiv E_t\left(V\left(x_{t+1}, y_{t+1}\right)\right)$.

We now show through a similar argument that the same Euler equation applies to any asset $i = 1, \cdots, m$,

$$E_t\left[\frac{\mathcal{W}_2\left(c\left(x_t, y_t\right), \nu\left(x_t, y_t\right)\right)}{\mathcal{W}_1\left(c\left(x_t, y_t\right), \nu\left(x_t, y_t\right)\right)} \mathcal{W}_1\left(c\left(x_{t+1}, y_{t+1}\right), \nu\left(x_{t+1}, y_{t+1}\right)\right)\left(1 + r_i\left(y_{t+1}\right)\right)\right] = 1. \quad \text{(9A.3)}$$

Indeed, we have:

$$V(x_t, y_t) = \max_{c_t, \omega_t} \mathcal{W}(c_t, E_t(V(x_{t+1}, y_{t+1}))) = \max_{\theta_{t+1}} \mathcal{W}(x_t - \sum P_i \theta_{i,t+1}, E_t(V(x_{t+1}, y_{t+1}))),$$

where $x_{t+1} = \sum (P_{i,t+1} + D_{i,t+1}) \theta_{i,t+1}$ and where the portfolio choice $\theta_{i,t+1}$ is made at time t. The first order conditions are

$$\theta_{i,t+1} : 0 = -\mathcal{W}_1(\cdot) P_{i,t} + \mathcal{W}_2(\cdot) E_t[V_1(x_{t+1}, y_{t+1})(P_{i,t+1} + D_{i,t+1})].$$

By replacing Eq. (9A.2) into the previous equation leaves

$$E\left[\frac{\mathcal{W}_2(c(x_t, y_t), \nu(x_t, y_t))}{\mathcal{W}_1((x_t, y_t), (x_t, y_t))} \mathcal{W}_1(c(x_{t+1}, y_{t+1}), \nu(x_{t+1}, y_{t+1})) \frac{P_{i,t+1} + D_{i,t+1}}{P_{i,t}}\right] = 1.$$

Derivation of Eq. (9.7). We need to explicitly determine the stochastic discount factor in Eq. (9A.3)

$$m(x_t, y_t; x_{t+1}, y_{t+1}) = \frac{\mathcal{W}_2(c(x_t, y_t), \nu(x_t, y_t))}{\mathcal{W}_1(c(x_t, y_t), \nu(x_t, y_t))} \mathcal{W}_1(c(x_{t+1}, y_{t+1}), \nu(x_{t+1}, y_{t+1})).$$

Note that

$$\begin{aligned} \mathcal{W}_1(c, \nu) &= \left(c^\rho + \beta((1-\gamma)\nu)^{\frac{\rho}{1-\gamma}}\right)^{\frac{1-\gamma}{\rho}-1} c^{\rho-1} = \mathcal{W}(c, \nu)^{\frac{1-\gamma-\rho}{1-\gamma}} (1-\gamma)^{\frac{1-\gamma-\rho}{1-\gamma}} c^{\rho-1} \\ \mathcal{W}_2(c, \nu) &= \left(c^\rho + \beta((1-\gamma)\nu)^{\frac{\rho}{1-\gamma}}\right)^{\frac{1-\gamma}{\rho}-1} \beta((1-\gamma)\nu)^{\frac{\rho}{1-\gamma}-1} \end{aligned} \tag{9A.4}$$

such that

$$m(x_t, y_t; x_{t+1}, y_{t+1}) = \frac{\mathcal{W}_2(c_t, \nu_t)}{\mathcal{W}_1(c_t, \nu_t)} \mathcal{W}_1(c_{t+1}, \nu_{t+1}) = \beta\left(\frac{\nu_t}{\mathcal{W}(c_{t+1}, \nu_{t+1})}\right)^{\frac{\rho}{1-\gamma}-1} \left(\frac{c_{t+1}}{c_t}\right)^{\rho-1},$$

where $\nu_t \equiv \nu(x_t, y_t)$.

Along any optimal consumption path, $V(x, y) = \mathcal{W}(c(x, y), \nu(x, y))$. Therefore,

$$m(x_t, y_t; x_{t+1}, y_{t+1}) = \beta\left(\frac{E_t(V(x_{t+1}, y_{t+1}))}{V(x_{t+1}, y_{t+1})}\right)^{\frac{\rho}{1-\gamma}-1} \left(\frac{c_{t+1}}{c_t}\right)^{\rho-1}. \tag{9A.5}$$

We are left with evaluating the term $\frac{E_t(V(x_{t+1}, y_{t+1}))}{V(x_{t+1}, y_{t+1})}$. By the envelope condition in (9A.2),

$$\begin{aligned} V_1(x_t, y_t) &= \mathcal{W}_1(c(x_t, y_t), E_t(V(x_{t+1}, y_{t+1}))) \\ &= \mathcal{W}(c(x_t, y_t), \nu(x_t, y_t))^{\frac{1-\gamma-\rho}{1-\gamma}} (1-\gamma)^{\frac{1-\gamma-\rho}{1-\gamma}} c(x_t, y_t)^{\rho-1} \\ &= V(x_t, y_t)^{\frac{1-\gamma-\rho}{1-\gamma}} (1-\gamma)^{\frac{1-\gamma-\rho}{1-\gamma}} c(x_t, y_t)^{\rho-1}, \end{aligned}$$

where the first equality follows by Eq. (9A.2), the second by the first of Eqs. (9A.4), and the last is the optimality condition. We conjecture that $V(x, y) = b(y) \frac{x^{1-\gamma}}{1-\gamma}$, such that the previous expression delivers the agent's optimal consumption:

$$c(x_t, y_t) = a(y_t) x_t, \quad a(y) \equiv b(y)^{\frac{\rho}{(1-\gamma)(\rho-1)}}. \tag{9A.6}$$

Therefore,

$$x_{t+1} = (1 - a(y_t)) x_t (1 + r_M(y_{t+1})), \tag{9A.7}$$

and

$$\frac{E_t\left(V\left(x_{t+1},y_{t+1}\right)\right)}{V\left(x_{t+1},y_{t+1}\right)}=\frac{E_t\left(b\left(y_{t+1}\right)\left(1+r_M\left(y_{t+1}\right)\right)^{1-\gamma}\right)}{b\left(y_{t+1}\right)\left(1+r_M\left(y_{t+1}\right)\right)^{1-\gamma}}. \tag{9A.8}$$

Along any optimal path, $V\left(x,y\right)=\mathcal{W}\left(c\left(x,y\right),\nu\left(x,y\right)\right)$. Using the expression of $\mathcal{W}$ in Eq. (9.6) and again the conjecture made on V,

$$\begin{aligned}V_t&=b\left(y_t\right)\frac{x_t^{1-\gamma}}{1-\gamma}\\&=\frac{1}{1-\gamma}\left(a\left(y_t\right)^{\rho}x_t^{\rho}+\beta E_t\left(b\left(y_{t+1}\right)x_{t+1}^{1-\gamma}\right)^{\frac{\rho}{1-\gamma}}\right)^{\frac{1-\gamma}{\rho}}\\&=\frac{1}{1-\gamma}\left(a\left(y_t\right)^{\rho}x_t^{\rho}+\beta E_t\left(b\left(y_{t+1}\right)\left(1-a\left(y_t\right)\right)^{1-\gamma}x_t^{1-\gamma}\left(1+r_M\left(y_{t+1}\right)\right)^{1-\gamma}\right)^{\frac{\rho}{1-\gamma}}\right)^{\frac{1-\gamma}{\rho}},\end{aligned}$$

and rearranging terms leaves:

$$E_t\left(b\left(y_{t+1}\right)\left(1+r_M\left(y_{t+1}\right)\right)^{1-\gamma}\right)=\beta^{-\frac{1-\gamma}{\rho}}\left(\frac{a\left(y_t\right)}{1-a\left(y_t\right)}\right)^{\frac{(1-\gamma)(\rho-1)}{\rho}}. \tag{9A.9}$$

Moreover, using the definition of $a\left(y\right)$ in Eq. (9A.6) leaves

$$b\left(y_{t+1}\right)\left(1+r_M\left(y_{t+1}\right)\right)^{1-\gamma}=\left(a\left(y_{t+1}\right)\left(1+r_M\left(y_{t+1}\right)\right)^{\frac{\rho}{\rho-1}}\right)^{\frac{(1-\gamma)(\rho-1)}{\rho}}. \tag{9A.10}$$

Plugging Eqs. (9A.9)-(9A.10) into Eq. (9A.8),

$$\begin{aligned}\frac{E_t\left(V\left(x_{t+1},y_{t+1}\right)\right)}{V\left(x_{t+1},y_{t+1}\right)}&=\beta^{-\frac{1-\gamma}{\rho}}\left(\frac{a\left(y_t\right)}{\left(1-a\left(y_t\right)\right)a\left(y_{t+1}\right)\left(1+r_M\left(y_{t+1}\right)\right)^{\frac{\rho}{\rho-1}}}\right)^{\frac{(1-\gamma)(\rho-1)}{\rho}}\\&=\beta^{-\frac{1-\gamma}{\rho}}\left(\left(\frac{c_{t+1}}{c_t}\right)^{-1}\frac{x_{t+1}}{\left(1-a\left(y_t\right)\right)x_t\left(1+r_M\left(y_{t+1}\right)\right)^{\frac{\rho}{\rho-1}}}\right)^{\frac{(1-\gamma)(\rho-1)}{\rho}}\\&=\beta^{-\frac{1-\gamma}{\rho}}\left(\left(\frac{c_{t+1}}{c_t}\right)^{-1}\frac{1}{\left(1+r_M\left(y_{t+1}\right)\right)^{\frac{1}{\rho-1}}}\right)^{\frac{(1-\gamma)(\rho-1)}{\rho}}\end{aligned}$$

where the first equality follows by Eq. (9A.6) and the second by Eq. (9A.7). The result follows by replacing the previous expression into Eq. (9A.5). ■

Proof of Eqs. (9.12) and (9.13). By the standard property that if $\tilde{y}$ is normally distributed, $\ln E(e^{\tilde{y}})=E(\tilde{y})+\frac{1}{2}var\left(\tilde{y}\right)$, we can elaborate on Eq. (9.10), obtaining,

$$\begin{aligned}0&=\ln E_t\left(e^{-\delta\theta-\frac{\theta}{\psi}\ln(c_{t+1}/c_t)+\theta R_{M,t+1}}\right)\\&=-\delta\theta-\frac{\theta}{\psi}E_t\left(\ln\frac{c_{t+1}}{c_t}\right)+\theta E_t(R_{M,t+1})+\frac{1}{2}\left(\left(\frac{\theta}{\psi}\right)^2\sigma_c^2+\theta^2\sigma_{R_M}^2-2\frac{\theta^2}{\psi}\sigma_{R_M,c}\right).\end{aligned} \tag{9A.11}$$

We do the same in Eq. (9.11), and obtain:

$$R_{f,t}=\delta\theta+\frac{\theta}{\psi}E_t\left(\ln\frac{c_{t+1}}{c_t}\right)-(\theta-1)E_t(R_{M,t+1})-\frac{1}{2}\left(\left(\frac{\theta}{\psi}\right)^2\sigma_c^2+(\theta-1)^2\sigma_{R_M}^2-2\frac{\theta\left(\theta-1\right)}{\psi}\sigma_{R_M,c}\right). \tag{9A.12}$$

By replacing Eq. (9A.12) into Eq. (9A.11), we obtain Eq. (9.12) in the main text.

To obtain the risk-free rate $R_{f,t}$ in Eq. (9.13), we replace the expression for $E_t(R_{M,t+1})$ in Eq. (9.12) into Eq. (9A.12). ■

9.15.2 Details regarding models of long-run risks

Proof of Eq. (9.22). By substituting the guess $z_t = a_0 + a_1 x_t$ into Eq. (9.21),

$$\begin{aligned}
0 &= (\kappa_0 - (1-\kappa_1)\, a_0 - \delta)\, \theta + \ln E_t \left(e^{-\frac{\theta}{\psi} \ln(c_{t+1}/c_t) + \theta \kappa_1 a_1 x_{t+1} - \theta a_1 x_t + \theta g_{t+1}} \right) \\
&= \theta \left(\left((\kappa_0 - (1-\kappa_1)\, a_0 - \delta) + \left(1 - \frac{1}{\psi}\right) \left(g_0 - \frac{1}{2}\sigma_c^2 \right) \right) \right) + \ln E_t \left(e^{\theta \left(1 - \frac{1}{\psi}\right) \epsilon_t + \theta \kappa_1 a_1 \eta_t} \right) \\
&+ \left((\kappa_1 \rho - 1)\, a_1 + \left(1 - \frac{1}{\psi}\right) \right) x_t \\
&\equiv \text{const}_1 + \text{const}_2 \cdot x_t,
\end{aligned}$$

where the second equality follows by Eqs. (9.19) and (9.20). Note, then, that this equality can only hold if the two constants, const_1 and const_2 are both zero. Imposing $\text{const}_2 = 0$ yields,

$$a_1 = \frac{1 - \frac{1}{\psi}}{1 - \kappa_1 \rho},$$

as in Eq. (9.22) in the main text. Imposing $\text{const}_1 = 0$, and using the solution for a_1, yields the solution for the constant a_0. ■

9.15.3 Continuous time

Duffie and Epstein (1992a,b) extend the framework on non-expected utility to continuous time. Heuristically, the continuation utility is the continuous time limit of,

$$v_t = \left(c_t^{\rho} \Delta t + e^{-\delta \Delta t} \left(E(v_{t+\Delta t}^{1-\gamma}) \right)^{\frac{\rho}{1-\gamma}} \right)^{1/\rho}.$$

Continuation utility v_t solves the following stochastic differential equation,

$$dv_t = \left(-f(c_t, v_t) - \frac{1}{2} A(v_t) \left\| \sigma_{vt} \right\|^2 \right) dt + \sigma_{vt} dB_t, \quad \text{with} \;\; v_T = 0$$

Now, (f, A) is the aggregator, with A being a variance multiplier, placing a penalty proportional to utility volatility $\left\| \sigma_{vt} \right\|^2$. The aggregator (f, A) is the continuous time counterpart to the aggregator $(W, \hat{v})$ of the discrete time case. The solution to the previous "stochastic differential utility" is:

$$v_t = E \left[\int_t^T \left(f(c_s, v_s) + \frac{1}{2} A(v_s) \left\| \sigma_{vs} \right\|^2 \right) ds \right].$$

For example we can take,

$$f(c, v) = \frac{c^{\rho} - \beta (\alpha v)^{\rho/\alpha}}{\rho (\alpha v)^{(\rho - \alpha)/\alpha}},$$

with $\rho \neq 0$ and $\alpha \leq 1$. The standard additive utility case is obtained once $A = 0$ and $\rho = \alpha$, in which case, $f(c, v) = u(c) - \beta v$ (see, e.g., Duffie and Epstein (1992, p. 367).

[In progress]

9.16 Appendix 2: Economies with heterogenous agents

Economies with a continuum of agents. We study asset prices in economies with complete markets by relying on a centralized decision mechanims that delivers the same outcome as that of the decentralized, market based. This approach was developed by Huang (1987), and extends the approach to deal with the classical static mode, explained in Part I (see Chapter 2).

We consider a set continuum of agents, A, indexed by an instantaneous utility function $u_a(c)$, where for each $a \in A$, c is consumption. The case with a discrete number of agents is dealt with a change in notation, and actually analyzed in some of the examples below. Because markets are complete, the market allocation is Pareto efficient, and by the second welfare theorem, it can be "implemented" by means of the following program,

$$\max_{c_a} E\left[\int_0^\infty e^{-\delta t}\left(\int_{a\in A} u_a(c_{at})\, f_a da\right) dt\right], \quad \text{s.t.} \quad \int_{a\in A} c_{at} da = D_t,$$

or, because there is no intertemporal transfer of resources,

$$U(D) \equiv \max_{(c_a)_{a\in A}} \int_{a\in A} u_a(c_a)\, f_a da, \quad \text{s.t.} \quad \int_{a\in A} c_a da = D, \tag{9A.P1}$$

where D is the aggregate endowment in the economy.

We explain how the equilibrium price system is determined as the Arrow-Debreu state price density in an economy with a single agent endowed with the aggregate endowment D, instantaneous utility function $U(D)$, and where the social weighting function f_a is the reciprocal of the marginal utility of income for agent $a \in A$. Indeed, the first order conditions for the program [9A.P1] are

$$u_a'(c_{at})\, f_a = \psi_t, \quad \text{for all } a \in A, \tag{9A.13}$$

where $u_a'(c_a)$ denotes the partial derivative of u_a with respect to c_a, and ψ_t is a Lagrange multiplier. Moreover, by differentiating U with respect to D,

$$U_C(D_t) = \int_{a\in A} u_a'(c_{at}) \frac{\partial c_{at}}{\partial D_t} f_a da = \psi_t \int_{a\in A} \frac{\partial c_{at}}{\partial D_t} da = \psi_t, \tag{9A.14}$$

where $U_C(D)$ denotes the partial derivative of U with respect to D. The second equality in Eq. (9A.14) follows by the optimality conditions (9A.13), and the third, by differentiating the constraint of the social plan [9A.P1]. Combining Eqs. (9A.13) and (9A.14) leaves

$$\frac{U_C(D_t)}{U_C(D_0)} = \frac{u_a'(c_{at})}{u_a'(c_{a0})}, \quad \text{for all } a \in A. \tag{9A.15}$$

On the other hand, the market general equilibrium allocation satisfies

$$e^{\delta t}\frac{\xi_t}{\xi_0} = \frac{u_a'(c_{at})}{u_a'(c_{a0})}, \quad \text{for all } a \in A. \tag{9A.16}$$

To show that the allocations in Eq. (9A.15) and in Eq. (9A.16) are the same, we need to show that

$$U_C(D_t) = e^{\delta t}\xi_t. \tag{9A.17}$$

We show that Eq. (9A.17) holds true once the social weight $f_a = \kappa_a^{-1}$, where κ_a is the marginal utility of income of agent a. Indeed, in this case, Eq. (9A.13) and the optimality conditions for the decentralized economy lead to,

$$\text{for all } a \in A, \quad u_a'(c_{at}) = \kappa_a \psi_t \quad \text{and} \quad u_a'(c_{at}) = \kappa_a e^{\delta t}\xi_t. \tag{9A.18}$$

By aggregating the first and the second of the previous conditions leaves:

$$D_t = \int_{a\in A} \varphi_a\left(\kappa_a \psi_t\right) da, \quad D_t = \int_{a\in A} \hat{\varphi}_a\left(\kappa_a e^{\delta t}\xi_t\right) da,$$

where φ_a ans $\hat{\varphi}_a$ denote the inverse functions for consumption, as implied by the social and private allocations in (9A.18). Eq. (9A.17) follows by an argument similar to that used to show Theorem 2.7 in Chapter 2, and by the envelope theorem in Eq. (9A.14). Therefore, the pricing kernel in this economy with heterogeneous agents and complete markets has the same pricing implication as the economy with a single agent as explained, viz

$$\frac{\xi_t}{\xi_0} = e^{-\delta t}\frac{U_C\left(D_t\right)}{U_C\left(D_0\right)}. \tag{9A.19}$$

The practical merit of this approach is that while the marginal utility of income is unobservable, the thusly constructed Arrow-Debreu state price density depends on the "infinite dimensional parameter", f, which could be calibrated to match selected quantitative features of consumption and asset price data.

We now apply this approach and derive the equilibrium conditions in two models and then, move to study the allocation process in an incomplete market setting.

"CATCHING UP WITH THE JONESES" (CHAN AND KOGAN, 2002). In this model, markets are complete, and we have that $A = [1, \infty]$ and the instantaneous utility of agent γ is, $u_\gamma\left(c_\gamma, x\right) = \left(c_\gamma / x\right)^{1-\gamma} / \left(1-\gamma\right)$, where x is the "standard of living of others," as explained in the main text.

The static optimization problem for the social planner in [9A.P1] can be written as,

$$u\left(D, x\right) = \max_{c_\gamma} \int_1^\infty \frac{\left(c_\gamma / x\right)^{1-\gamma}}{1-\gamma} f_\gamma d\gamma, \qquad \text{s.t. } \int_1^\infty \frac{c_\gamma}{x} d\gamma = \frac{D}{x}. \tag{9A.20}$$

The first order conditions for this problem lead to,

$$\left(\frac{c_\gamma}{x}\right)^{-\gamma} f_\gamma = \psi\left(\frac{D}{x}\right), \tag{9A.21}$$

where ψ is a Lagrange multiplier, a function of the aggregate endowment D, normalized by x. It is determined by Eq. (9.26), which is obtained by replacing Eq. (9A.21) into the budget constraint of the social planner, the second of Eqs. (9A.20). The value function in Eq. (9.25) of the main text follows by replacing Eq. (9A.21) into the maximed value of the intertemporal utility, the first of Eqs. (9A.20). General equilibrium allocations, and prices, are obtained by setting $f\left(\gamma\right)$ equal to the reciprocal of the marginal utility of income for agent γ.

The expression for the unit risk-premium in Eq. (9.27) follows by results given in Section 8.5.1 of Chapter 8,

$$\lambda\left(s\left(D, x\right)\right) = -\left(\frac{\partial^2 U\left(s\left(D, x\right)\right)}{\partial D^2} \Big/ \frac{\partial U\left(s\left(D, x\right)\right)}{\partial D}\right)\sigma_0 D, \quad s\left(D, x\right) \equiv \ln\frac{D}{x}, \tag{9A.22}$$

where $U\left(s\right)$ is the value function in Eq. (9.25). To evaluate the previous expression for λ, note, first, that:

$$\frac{\partial U\left(s\left(D, x\right)\right)}{\partial D} = -\frac{\psi'\left(s\left(D, x\right)\right)}{D}\int_1^\infty \frac{1}{\gamma} f_\gamma^{\frac{1}{\gamma}} \psi\left(s\left(D, x\right)\right)^{-\frac{1}{\gamma}} d\gamma. \tag{9A.23}$$

Moreover, by differentiating Eq. (9.26) with respect to D, using Eq. (9A.23), and rearranging terms, leads to $\partial U\left(s\left(D, x\right)\right)/\partial D = \psi\left(s\left(D, x\right)\right)/x$. Differentiating this expression for $\partial U/\partial D$ with respect to D again, produces:

$$\frac{\partial^2 U\left(s\left(D, x\right)\right)}{\partial D^2} = \frac{\psi'\left(s\left(D, x\right)\right)}{x}\frac{1}{D}. \tag{9A.24}$$

Replacing Eqs. (9A.23)-(9A.24) into Eq. (9A.22) yields Eq. (9.27) in the main text. In this fictitious representative agent economy, the short-term rate is the expectation of the stochastic discount factor. It equals, again by results given in Section 8.5.1 of Chapter 8,

$$\begin{aligned} r(D,s,x) &= \rho - g_0 \frac{\partial^2 U\left(s\left(D,x\right)\right)/\partial D^2}{\partial U\left(s\left(D,x\right)\right)/\partial D} D - \theta \frac{\partial^2 U\left(s\left(D,x\right)\right)/\partial D\partial x}{\partial U\left(s\left(D,x\right)\right)/\partial D} sx - \frac{1}{2}\sigma_0^2 \frac{\partial^3 U\left(s\left(D,x\right)\right)/\partial D^3}{\partial U\left(s\left(D,x\right)\right)/\partial D} D^2 \\ &= \rho + \frac{g_0}{\sigma_0}\lambda\left(s\left(D,x\right)\right) - \theta\left(1 + \frac{\psi'\left(s\left(D,x\right)\right)}{\psi\left(s\left(D,x\right)\right)}\right) s\left(D,x\right) - \frac{1}{2}\sigma_0^2\left(\frac{\psi''\left(s\left(D,x\right)\right) - \psi'\left(s\left(D,x\right)\right)}{\psi\left(s\left(D,x\right)\right)}\right). \end{aligned}$$

It is instructive to compare the first order conditions of the social planner in Eq. (9A.21) with those in the decentralized economy. Because markets are complete, the optimality conditions in the decentralized economy satisfy:

$$e^{-\delta t}\left(\frac{c_{\gamma t}^{\mathrm{m}}}{x_t}\right)^{-\gamma} = \kappa_\gamma \xi_t x_t, \tag{9A.25}$$

where κ_γ is the marginal utility of income for agent γ, as usual. Aggregating the market allocations c_γ^{m} in Eq. (9A.25), and solving for ξ_t yields,

$$D_t = \int_1^\infty c_{\gamma t}^{\mathrm{m}} d\gamma \Longrightarrow e^{\delta t}\xi_t x_t = \left(\frac{D_t}{x_t}\right)^{-\gamma}\left(\int_1^\infty \kappa_\gamma^{-1} d\gamma\right)^{\gamma}. \tag{9A.26}$$

By aggregating the social weighted allocations, $c_{\gamma t}^{\mathrm{s}}$ say, in Eq. (9A.21), with $f = \kappa^{-1}$,

$$D_t = \int_1^\infty c_{\gamma t}^{\mathrm{s}} d\gamma \Longrightarrow \psi\left(\frac{D_t}{x_t}\right) = \left(\frac{D_t}{x_t}\right)^{-\gamma}\left(\int_1^\infty \kappa_\gamma^{-1} d\gamma\right)^{\gamma}. \tag{9A.27}$$

Comparing the two implications in (9A.26) and (9A.27) leads to the conclusion that the social allocations with $f = \kappa^{-1}$, are the same as the private, and that the counterpart to Eq. (9A.19) is,

$$\frac{\xi_t}{\xi_0} = e^{-\delta t}\frac{x_t^{-1}\psi\left(\frac{D_t}{x_t}\right)}{x_0^{-1}\psi\left(\frac{D_0}{x_0}\right)} = e^{-\delta t}\frac{U_C\left(D_t, x_t\right)}{U_C\left(D_0, x_t\right)}.$$

Restricted stock market participation (Basak and Cuoco, 1998). Given Eq. (9.39), and results given in Section 8.5.1 of Chapter 8, the unit risk-premium, $\lambda(D,x)$, solves a fixed point problem:

$$\lambda(D,x) = -\frac{U_{11}(D,x)}{U_1(D,x)}\sigma_0 D + \frac{U_{12}(D,x)x}{U_1(D,x)}\lambda(D,x).$$

That is,

$$\lambda(D,x) = -\frac{U_1(D,x)U_{11}(D,x)}{U_1(D,x) - U_{12}(D,x)x} \cdot \frac{\sigma_0 D}{U_1(D,x)}. \tag{9A.28}$$

We claim that:

$$U_1(D,x) = u_p'(c_p^*), \text{ and } \frac{U_1(D,x)U_{11}(D,x)}{U_1(D,x) - U_{12}(D,x)x} = u_p''(c_p^*), \tag{9A.29}$$

where c_p^* and c_n^* are the social planner consumption allocations. By replacing Eqs. (9A.29) into Eq. (9A.28), and using the definition of u_p and s, leads to the expression for λ in Eq. (9.34). The expression for the short-term rate r in Eq. (9.33) can be found similarly, and again, through the results given in Section 8.5.1 of Chapter 8.

We now show that Eq. (9A.29) hold true. Consider the Lagrangean for the maximization problem in Eq. (9.38),

$$L = u_p\left(c_p\right) + x u_n\left(c_n\right) + \nu\left(c - c_p - c_n\right),$$

where ν is a Lagrange multiplier, $x = \frac{u_p'(\hat{c}_p)}{u_n'(\hat{c}_n)}$, and $\hat{c}_p$ and $\hat{c}_n$ are the market consumption allocations. The first order conditions for the social planner lead to social allocation functions $c_p^* = c_p(D,x)$ and $c_p^* = c_n(D,x)$, and Lagrange multiplier $\nu(D,x)$, satisfying:

$$u_p'(c_p(D,x)) = \nu(D,x) = x u_n'(c_n(D,x)). \tag{9A.30}$$

Accordingly, the value of the problem in Eq. (9.38) is:

$$U(D,x) = u_p(c_p(D,x)) + x u_n(c_n(D,x)),$$

such that

$$\begin{aligned} U_1(D,x) &= u_p'(c_p(D,x))\frac{\partial c_p(D,x)}{\partial D} + x u_n'(c_n(D,x))\frac{\partial c_n(D,x)}{\partial D} \\ &= u_p'(c_p(D,x))\left(\frac{\partial c_p(D,x)}{\partial D} + \frac{\partial c_n(D,x)}{\partial D}\right) \\ &= u_p'(c_p(D,x)), \end{aligned} \tag{9A.31}$$

where the second equality follows by the first order conditions in Eq. (9A.30), and the third equality holds by differentiating the equilibrium condition

$$D = c_p(D,x) + c_n(D,x), \tag{9A.32}$$

with respect to D.

Eq. (9A.31) establishes the first claim in Eqs. (9A.29). To prove the second claim, invert, first, the first order condition with respect to ν, obtaining, $c_p(D,x) = u_p'^{-1}[\nu(D,x)]$ and $c_n(D,x) = u_n'^{-1}[\nu(D,x)/x]$. Replace, then, these inverse functions into Eq. (9A.32),

$$D = u_p'^{-1}[\nu(D,x)] + u_n'^{-1}[\nu(D,x)/x], \tag{9A.33}$$

where, by Eq. (9A.32) and Eq. (9A.31),

$$\nu(D,x) = u_p'(c_p(D,x)) = U_1(D,x). \tag{9A.34}$$

Differentiating Eq. (9A.33) with respect to x and D, and using Eq. (9A.34), leaves:

$$0 = \frac{1}{u_p''(c_p(D,x))}U_{12}(D,x) + \frac{1}{u_n''(c_n(D,x))}\frac{U_{12}(D,x)\,x - U_1(D,x)}{x^2}, \tag{9A.35}$$

and

$$1 = \frac{1}{u_p''(c_p(D,x))}U_{11}(D,x) + \frac{1}{u_n''(c_n(D,x))}\frac{U_{11}(D,x)}{x}. \tag{9A.36}$$

Replacing Eq. (9A.36) into Eq. (9A.35) leaves:

$$0 = \frac{U_{12}(D,x)}{u_p''(c_p(D,x))} + \left(\frac{1}{U_{11}(D,x)} - \frac{1}{u_p''(c_p(D,x))}\right)\frac{U_{12}(D,x)\,x - U_1(D,x)}{x}.$$

The second relation in Eqs. (9A.29) follows by rearranging terms in the previous relation.

Extinction I: derivation of Eq. (9.67). Denote with W_{0i} the initial wealth of agent i. Note that the budget constraints faced by the two agents are: $W_{0i} = E(\xi_T c_{iT})$, $i = 1,2$, such that by using the expressions for ξ_T and c_{it} in Eqs. (9.65) and (9.66),

$$1 = \frac{W_{02}}{W_{01}} = \frac{E(\xi_T c_{2T})}{E(\xi_T c_{1T})} = \frac{E\left((b\eta_T)^{1/\gamma}(1+(b\eta_T)^{1/\gamma})^{\gamma-1}D_T^{1-\gamma}\right)}{E\left((1+(b\eta_T)^{1/\gamma})^{\gamma-1}D_T^{1-\gamma}\right)}, \tag{9A.37}$$

where the first equality follows by the assumptions that rational and irrational agents have the same initial endowments.

Let us introduce a new probability, $\breve{P}$, defined through the Radon-Nikodym derivative,

$$\left.\frac{d\breve{P}}{dP}\right|_{\mathbb{F}_T} = \frac{D_T^{1-\gamma}}{E(D_T^{1-\gamma})},$$

such that Eq. (9A.37) can be re-written as

$$\breve{E}\left((b\eta_T)^{1/\gamma}\left(1+(b\eta_T)^{1/\gamma}\right)^{\gamma-1}\right) = \breve{E}\left(\left(1+(b\eta_T)^{1/\gamma}\right)^{\gamma-1}\right), \tag{9A.38}$$

where $\breve{E}$ denotes the expectation under $\breve{P}$. By Girsanov's theorem, we have that under $\breve{P}$, η_t is solution to

$$\eta_T = e^{\left((1-\gamma)\upsilon\sigma^2-\frac{1}{2}\sigma^2\upsilon^2\right)T+\sigma\upsilon\breve{W}_T},$$

where $\breve{W}_t$ is a Brownian motion under $\breve{P}$. That is, utilizing the expression for b in Eq. (9.67), Eq. (9A.38) can be written as

$$\breve{E}\left(\breve{\eta}_T^{1/\gamma}(1+\breve{\eta}_T^{1/\gamma})^{\gamma-1}\right) = \breve{E}\left((1+\breve{\eta}_T^{1/\gamma})^{\gamma-1}\right), \quad \breve{\eta}_T \equiv e^{-\frac{1}{2}\sigma^2\upsilon^2T+\sigma\upsilon\breve{W}_T}. \tag{9A.39}$$

KRWW show indeed that Eq. (9A.39) holds true.

Extinction II: derivation of Eq. (9.68). We determine the equilibrium asset price in the logarithmic preferences case. We have:

$$S_t = E_t\left(\xi_T D_T\right) = \frac{E_t\left(1+\eta_T\right)}{E_t\left(\left(1+\eta_T\right)D_T^{-1}\right)} = \frac{1+\eta_t}{E_t\left(\left(1+\eta_T\right)D_T^{-1}\right)},$$

where the third equality follows by the martingale property of η_τ. To determine the denominator of the previous ratio, note that

$$\begin{aligned}
E_t\left[\left(1+\eta_T\right)\left(\frac{D_T}{D_t}\right)^{-1}\right] &= e^{-\mu(T-t)} + E_t\left[\eta_T\left(\frac{D_T}{D_t}\right)^{-1}\right] \\
&= e^{-\mu(T-t)} + \eta_t E_t^2\left[\left(\frac{D_T}{D_t}\right)^{-1}\right] \\
&= e^{-\mu(T-t)} + \eta_t e^{-\mu(T-t)}\cdot e^{-\upsilon\sigma^2(T-t)},
\end{aligned}$$

where the second equality follows by a change in probability and the third, by Girsanov's theorem and Eq. (9.62). Therefore,

$$S_t = D_t\frac{1+\eta_t}{E_t\left(\left(1+\eta_T\right)\left(\frac{D_T}{D_t}\right)^{-1}\right)} = D_t\frac{1+\eta_t}{e^{-\mu(T-t)}\left(1+\eta_t\cdot e^{-\upsilon\sigma^2(T-t)}\right)}. \tag{9A.40}$$

The price prevailing in a rational market, S_t^*, is determined once we set $\upsilon = 0$ and $\eta_\tau = 1$ for all τ, such that by Eq. (9A.40),

$$S_t^* = D_t\frac{1}{e^{-\mu(T-t)}}.$$

Replacing this value for S_t^* back to Eq. (9A.40) produces Eq. (9.68).

9.17 Appendix 3: Knightian uncertainty

Smooth ambiguity aversion. Given the assumptions on h in the main text, we have:

$$\begin{aligned} E_{\mathcal{M}}\left[h\left(E_{\mu}\left(-e^{-\tau W_x}\right)\right)\right] &= -E\left[\left(E\left(e^{-\tau W_x}\middle|\,\mu\right)\right)^{\alpha}\right] \\ &= -E\left[\exp\left(-\alpha\tau\left(E\left(f\middle|\,\mu\right)-p\right)x+\frac{1}{2}\alpha\tau^2\mathrm{var}\left(f\middle|\,\mu\right)x^2\right)\right]. \end{aligned}$$

Note that $E\left(f\middle|\,\mu\right)$ is normally distributed with, say, mean $\mu_E \equiv E\left[E\left(f\middle|\,\mu\right)\right]$ and variance $\sigma_E^2 \equiv \mathrm{var}\left[E\left(f\middle|\,\mu\right)\right]$, such that,

$$E_{\mathcal{M}}\left[h\left(E_{\mu}\left(-e^{-\tau W_x}\right)\right)\right] = -E\left[\exp\left(-\alpha\tau\left(\mu_E-p\right)x+\frac{1}{2}\alpha^2\tau^2\sigma_E^2x^2+\frac{1}{2}\alpha\tau^2\mathrm{var}\left(f\middle|\,\mu\right)x^2\right)\right]. \quad (9A.41)$$

By the Law of Iterated Expectations,

$$\mu_E \equiv E\left[E\left(f\middle|\,\mu\right)\right] = E\left(f\right) = \mu_0.$$

Regarding σ_E^2, note that by the Law of Total Variance,

$$\mathrm{var}\left(f\right) = \mathrm{var}\left[E\left(f\middle|\,\mu\right)\right] + E\left[\mathrm{var}\left(f\middle|\,\mu\right)\right] \equiv \sigma_E^2 + \mathrm{var}\left(f\middle|\,\mu\right).$$

Replacing the values of μ_E and σ_E^2 implied by the previous two equalities into Eq. (9A.41) leaves Eq. (9.94).

Solution of the multiple likelihood model. Consider the APT equation delivering $\mathcal{E}_t^h$, i.e., the asset expected returns under the distorted probability in (9.96),

$$\begin{aligned} \mathcal{E}^h &= \frac{L^hP}{P}+\frac{D}{P} \\ &= r_t+\gamma\sigma_0\cdot\mathrm{Vol}_t \\ &= \rho+\gamma x-\frac{1}{2}\gamma\left(\gamma+1\right)\sigma_0^2+\gamma\sigma_0\left(\sigma_0+\tfrac{p'(x)}{p(x)}v\left(x\right)\right) \end{aligned} \quad (9A.42)$$

where L^h denotes the infinitesimal generator of $(D,\hat{g})$ in (9.100), and the last two lines follow by Eqs. (9.101). That is, the price-dividend ratio is solution to the following partial differential equation:

$$0 = Lp\left(x\right)+\sigma_0v\left(x\right)\left(1-\gamma\right)p'\left(x\right)-\left(\rho-\left(1-\gamma\right)x-\frac{1}{2}\gamma\left(\gamma+1\right)\sigma_0^2+\gamma\sigma_0^2\right)p\left(x\right)+1.$$

Conjecture that $p\left(x\right)=\alpha_0+\alpha_1x$, for two positive constants. By plugging this guess back to the previous equation confirms that the price-dividend ratio is indeed affine in the expected growth. To determine α_0 and α_1, one may proceed through the following intuitive arguments. Note that

$$p\left(\hat{g}_0\right) = \int_0^{\infty}e^{-\rho t}E^{\hat{h}}\left(\frac{D_t}{D_0}\right)^{1-\gamma}dt = \pi_0\frac{1}{\rho-\left(1-\gamma\right)\left(\bar{\theta}^{\hat{h}}-\frac{1}{2}\gamma\sigma_0^2\right)}+\left(1-\pi_0\right)\frac{1}{\rho-\left(1-\gamma\right)\left(\underline{\theta}^{\hat{h}}-\frac{1}{2}\gamma\sigma_0^2\right)},$$

where the first equality is a standard evaluation formula, and the second relies on Eq. (9.100). Replacing the definition of expected dividend growth into the previous equation, $\hat{g}_t=\pi_t\bar{\theta}^{\hat{h}}+\left(1-\pi_t\right)\underline{\theta}^{\hat{h}}$, delivers Eq. (9.102).

Finally, note that he expression for the price-dividend ratio in Eq. (9.102) simplifies when ambiguity aversion is the same in all states of nature. Precisely, assume that $\eta\left(\underline{\theta}\right)=\eta\left(\bar{\theta}\right)\equiv\hat{\eta}$, in which case $\mathcal{A}_t$

collapses to $-\sqrt{\bar{\eta}}$, and $g_t = \hat{g}_t + \sqrt{\bar{\eta}}\sigma_0$, such that the aggregate dividends under the distorted dynamics are

$$\begin{cases} \dfrac{dD_t}{D_t} = \left(g_t - \sqrt{\bar{\eta}}\sigma_0\right) dt + \sigma_0 dW_t \\ dg_t = \frac{1}{\sigma_0}(g_t - \underline{\theta})(\bar{\theta} - \gamma_t) dW_t \end{cases}$$

and the price-dividend ratio is

$$p\left(g_t\right) = \frac{1}{\bar{\theta} - \underline{\theta}} \left(\frac{g_t - \underline{\theta}}{\rho_\eta - (1-\gamma)\left(\bar{\theta} - \frac{1}{2}\gamma\sigma_0^2\right)} + \frac{\bar{\theta} - g_t}{\rho_\eta - (1-\gamma)\left(\underline{\theta} - \frac{1}{2}\gamma\sigma_0^2\right)} \right). \tag{9A.43}$$

where $\rho_\eta \equiv \rho + (1-\gamma)\sqrt{\bar{\eta}}\sigma_0$. The analysis summarized by Figure 9.2 is based on this expression for the price dividend ratio.

9.18 Appendix 4: Credit rationing

We provide proofs of selected results regarding the Stiglitz and Weiss (1981) model as well as a parametric example to illustrate the model.

Risk increases with loan rates. Consider the definition of Π in Eq. (9.130) and the break-even level of risk $\hat{\theta}$, defined as the function of the loan rate r that solves the following equation:

$$\begin{aligned}\hat{\theta}(r) : 0 &= \bar{\Pi}(r, \hat{\theta}(r)) \\ &\equiv E_{\hat{\theta}(r)}(\Pi) = \int_0^{B(1+r)-C} (-C) f(R; \hat{\theta}(r)) dR + \int_{B(1+r)-C}^{\infty} (R - B(1+r)) f(R; \hat{\theta}(r)) dR,\end{aligned} \tag{9A.44}$$

such that

$$0 = \frac{\partial \bar{\Pi}(r, \hat{\theta}(r))}{\partial r} + \frac{\partial \bar{\Pi}(r, \hat{\theta}(r))}{\partial \theta} \hat{\theta}'(r). \tag{9A.45}$$

The first term on the R.H.S. of the previous equality is always negative, indicating as it does that the expected firm's profits lower with the cost of capital,

$$\frac{\partial \bar{\Pi}(r, \hat{\theta}(r))}{\partial r} = -B \int_{B(1+r)-C}^{\infty} f(R; \hat{\theta}(r)) dR.$$

The second term in Eq. (9A.45) must be, then, always positive, but because $\frac{\partial \bar{\Pi}(r, \hat{\theta}(r))}{\partial \theta} > 0$, it must be that $\hat{\theta}'(r) > 0$.

A parametric model of credit rationing. Assume that there exists a continuum of projects in $(0, 1)$ and that the gross return on any project i can be either $R_i > 0$ with probability $1 - \theta_i$ or zero with complementary probability, an assumption borrowed from De Meza and Webb (1987). In the absence of borrowing, all projects have the same expected return, i.e., $\bar{R} \equiv (1 - \theta_i) R_i$ for each i. We assume that $R_i > B(1 + r)$ for all i, which implies that $\bar{R} > B(1 + r)$: the expected return on the project is at least as large as that on banking, that is, entrepreneurs "know how" to produce, and they know better than bankers do. The variance of the project is $\theta_i (1 - \theta_i) R_i^2 = \bar{R}^2 \frac{\theta_i}{1-\theta_i}$, which is increasing in the probability of project failure θ_i; a change in θ_i is a mean-preserving spread.

The expected return to any firm that carries a project with risk θ_i is

$$\begin{aligned}\bar{\Pi}(r, \theta_i) \equiv E_{\theta_i}(\Pi) &= (-C)\theta_i + (R_i - B(1+r))(1 - \theta_i) \\ &= \bar{R} - B(1+r) + (B(1+r) - C)\theta_i,\end{aligned}$$

and parallels the expected return in the continuous case (see Eq. (9A.44)). We require that $C < B(1 + r)$ (which parallels the "convexity" property in the main text), such that $\frac{\partial \bar{\Pi}(r,\theta)}{\partial \theta} > 0$, and we also have $\frac{\partial \bar{\Pi}(r,\theta)}{\partial r} = -B(1 - \theta) < 0$, provided that $\theta < 1$. Therefore, and by the same reasoning as that underlying Eq. (9A.45), the credit market has adverse selection in that $\hat{\theta}'(r) > 0$, provided the cutoff function $\hat{\theta}(r) < 1$. We provide regularity conditions on $\hat{\theta}(r)$ below.

We slightly generalize the analysis in the main text, assuming that the cutoff function $\hat{\theta}(r)$ satisfies $(1 + r_f) W = \bar{\Pi}(r, \hat{\theta}(r))$, where r_f is the return on a safe asset, and W is the wealth of the investor, which we take equal to the collateral that he posts, $W = C$. We shall explain the rationale of this assumption in a moment. Under this assumption, the cutoff function satisfies

$$\hat{\theta}(r) = \frac{C(1 + r_f) + B(1 + r) - \bar{R}}{B(1 + r) - C}, \quad \bar{R} \in (C(2 + r_f), B(1 + r) + C(1 + r_f)),$$

where the condition on $\bar{R}$ ensures that $\hat{\theta}(\cdot) \in (0,1)$. The economic meaning of the lower bound on $\bar{R}$ is that the project is "productive," in that it is on average higher than the "unproductive collateral." Under this condition, $\hat{\theta}(\cdot) < 1$: intuitively, progressively more risky entrepreneurs apply for a loan as the wedge between $\bar{R}$ and C tightens. The upper bound ensures that sufficiently low risky projects (i.e., θ_i small) cannot even break-even. The assumption the entrepreneur has access to a safe technology is made to guarantee that, still, $\bar{R} > B(1+r)$, as discussed above.

We assume that θ is drawn from a distribution with a given density $b(\theta)$, such that the posterior density, i.e., the density the bank estimates based on the adverse selection outcome ($\theta \geq \hat{\theta}(r)$), is

$$b_r(\theta) \equiv \frac{b(\theta)}{\int_{\hat{\theta}(r)}^{1} b(\theta)\, d\theta}, \quad \theta \in (\hat{\theta}(r), 1).$$

Conditioning on a given risk θ, the bank's expected profits are $\bar{\Pi}_b(r;\theta) \equiv E_\theta(\Pi_b)$, where Π_b is as in Eq. (9.131) of the main text. That is,

$$\bar{\Pi}_b(r;\theta) = C\theta + B(1+r)(1-\theta),$$

such that the posterior profits for the bank are

$$\bar{\Pi}_b(r) \equiv \int_{\hat{\theta}(r)}^{1} \bar{\Pi}_b(r;\theta)\, b_r(\theta)\, d\theta.$$

We set $B = 100$, $C = 50$, $\bar{R} = 160$ and $r_f = 0.03$. Regarding $b(\cdot)$, we consider a simple situation, which parallels one in Stiglitz and Weiss (1981), in which there are only two types, i.e., $\theta = \theta_l = 0.20$ with mass $\lambda = 0.95$, and $\theta = \theta_u = 0.30$ with mass $1-\lambda$. The posterior profits for the bank are

$$\bar{\Pi}_b(r) = \begin{cases} C\bar{\theta} + B(1+r)\left(1-\bar{\theta}\right) & \text{for all } r : \hat{\theta}(r) \leq \theta_l \\ C\theta_u + B(1+r)(1-\theta_u) & \text{for all } r : \theta_l < \hat{\theta}(r) \leq \theta_u \\ 0 & \text{for all } r : \theta_u < \hat{\theta}(r) \end{cases}$$

where $\bar{\theta} = \lambda\theta_l + (1-\lambda)\theta_u$, and the loans demand schedule is

$$L^d(r) = \begin{cases} B & \text{for all } r : \hat{\theta}(r) \leq \theta_l \\ B(1-\lambda) & \text{for all } r : \theta_l < \hat{\theta}(r) \leq \theta_u \\ 0 & \text{for all } r : \theta_u < \hat{\theta}(r) \end{cases}$$

The top panel of Figure 9A.1 depicts the banks' posterior returns, defined as $\rho^b(r) \equiv \frac{\bar{\Pi}_b(r)}{B} - 1$, as a function of the loan rate r. Note that $\rho^b(r)$ achieves a local maximum at $r = 23.10\%$, which is the threshold level of the loan rate after which the good applicants drop out of the market. For values of the loan rate higher than the threshold, the banks' posterior returns experience a discrete drop, reflecting the deterioration of the borrowers' creditworthiness. In other words, if the loan rate had to slightly increase from below 23.10%, the banks' posterior returns would plummet by more than 6 percentage points as a result of adverse selection. This extreme prediction is due to highly asymmetric distribution of loan applicants.

The bottom panel of Figure 9A.1 plots supply and demand of loans, where we assume that the loans supply satisfies: $L^s(r) \equiv 0.9\bar{\Pi}_b(r)$. We justify this functional form by positing that the loans supply is an increasing function of $\rho^b(r)$, notably $L^s(r) = 0.9\left(1+\rho^b(r)\right)B$. The properties of the loans supply are obviously inherited by those of the banks' posterior returns and imply the existence of a *local* maximum at about 23%. However, the *global* maximum is achieved at $r = 33.57\%$, which is the loan rate at which type θ_u loses his interest as a potential borrower. In this case, $\rho^b(r^*) = 8.5\%$. The previous characterization of the global maximum does not depend on parameter values and shall be formally proved below.

The fact that the global maximum is corner is not a merely technical detail. There would be credit rationing if the banks' space choice did not include loan rates in a neighborhood of $r^* = 33.57\%$, for example, all rates $r < 30\%$. In this case, the banks' profits would be maximized at the local maximum, $r = 23.10\%$, which would imply credit rationing. Naturally, it is problematic to justify these constraints. At least theoretically, a better route to credit rationing would perhaps be one arising assuming banks face Bertrand competition. Suppose a bank fixes a loan rate equal to r^* to maximize its profits. Then, competitors would lower loan rates to attract borrowers and the process would continue up to the point where the banks' expected profits equal the operating costs. For example, if the operaring costs absorbe 3.5% of the posterior profits, there would be two equilibria (the loan rates at which the posterior profits intersect with the 3.5% line in the top panel of Figure 9A.1). In this example, credit rationing would arise in both equilibria. The assumption of Bertrand competition seems to be empirically questionable, however.

The difficulty to ensure that an interior maximum of the banks' posterior profits is also a global maximum is not specific to this model, but a a specific case of results known since at least Riley (1987). Let us characterize r^* in the example of this appendix. First, note, obviously, that $\bar{\Pi}_b(r;\theta_i) = \bar{R} - \bar{\Pi}(r;\theta_i)$, $i \in \{l,u\}$, and that $\bar{\Pi}(r;\theta_i) \geq C(1+r_f)$ for all $\theta_i \geq \hat{\theta}(r)$, with a strict inequality for $\theta_i = \theta_u$ and $r < r_*$, where r_* is defined as the loan rate at which the riskiest firm (type θ_u) ceases to demand funds, $\hat{\theta}(r_*) = \theta_u$. Therefore, $E(\bar{\Pi}(r;\theta)\big|\,\theta_i \geq \hat{\theta}(r)) > C(1+r_f)$ for $r < r_*$, such that the posterior profits for the bank are, for a given $r < r_*$,

$$E(\bar{\Pi}_b(r;\theta_i)\big|\,\theta_i \geq \hat{\theta}(r)) = \bar{R} - E(\bar{\Pi}(r;\theta)\big|\,\theta_i \geq \hat{\theta}(r)) \leq \bar{R} - C(1+r_f),$$

where the equality holds because the expected return $\bar{R}$ on a project is the same, independent of any risk. However, we have that $\bar{\Pi}(r_*;\theta_u) = \bar{\Pi}(r_*;\hat{\theta}(r_*)) = C(1+r_f)$, such that the previous weak inequality holds as an equality for $r = r^*$. Therefore, the unique global maximum is achieved by $r = r^*$.

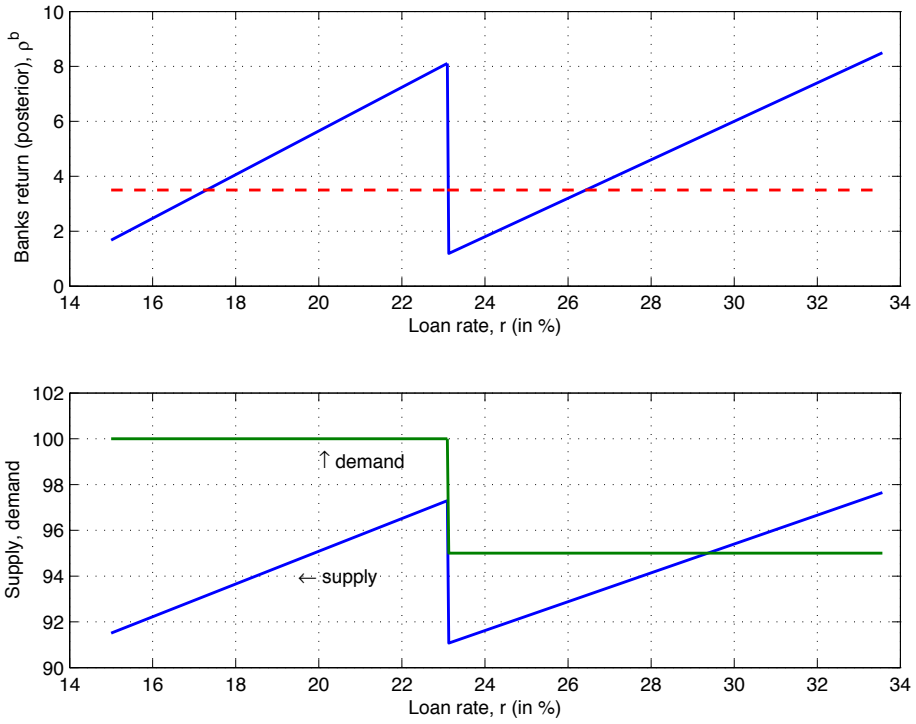


FIGURE 9A.1. Adverse selection and credit markets.

References

Abel, A.B. (1990): "Asset Prices under Habit Formation and Catching Up with the Joneses." *American Economic Review Papers and Proceedings* 80, 38-42.

Abel, A.B. (1999): "Risk Premia and Term Premia in General Equilibrium." *Journal of Monetary Economics* 43, 3-33.

Adrian, T. and H. S. Shin (2011): "Financial Intermediaries and Monetary Economics." In B. M. Friedman and M. Woodford (Editors): *Handbook of Monetary Economics* (North-Holland Elsevier), Vol 3A, Chapter 12, 601-650.

Alvarez, F. and U.J. Jermann (20??):

Bansal, R. and A. Yaron (2004): "Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles." *Journal of Finance* 59, 1481-1509.

Basak, S. (2000): "A Model of Dynamic Equilibrium Asset Pricing with Heterogeneous Beliefs and Extraneous Risk." *Journal of Economic Dynamics and Control* 24, 63-95.

Basak, S. (2005): "Asset Pricing with Heterogeneous Beliefs." *Journal of Banking and Finance* 29, 2849-2881.

Basak, S. and D. Cuoco (1998): "An Equilibrium Model with Restricted Stock Market Participation." *Review of Financial Studies* 11, 309-341.

Bernanke, B.S. (1983): "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression." *American Economic Review* 73, 257-276.

Bernanke, B.S. and M. Gertler (1989): "Agency Costs, Net Worth, and Business Fluctuations." *American Economic Review* 79, 14-31.

Bernanke, B.S. (2004): The Great Moderation: Remarks by the Federal Reserve Board Governor, The Meetings of the Eastern Economic Association, Washington, DC, February 20.

Bernanke, B. S., M. Gertler and S. Gilchrist (1999): "The Financial Accelerator in a Quantitative Business Cycle Framework." In J.B. Taylor and M. Woodford (Editors): *Handbook of Macroeconomics* (North-Holland Elsevier), Vol. 1C, Chapter 21, 1341-1393.

Berrada, T. (2006): "Incomplete Information, Heterogeneity, and Asset Pricing." *Journal of Financial Econometrics* 4, 136-160.

Black, F. (1976): "Studies of Stock Price Volatility Changes." *Proceedings of the 1976 Meeting of the American Statistical Association*, 177-81.

Boldrin, M., L. Christiano and J. Fisher (2001): "Habit Persistence, Asset Returns and the Business Cycle." *American Economic Review* 91, 149-166.

Borio, C., C. Furfine, and P. Lowe (2001): "Procyclicality of the Financial System and Financial Stability: Issues and Policy Options." Bank for International Settlements working paper no. 1.

Brunnermeier, M. K., T. M. Eisenbach and Y. Sannikov (2012): "Macroeconomics with Financial Frictions: A Survery." Working Paper Princeton University.

Brunnermeier, M. K. and Y. Sannikov (2013): "A Macroeconomic Model with a Financial Sector." Working Paper Princeton University.

Buraschi, A. and A. Jiltsov (2006): "Model Uncertainty and Option Markets with Heterogeneous Beliefs." *Journal of Finance* 61, 2841-2897.

Campbell, J.Y., A. W. Lo and C. MacKinlay (1997): *The Econometrics of Financial Markets*. Princeton: Princeton University Press.

Campbell, J.Y. (2003): "Consumption-Based Asset Pricing." In Constantinides, G. M., M. Harris and R. M. Stulz (Editors): *Handbook of the Economics of Finance* (North-Holland Elsevier), Vol 1B, Chapter 13, 803-887.

Campbell, J.Y., and J.H. Cochrane (1999): "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior." *Journal of Political Economy* 107, 205-251.

Campbell, J.Y. and R. Shiller (1988): "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors." *Review of Financial Studies* 1, 195–228.

Chan, Y.L. and L. Kogan (2002): "Catching Up with the Joneses: Heterogeneous Preferences and the Dynamics of Asset Prices." *Journal of Political Economy* 110, 1255-1285.

Christie, A.A. (1982): "The Stochastic Behavior of Common Stock Variances: Value, Leverage, and Interest Rate Effects." *Journal of Financial Economics* 10, 407-432.

Cochrane, J. H., F. A. Longstaff, and P. Santa-Clara (2008): "Two Trees." *Review of Financial Studies* 21, 347-385.

Constantinides, G.M. and D. Duffie (1996): "Asset Pricing with Heterogeneous Consumers." *Journal of Political Economy* 104, 219-240.

Constantinides, G.M., J.B. Donaldson and R. Mehra (2002): "Juniors Can't Borrow: a New Perspective on the Equity Premium Puzzle." *Quarterly Journal of Economics* 117, 269-296.

Cujean, J. and M. Hasler (2001): "Fear of Recessions, Heterogenous Beliefs, and Stock Price Under/Over-Reaction." Working Paper Swiss Finance Institute EPFL.

Cvitanić, J., and S. Malamud (2011): "Price Impact and Portfolio Impact." *Journal of Financial Economics* 100: 201-225.

Danielsson, J., J.-P. Zigrand and H. S. Shin (2011): "Balance Sheet Capacity and Endogenous Risk." Working Paper London School of Economics and Princeton University.

De Meza, D. and D.C. Webb (1987): "Too Much Investment: A Problem of Asymmetric Information." *Quarterly Journal of Economics* 102, 281-292.

Detemple, J. and Murthy, S. (1994): "Intertemporal Asset Pricing with Heterogeneous Beliefs" *Journal of Economic Theory* 62, 294-320.

Dow, J. and S. Werlang (1992): "Uncertainty Aversion, Risk Aversion, and the Optimal Choice of Portfolio." *Econometrica* 60, 197-204.

Duffie, D. (1992): "The Nature of Incomplete Security Markets." In J-J Laffont (Editor): *Advances In Economic Theory, 6th World Congress*, Vol. II, Chapter 4, 214-262.

Duffie, D. and L.G. Epstein (1992a): "Asset Pricing with Stochastic Differential Utility." *Review of Financial Studies* 5, 411-436.

Duffie, D. and L.G. Epstein (with C. Skiadas) (1992b): "Stochastic Differential Utility." *Econometrica* 60, 353-394.

Dumas, B., A. Kurshev and R. Uppal (2009): "Equilibrium Portfolio Strategies in the Presence of Sentiment Risk and Excess Volatility." *The Journal of Finance* 64, 579-629.

Ellsberg, D. (1961): "Risk, Ambiguity and the Savage Axioms." *Quarterly Journal of Economics* 75, 643-69.

Epstein, L.G. and S.E. Zin (1989): "Substitution, Risk-Aversion and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework." *Econometrica* 57, 937-969.

Epstein, L.G. and S.E. Zin (1991): "Substitution, Risk-Aversion and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis." *Journal of Political Economy* 99, 263-286.

Fisher, I. (1933): "The Debt-Deflation Theory of Great Depressions." *Econometrica* 1, 337-57.

Friedman, M. (1953): *The Case for Flexible Exchange Rates*—Essays in Positive Economics. Chicago: University of Chicago Press.

Friedman, M. and A. J. Schwartz (1963): *A Monetary History of the United States: 1867-1960.* Princeton, NJ: Princeton University Press.

Gallmeyer, M., Aydemir, A.C. and B. Hollifield (2007): "Financial Leverage and the Leverage Effect: A Market and a Firm Analysis." working paper Carnegie Mellon.

Geanakoplos, J. (2010): "The Leverage Cycle." In D. Acemoglu, K. Rogoff and M. Woodford (Editors): *NBER Macroeconomic Annual 2009* (University of Chicago Press), Vol 24, 1-65.

Gertler, M. and N. Kiyotaki (2011): "Financial Intermediation and Credit Policy in Business Cycle Analysis." In B. M. Friedman and M. Woodford (Editors): *Handbook of Monetary Economics* (North-Holland Elsevier), Vol 3A, Chapter 11, 547-599.

Gilboa, I. and M. Marinacci (2011): "Ambiguity and the Bayesian Paradigm." In *Advances in Economics and Econometrics: Theory and Applications*, Tenth World Congress of the Econometric Society. Cambridge: Cambridge University Press.

Gilboa, I. and D. Schmeidler (1989): "Maxmin Expected Utility with a Non-Unique Prior." *Journal of Mathematical Economics* 18, 141-153.

Guvenen, F. (2009): "A Parsimonious Macroeconomic Model for Asset Pricing." *Econometrica* 77, 1711-1740.

Harrison, J.M. and D.M. Kreps (1978): "Speculative Investor Behavior in a Stock Market with Heterogeneous Expectations." *Quarterly Journal of Economics* 92, 323-36.

Hansen, L. P. and T. J. Sargent (2008): *Robustness*. Princeton: Princeton University Press.

He, Z. and A. Krishnamurthy (2012): "Intermediary Asset Pricing." Forthcoming in the *American Economic Review*.

Heaton, J. and D.J. Lucas (1996): "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing." *Journal of Political Economy* 104, 443-487.

Huang, C.-f. (1987): "An Intertemporal General Equilibrium Asset Pricing Model: the Case of Diffusion Information." *Econometrica* 55, 117-142.

Hugonnier, J. and R. Prieto (2012): "Arbitrageurs, Bubbles, and Credit Conditions." Working paper SFI- EPFL (Lausanne) and Boston University.

Jermann, U.J. (1998): "Asset Pricing in Production Economies." *Journal of Monetary Economics* 41, 257-276.

Karatzas, I. and S.E. Shreve (1991): *Brownian Motion and Stochastic Calculus*. New York: Springer Verlag.

Keynes, J.M. (1921): *A Treatise on Probability*. London: MacMillan and Co.

Kiyotaki, N. (1998): "Credit and Business Cycles." *Japanese Economic Review* 49, 18-35.

Kiyotaki, N. and J. Moore (1997): "Credit Cycles." *Journal of Political Economy* 105, 211-248.

Klibanoff, P., M. Marinacci, and S. Mukerji (2005): "A Smooth Model of Decision Making under Ambiguity." *Econometrica* 73, 1849-1492.

Knight, F.H. (1921): *Risk, Uncertainty, and Profit*. New York: Houghton Mifflin.

Kogan, L., S. Ross, J. Wang, and M. Westerfield (2006): "The Survival and Price Impact of Irrational Traders." *Journal of Finance* 61, 195-229.

Kyle, A.S. and F.A. Wang (1997): "Speculation Duopoly with Agreement to Disagree: Can Overconfidence Survive the Market Test?" *Journal of Finance* 52, 2073-90.

Leippold, M., F. Trojani and P. Vanini (2008): "Learning and Asset Prices under Ambiguous Information." *Review of Financial Studies* 21, 2565-2597.

Lettau, M. (2002): "Idiosyncratic Risk and Volatility Bounds, or, Can Models with Idiosyncratic Risk Solve the Equity Premium Puzzle?" *Review of Economics and Statistics* 84, 376-380.

Liptser, R.S. and A.N. Shiryaev (2001): *Statistics of Random Processes–Vol. II (Applications)*. Berlin, Springer-Verlag.

Lucas, R.E. (19??):

Lucas, D.J. (1994): "Asset Pricing with Undiversifiable Income Risk and Short Sales Constraints: Deepening the Equity Premium Puzzle." *Journal of Monetary Economics* 34, 325-341.

Mankiw, N.G. (1986): "The Equity Premium and the Concentration of Aggregate Shocks." *Journal of Financial Economics* 17, 211-219.

Mankiw, N.G. and S.P. Zeldes (1991): "The Consumption of Stockholders and Non-Stockholders." *Journal of Financial Economics* 29, 97-112.

Markovitz, H. (1952): "Portfolio Selection." *Journal of Finance* 7, 77-91.

Martin, I. (2011): "The Lucas Orchard." Working Paper Stanford University.

Mele, A. and F. Sangiorgi (2015): "Uncertainty, Information Acquisition and Price Swings in Asset Markets." *Review of Economic Studies* 82, 1533-1567.

Menzly, L., T. Santos and P. Veronesi (2004): "Understanding Predictability." *Journal of Political Economy* 112, 1, 1-47.

Nelson, D.B. (1991): "Conditional Heteroskedasticity in Asset Returns: A New Approach." *Econometrica* 59, 347-370.

Odean, T. (1998): "Volume, Volatility, Price, and Profit When All Traders Are Above Average." *Journal of Finance* 53, 1887-1934.

Pavlova, A. and R. Rigobon (2008): "The Role of Portfolio Constraints in the International Propagation of Shocks." *Review of Economic Studies* 75, 1215-1256.

Rawls, J. (1971): *A Theory of Justice*. Cambridge: Harvard University Press.

Riley, J.G. (1987): "Credit Rationing: A Further Remark." *American Economic Review* 77, 224-227.

Rouwenhorst, G. K. (1995): "Asset Returns and Business Cycles." In Cooley, T.F. (Ed.): *Frontiers of Business Cycle Research*, Princeton University Press, 294-330.

Scheinkman, J.A. and W. Xiong (2002): "Overconfidence and Speculative Bubbles." *Journal of Political Economy* 111, 1183-1219.

Schmeidler, D. (working paper, 1982, published in 1989): "Subjective Probability and Expected Utility without Additivity." *Econometrica* 57, 571-587.

Schmeidler, D. (1986): "Integral Representation without Additivity." *Proceedings of the American Mathematical Society* 97, 255-261.

Schwert, G.W. (1989a): "Why Does Stock Market Volatility Change Over Time?" *Journal of Finance* 44, 1115-1153.

Schwert, G.W. (1989b): "Business Cycles, Financial Crises and Stock Volatility." *Carnegie-Rochester Conference Series on Public Policy* 31, 83-125.

Shin, H. S. (2010): *Risk and Liquidity*. Clarendon Lectures in Finance. Oxford: Oxford University Press.

Stiglitz, J.E. and A. Weiss (1981): "Credit Rationing in Markets with Imperfect Information." *American Economic Review* 71, 393-410.

Tallarini, T. (2000): "Risk-Sensitive Real Business Cycles." *Journal of Monetary Economics* 45, 507-32.

Telmer, C.I. (1993): "Asset-Pricing Puzzles and Incomplete Markets." *Journal of Finance* 48, 1803-1832.

Veronesi, P. (2000): "How Does Information Quality Affect Stock Returns?" *Journal of Finance* 55, 807-837.

Wachter, J.A. (2006): "A Consumption-Based Model of the Term Structure of Interest Rates." *Journal of Financial Economics* 79, 365-399.

Wald, A. (1950): *Statistical Decision Functions*. New York: John Wiley.

Weil, Ph. (1989): "The Equity Premium Puzzle and the Risk-Free Rate Puzzle." *Journal of Monetary Economics* 24, 401-421.

Weil, Ph. (1992): "Equilibrium Asset Prices with Undiversifiable Labor Income Risk." *Journal of Economic Dynamics and Control* 16, 769-790.

Xiouros, C. and F. Zapatero (2010): "The Representative Agent of an Economy with External Habit Formation and Heterogeneous Risk Aversion." *Review of Financial Studies* 23, 3017-3047.

Zapatero, F. (1998): "Effects of Financial Innovations on Market Volatility when Beliefs are Heterogeneous." *Journal of Economic Dynamics and Control* 22, 597-626.

10
Information and other market frictions

10.1 Introduction

In the economies of the previous chapters, the equilibrium outcomes do not convey more information than that available to each agent because information, whilst sometimes incomplete, is disseminated *symmetrically* across decision makers—for example, asset prices aggregate information that the agents already know, such that the agents' inference on the asset fundamentals would not improve were it also based on the equilibrium prices.

This chapter considers markets in which equilibrium outcomes aggregate information dispersed across the agents, which the very same agents find useful while updating beliefs. This information is useful because the pieces of information agents have access to are not the same, such that, now, asset prices contain information about the fundamentals that some agents might not directly observe, and which are made publicly available (so to speak) through trading activity. We study markets with *asymmetric* information, in which some agents have more precise information than others, and markets with *differential* information, in which agents know different pieces of information that have the same quality. In the economies of the previous chapters, the price only determines the budget constraints. While the price still determines budget constraints in the markets that we analyze in this chapter, this price now plays an additional, fundamental role: it conveys information to investors.

Note how subtle the equilibrium concept needs to be in the markets of this chapter. If agents find it useful to condition their choices upon the information conveyed by an equilibrium outcome, these very same choices may well affect the information contained in the equilibrium outcome, over a fixed point. Provided it exists, this fixed point leads to implications regarding the informational role of asset prices, that is, the equilibrium amount of information. In the asymmetric information case, we say that the price "transmits" information from the more informed investors to the less. In the differential information case, we say that the price "aggregates" information dispersed amongst investors. Both cases play outstanding roles in financial economics.

Consider the asymmetric information case: if uninformed investors can learn from the equilibrium price, what are the incentives left to purchase information? In other words, what is the value of information? The differential information case is equally important. The welfare

theorems reviewed in Chapter 2 suggest that a Pareto optimal allocation can be "centralized." However, this solution cannot be implemented when portfolio decisions are made on the basis of "local" information. The market solution proves useful as the price would now aggregate dispersed information, making it available to investors initially not having direct access to it. These ideas go back to Hayek (1945) at least.

There is indeed a number of conceptual issues that arise in these markets. Namely, how much information does a price need to convey for an equilibrium to exist? Intuitively, if the asset price conveys too precise information, and becomes a sufficient statistics for the overall information available in the marketplace, the agents would not even need to condition upon their own information while formulating portfolio decisions. But then, if agents trade while only relying on the equilibrium price (and not on their own signals), how can, then, the price aggregate information? This is the Grossman (1976) paradox. Furthermore, it is intuitive that the price cannot reveal all private information when information is costly, for otherwise there would not be incentives left to purchase information in the first place: it is the Grossman and Stiglitz (1980) paradox.

A standard approach to deal with these conceptual difficulties is to assume that prices convey "noisy" signals about the information that investors have. As Black (1986) discussed, noise makes markets function when information problems would otherwise lead them not to arise in the first place. The mechanism is simple. In equilibrium, the price incorporates the information on which informed investors trade, but also other factors that are possibly unrelated to the fundamentals, such as liquidity shocks—"noise." In this case, an uninformed investor cannot tell whether a large asset price swing is due to information or to a liquidity shock, because now the price only *partially reveals* the information informed investors have. So uninformed investors learn from the price, but the information conveyed by the price is imperfect, such that informed investors still have superior information, even after the uninformed learning. Thus, it can pay to purchase information. All in all, *partial* revelation of information is the key to make asset markets work in this context.

The equilibrium concept to deal with these markets is an extension of the Rational Expectations Equilibrium (REE) the previous chapters have relied upon, and is called Noisy REE (NREE). It appears that macroeconomics contains the first example of an an equilibrium in which agents confuse fundamentals with noise. In an attempt to explain the relation between the conduct of monetary policy and the business cycle, Phelps (1970) and Lucas (1972) consider an economy in which agents have imperfect information regarding the fundamentals of the economy. This information-based approach to the business cycle is summarized in Lucas (1981), and was somehow abandoned in favor of the real business cycle theory (see Chapter 3), perhaps because it may be unrealistic to hypothesize that information is the main engine of macroeconomic fluctuations. However, the interplay between macroeconomics and information helped Lucas (1972) develop the notion of rational expectations within a "noisy" economy. In Section 10.2, we present a simplified version of the Lucas framework, which helps pave the way to the study of asset markets in subsequent sections.

A central theme of this chapter is the role of information in asset markets. In many of the models of this chapter, noise is what makes these markets work, as explained. Liquidity shocks are the most natural example that illustrates this concept. In fact, the models in this chapter make sharp predictions on how liquidity shocks affect asset prices. For example, we shall explain, in non-competitive markets (markets beyond the standard NREE), uninformed trades may have a price impact because market makers confuse information with noise: being unsure about the nature of the orders they see (information-driven or liquidity shocks), market makers price the

assets in a way that even a liquidity shock (noise) gets impounded in equilibrium. It is a clear case of adverse selection: the prices set by market makers attract the "wrong trades," in that they will be potentially buying from those who hold bad news about the asset and selling to those who hold good news. However, these markets work because the market makers are not sure that each of the orders they receive is informed or noise. But then, liquidity motivated trades are hurt because market makers would set prices taking into account the fact they might not be such. Liquidity in asset markets therefore constitutes the other side of the same coin (against information) in this chapter.

The core of this chapter is then to analyze how information affects asset markets. Section 10.3 discusses the classical notions of informational efficiency in the early empirical literature, and how the subsequent information literature has helped shape these notions. This literature is reviewed next. First, Sections 10.4 and 10.5 explain that the standard notions of a Walrasian equilibrium or REE are not sufficient to deal with markets in which agents have asymmetric or diverse information. Section 10.6 then deals with NREE. Sections 10.7 and 10.8 cover markets with non-competitive players.

While information does play a fundamental role to explain prices and liquidity in the "microstructure" of asset markets, information cannot be the only driver of market liquidity. Liquidity in markets driven by macroeconomic news (government bonds, for example) cannot be only driven by investors with superior or diverse information. "Search" and "bargaining" are alternative explanations in these markets. Section 10.9 studies liquidity in markets where information plays a more limited role, with a focus shifted to the search nature of OTC markets.

Section 10.10 concludes the chapter and examines a number of additional mechanisms that could potentially affect the asset price formation process, such as the presence of irrational (or noise) traders and additional capital market imperfections that lead to limits to arbitrage.

10.2 Prelude: imperfect information in macroeconomics

We consider a simplified version of the model developed by Lucas (1973), in which n goods are produced in n distinct "islands." Let y_i^s denote log-production supplied in the i-th island. (All variables are in logs, in this section.) Below, we explain which sources of randomness affect this economy (see Eqs. (10.3) and (10.4)). It is assumed that production supplied is set so as to equal the expected wedge of the (log-) price in the island, p_i, over the average price in the economy, p,

$$y_i^s = E\left(p_i - p | \, p_i\right), \quad \text{where} \quad p = \frac{1}{n}\sum_{j=1}^{n} p_j. \tag{10.1}$$

Eq. (10.1) follows, approximately, once we assume that the average price, p, is common knowledge, as for example in the monopolistic competition model of Blanchard and Kiyotaki (1987). When p is not common knowledge as in the analysis of this section, Eq. (10.1) can still be thought of as arising through a plausible decision mechanism. The specific functional form for the average price is the most important approximation made while deriving Eq. (10.1) based on a rigorous micro-founded setup. Appendix 4 to this chapter contains a discussion of these issues.

Information is disseminated differentially (not asymmetrically), in that producers in the i-th island are not aware of the price in the remaining islands, and make statistical inference on economic developments occurring in the other islands with the same precision. We conjecture

and, later, verify, that all variables, exogeneous and endogeneous, are normally distributed. We shall show that this normality property implies the price index p gathers all information efficiently, i.e. p is a sufficient statistics for all that information.

By the projection theorem reviewed in Appendix 1 to this chapter, we have that:

$$y_i^s = E\left(p_i - p | \, p_i\right) \equiv \beta\left(p_i - E\left(p\right)\right),$$

where we have used the fact that information is symmetrically disseminated and, then, (i) the expectation $E\left(p_i\right) = E\left(p_j\right) = E\left(p\right)$ for every i and j, and (ii) both the numerator and denominator of the ratio, $\beta \equiv \frac{cov(p_i - p, p_i)}{var(p_i)}$, are the same across all islands. This coefficient will be determined below, as a part of the equilibrium.

Aggregating across all islands, yields the celebrated *Lucas supply* equation:

$$y^s \equiv \frac{1}{n}\sum_{j=1}^{n} y_j^s = \beta\left(p - E\left(p\right)\right). \tag{10.2}$$

Next, assume the demand for the good produced in the i-th island is given by:

$$y_i^d = m - p + u_i - \theta\left(p_i - p\right), \quad \text{where} \quad u_i \sim N\left(0, \sigma_u^2\right), \tag{10.3}$$

and m is money supply, which we assume to satisfy:

$$m = E\left(m\right) + \epsilon, \quad \text{where} \quad \epsilon \sim N\left(0, \sigma_\epsilon^2\right). \tag{10.4}$$

Finally, we assume that $E\left(u_i \epsilon\right) = 0$, and that u_i are sectoral shocks in that: $\sum_{j=1}^{n} u_j = 0$. The functional form for the demand function, y_i^d, follows after assuming the goods in the islands are imperfect substitutes (see Appendix 3 to this chapter).

The equilibrium price in the islands plays two roles in this economy. A first, and standard role, is to clear the markets, being such that $y_i^s = y_i^d$, or:

$$\beta\left(p_i - E\left(p\right)\right) = m - p + u_i - \theta\left(p_i - p\right), \quad \text{for all } i. \tag{10.5}$$

The second role the price performs is to convey information to agents regarding the two shocks: (i) the macroeconomic, monetary shock, ϵ; and (ii) the real shocks occurred in all the islands, u_j, $j = 1, \cdots, n$. We conjecture that the only real shock that determines the price in the i-th island is u_i, i.e., that the price is a function $p_i \equiv p\left(\epsilon, u_i\right)$. We also conjecture this price is affine, in ϵ and u_i, viz

$$p\left(\epsilon, u_i\right) = a + b\epsilon + cu_i, \tag{10.6}$$

where the coefficients a, b and c have to be determined in equilibrium. Under these conditions, the average price is a function $p \equiv \bar{p}\left(\epsilon\right)$ satisfying

$$\bar{p}\left(\epsilon\right) = a + b\epsilon. \tag{10.7}$$

Let us replace Eqs. (10.6), (10.7) and (10.4) into the equilibrium condition, Eq. (10.5). By rearranging terms,

$$0 = \left(\beta b + b - 1\right)\epsilon + \left(\beta c + \theta c - 1\right)u_i + a - E\left(m\right).$$

The previous equation has to hold for all ϵ and u_i. Therefore,

$$a = E\left(m\right),$$

and the coefficients for ϵ and u_i must both equal zero, leading to the following expressions for b and c:

$$b = \frac{1}{1+\beta}, \quad c = \frac{1}{\theta+\beta}. \tag{10.8}$$

We are left with determining β, which given Eqs. (10.6)-(10.7), and Eq. (10.8), is easily shown to equal:

$$\beta = \frac{cov\left(p_i - p, p_i\right)}{var\left(p_i\right)} = \frac{\sigma_u^2}{\sigma_u^2 + \left(\frac{\theta+\beta}{1+\beta}\right)^2 \sigma_\epsilon^2}. \tag{10.9}$$

The previous equation has a unique and positive fixed point for β, which can then be replaced back into Eqs. (10.8), yielding the solutions for b and c, which are both positive.

We can now figure out the implications of this equilibrium. By replacing Eqs. (10.6)-(10.7) into the Lucas supply equation (10.2), leaves:

$$y^s = \beta b \epsilon.$$

This is Lucas celebrated neutrality result. Anticipated monetary policy, $E\left(m\right)$, does not affect the equilibrium outcome, y^s. It is the monetary shock that affects y^s. Agents in any island do not observe the price in the remaining islands and, hence, the aggregate price level, p. Therefore, they are unable to tell whether an increase in the price of the good they produce, p_i, is due to a real shock, u_i, or to a monetary shock, ϵ. In other words, they cannot disentangle a monetary shock from a real shock. If the agents were informed about real shocks, they would of course infer ϵ, and a monetary shock would not exert any effect on the equilibrium production.

In other words, in equilibrium, the price difference is $p_i - p = cu_i$, which does not depend on ϵ. It is a "dichotomy" prediction reminiscent of the classical theory. Note, however, that $p_i - p$ is not observed, as p is not observed. Instead, the producers in the i-th island can only make inference on,

$$p_i - E\left(p \middle| p_i\right) = \frac{c^2\sigma_u^2}{b^2\sigma_\epsilon^2 + c^2\sigma_u^2}\left(b\epsilon + cu_i\right).$$

The previous term co-varies positively with the observed price, p_i, $cov\left(p_i - p, p_i\right) = c^2\sigma_u^2$. The covariance is zero precisely when the assumption is removed of imperfect knowledge regarding the real shocks, $\sigma_u^2 = 0$, in which case $\beta = 0$. In contrast, and assuming imperfect knowledge, producers act so as to compensate for their partial lack of knowledge, and produce to the maximum extent they can justify, on the basis of the positive statistical co-movements, $cov\left(p_i - p, p_i\right) > 0$. Note, if $E\left(m\right) = m_{-1}$, i.e. money supply in the previous period, then from Eq. (10.7), the inflation rate, $p - p_{-1} = b\epsilon + \left(1-b\right)\epsilon_{-1}$. Therefore, output and inflation are positively correlated, a "Phillips curve," which policy makers cannot exploit anyway, as anticipated monetary policy, $E\left(m\right)$, is rationally "factored out," and does not affect output. This is the essence of the Lucas critique (Lucas, 1977).

In the next sections, we analyze asset markets that work due to similar mechanisms, in the following sense. Why should we ever buy some assets from those people insisting in selling them? Trading seems to be a difficult phenomenon to explain in a world with imperfect information. Yet trading does occur, if imperfect information has the same nature as that of the Phelps-Lucas model. Agents might well be imperfectly informed about the nature of, say, unusually high market orders. For example, sell orders might arrive to the market, either because the asset is a lemon or because the agents selling the assets are hit by a liquidity shock. In the models of this chapter, an equilibrium with rational expectation exists, precisely because of

this "noise"—liquidity, in this example. There is a chance the sell order arrives to the market, simply because the agents selling it are hit by a liquidity shock. Imperfectly informed agents, therefore, might be willing to buy, if it is in their interest to do so.

10.3 Informational efficiency: roadmap

Informational efficiency is an attribute of asset markets that we gauge through equilibrium outcomes, typically the price. It is an old and somehow controversial topic in financial economics. We would say asset markets are informationally efficient if prices reflect available information, accurately and rapidly. This definition is obviously loose, but made purposedly so, as our objective, now, is to illustrate how the process of narrowing it down leads to topics that have made the object of controversy. In particular, we need to qualify the (i) type and (ii) quality of information embedded into and revealed by the price: What type of information does the price reveal? How accurately can the price convey information?

The first question, relating to the type of information, has been addressed in a famous contribution by Fama (1970), who considers three forms of informational efficiency. First, *strong* efficiency, arising when the price reflects all private and public information. Second, *semi-strong* efficiency, the situation in which the price conveys all public information. Third, *weak* efficiency, arising when the price only conveys information regarding past (price) data.

At least initially, the motivation to define "efficient" markets (in the informational sense) was to illustrate that markets cannot be in a state of "disequilibrium." For example, if asset prices tend to be high on Friday and to decline on Monday, a profitable trading opportunity might seem to arise. An equilibrium, the motivation goes, is informationally efficient (weakly so, in this example), should the average return from this opportunity be small enough to discourage trading on it. We know this reasoning has fallacies. Even if the average gain is statistically large, we might have no agent attempting at it, due to risk-aversion or trading frictions. Therefore, we would never know whether even a potentially large gain (in statistical terms) is a market inefficiency or rather, say, compensation for risk, a classical "joint hypothesis" problem.

The theoretical literature has refined the notion of informational efficiency, by shifting its focus on the second question formulated at the beginning of this section: How accurately *can* the price convey information? The approach followed to address this question relies on models in which rational agents obviously trade on their information, but also on the information the price reveals in equilibrium, which depends on the agents' portfolio decisions, over a fixed point, as anticipated in the Introduction. This fixed point can lead to asset prices that are *fully revealing*, in that they reveal all private and public information. These models rely on rational agents and not surprisingly, predict that no money could be ever "left on the table." We also have a refinement of this fully revealing concept: we say markets with fully revealing prices are strongly informationally efficient if the prices reveal a sufficient statistics for all the private information. Note that strong efficiency in this theoretical sense differs from its meaning in the empirical literature, signifying as it does, now, that a simple statistics (say, the average of the signals dispersed across the agents' population) is enough to forecast the asset fundamentals.

As discussed in the Introduction, markets with fully revealing prices are problematic. They lead to paradoxes. First, the Grossman (1976) paradox. If markets are strongly efficient, the agents should abandon their own signals while formulating portfolio decisions, although in this case the price could not, obviously, contain any piece of *private* information, contradicting the initial presumption that markets are strongly efficient.

Second, the Grossman and Stiglitz (1980) paradox. If markets are strongly efficient, informed agents make losses once information is costly, and would rather become "uninformed," free-riding on an informative price, although then, in this case, the price could not, obviously, contain information anymore. To resolve this paradox, Grossman and Stiglitz (1980) propose markets which they famously describe (p. 393) to be ones in which there is an "equilibrium degree of disequilibrium." That is, in their model, prices cannot be fully revealing, but *partially revealing*, meaning that the informed agents do not give all of their information away to the uninformed. "Disequilibrium" means prices are not fully revealing, and "equilibrium degree of disequilibrium" means that the price informativeness depends on how many agents are informed, which is an endogenous variable in the model. Note, now, that "disequilibrium" allows markets to function, a perspective somehow distinct from the early attempts to define efficiency in the empirical literature.

The following three sections aim to formalize these ideas. Section 10.4 shows that money is indeed "left on the table" once agents take portfolio decisions while ignoring the information content of asset prices. Then, we introduce progressively more appropriate notions of equilibrium in which prices are fully (Section 10.5) and partially (Section 10.6) revealing.

10.4 Walrasian equilibria as informationally inefficient outcomes

We consider a two-period market in which in the second period, a riskless asset (the numéraire) yields a gross return equal to R, and a risky asset pays off $d \sim N\left(\bar{d}, \sigma_d^2\right)$. The budget constraint faced by each agent is $W = \left(W_0 - (x - \omega_0)\, p\right) R + xd$, where W_0 is the agent's initial wealth, and ω_0 denotes the initial endowment of the risky asset, p is the price of the risky asset, x is the asset holding and, finally, W is the agent's terminal wealth. We assume that each investor maximizes a CARA expected utility function against terminal wealth, $E\left(-e^{-\tau W} \middle| \mathbb{F}_i\right)$, $\tau > 0$, where $\mathbb{F}_i$ denotes the information set available to agent i.

We derive the equilibrium in this market while assuming asymmetric information, and treat differential information as a special case of this setting. We assume that there is a total of n agents. Of these n agents, n_I are informed, in that they observe a signal on the asset value equal to

$$s^i = d + u^i, \quad i = 1, \cdots, n_I, \tag{10.10}$$

where the noise component of the signal s^i is $u^i \sim N\left(0, \sigma_u^2\right)$, for each investor i.

Let π_d and π_u denote the precisions of the dividend and noise distribution, i.e. $\pi_d \equiv \sigma_d^{-2}$ and $\pi_u \equiv \sigma_u^{-2}$. The demand for the risky asset formulated by any informed investor is

$$\begin{aligned} \hat{x}_I^i &= \arg\max_x E\left(-e^{-\tau W} \middle| s^i\right) \\ &= \arg\max_x \left(-e^{-\tau E\left(W|s^i\right) + \frac{1}{2}\tau^2 var\left(W|s^i\right)}\right) \\ &= \gamma \pi_{d|s^i} \left(E\left(d \middle| s^i\right) - Rp\right), \end{aligned} \tag{10.11}$$

where the second line follows because both d and s^i are normally distributed, and $\gamma \equiv \tau^{-1}$, interpreted as *risk-tolerance*, and by the projection theorem,

$$E\left(d \middle| s^i\right) = \bar{d} + \frac{\pi_u}{\pi_d + \pi_u}\left(s^i - \bar{d}\right), \tag{10.12}$$

and $\pi_{d|s^i}$ denotes the precision of the asset payoff d conditional upon having observed s^i,

$$\pi_{d|s^i} = \pi_d + \pi_u, \quad \text{for all } i. \tag{10.13}$$

There are also n_U uninformed agents, who do not observe the signal s^i in Eq. (10.10). Their demand for the risky asset is $\hat{x}_U^i$, say, and is a special case of the informed asset demand in Eq. (10.11), namely for $\pi_u = 0$,

$$\hat{x}_U \equiv \hat{x}_U^i = \gamma \pi_d \left(\bar{d} - Rp\right). \tag{10.14}$$

The equilibrium price of the risky asset is found by aggregating demand, setting supply equal to aggregate demand,

$$z \equiv n\omega_0 = \sum_{i=1}^{n_I} \hat{x}_I^i + n_U \hat{x}_U, \tag{10.15}$$

and solving for p after replacing Eq. (10.11) and Eq. (10.14) into Eq. (10.15) and using Eqs. (10.13)-(10.12), yielding,

$$p = \frac{\bar{d}}{R} + \frac{\pi_u n_I \left(\hat{s} - \bar{d}\right)}{R\left(\left(\pi_d + \pi_u\right) n_I + \pi_d n_U\right)} - \frac{z}{R\gamma\left(\left(\pi_d + \pi_u\right) n_I + \pi_d n_U\right)}, \tag{10.16}$$

where $\hat{s}$ denotes the average signal, viz

$$\hat{s} \equiv \frac{1}{n_I} \sum_{i=1}^{n_I} s^i. \tag{10.17}$$

The equilibrium price in Eq. (10.16) has three components. The first is the discounted payoff. The second term reveals that the price aggregates information dispersed across the informed investors through the average signal, $\hat{s}$. That is, $\hat{s}$ is a sufficient statistics for all of the signals observed by each informed agent, $(s^i)_{i=1}^{n_I}$, with respect to the equilibrium price. This second term thus adds or subtracts value according to whether the average signal, $\hat{s}$, is higher than the unconditional guess, $\bar{d}$. The third term is a risk-premium: the higher the average supply, the higher the risk the agents have to bear in equilibrium, which is evaluated proportionally to their risk aversion, $\gamma^{-1} = \tau$.

The main issue with a Walrasian equilibrium is that while the price conveys information informed investors have, the uninformed investors do not condition upon this informative price while formulating their asset demand in Eq. (10.14). They only use the price to determine their budget constraint. The resulting equilibrium is conceptually problematic as a result. Consider, for example a market with a large number of agents. In this market, we have that by the Law of Large Numbers,

$$\operatorname*{plim}_{n\to\infty} p = \bar{p} \equiv \frac{\bar{d}}{R} + \frac{\pi_u \bar{\lambda}\left(d - \bar{d}\right)}{R\left(\bar{\lambda}\left(\pi_d + \pi_u\right) + \left(1 - \bar{\lambda}\right)\pi_d\right)} - \frac{\omega_0}{R\gamma\left(\bar{\lambda}\left(\pi_d + \pi_u\right) + \left(1 - \bar{\lambda}\right)\pi_d\right)},$$

where $\bar{\lambda} \equiv \lim_{n\to\infty} \frac{n_I}{n}$. That is, in this limit, the price perfectly reveals the asset payoff, d, provided of course the proportion of informed agents is asymptotically significant, $\bar{\lambda} > 0$.

It is an arbitrage opportunity. Any investor who understands this market (including any uninformed investor!), can make large profits whenever $R\bar{p} \neq d$. He will observe the price $\bar{p}$, and infer d. For example, if $R\bar{p} < d$, he will borrow $\bar{p}$ and use this to invest in the risky asset. In the second period, he will pay back $R\bar{p}$ and receive the asset payoff d, with a sure profit equal to $d - R\bar{p} > 0$.

10.5 Rational Expectations Equilibrium

If the equilibrium price reveals information, any agent will find it convenient to update his knowledge about the asset payoff, d, while conditioning on his information set, comprising both his own signal s^i and the equilibrium price. We term the equilibrium resulting in this market as *rational expectations equilibrium* (REE, henceforth). The term "rational expectations" is used to indicate that the agents correctly understand that the equilibrium asset price incorporates information disseminated across the market, and make statistical inference and take investment decisions accordingly. We shall return to this point in a moment.

We solve for the equilibrium as follows. First, we conjecture the equilibrium price does indeed incorporate information on the average signal, $\hat{s}$, as the equilibrium price does in the Walrasian case in (10.16),

$$p = a + b\left(\hat{s} - \bar{d}\right), \tag{10.18}$$

for two constants a and b. In this market, then, a state of the world is a random variable that is the average signal, and the asset price is a function of this state: it takes values according to the realization of $\hat{s}$, thereby leading to *endogenous beliefs*, or *rational expectations* (see Chapter 2) in that the statistical properties of the equilibrium price are fully consistent with the probability distribution that agents have regarding the very same equilibrium price.

Naturally, the two constants a and b cannot be the same as those in Eq. (10.16): if the agents make inference about the asset payoff while also conditioning on the equilibrium price, this equilibrium price should then differ from that arising in the Walrasian case. To determine the equilibrium, first note that the informed agents formulate a demand function equal to

$$\hat{x}_I^i = \gamma \pi_{\,d|s^i,\hat{s}} \left(E\left(d|\, s^i, \hat{s}\right) - Rp\right). \tag{10.19}$$

This demand schedule generalizes that in Eq. (10.11), in that the conditioning information now contains both the signal available to the informed investor, s^i, and the equilibrium price in (10.18) or, equivalently, and conjecturing that $b \neq 0$, the average signal, $\hat{s}$. However, it is easy to check that, by the projection theorem,

$$E\left(d|\, s^i, \hat{s}\right) = E\left(d|\, \hat{s}\right) = \bar{d} + \frac{n\pi_u}{\pi_d + n\pi_u}\left(\hat{s} - \bar{d}\right), \tag{10.20}$$

and,

$$\pi_{\,d|s^i,\hat{s}} = \pi_{\,d|\hat{s}} = \pi_d + n\pi_u. \tag{10.21}$$

That is, the average signal is a sufficient statistics for the distribution of the asset payoff. We shall discuss the economic implications of this conclusion below, once we will have solved for the equilibrium.

Regarding the uninformed investors, their demand schedule differs from that in (10.14), in that they will update their beliefs about the asset payoff after having learnt the price realization, p in (10.18) or, equivalently, and conjecturing that $b \neq 0$, the average signal, $\hat{s}$. But then note that each uninformed agent will have exactly the same demand schedule as the informed in (10.19), due to Eqs. (10.20)-(10.21), viz,

$$\hat{x}_I^i = \hat{x}_U^j = \gamma \pi_{\,d|\hat{s}} \left(E\left(d|\, \hat{s}\right) - Rp\right), \quad i = 1, \cdots, n_I, \quad j = 1, \cdots, n_U.$$

We determine the coefficients a and b of the equilibrium price in (10.18) by plugging the previous demand schedules into the market clearing condition, Eq. (10.15), leaving,

$$p = \frac{\bar{d}}{R} + \frac{n\pi_u}{R\left(\pi_d + n\pi_u\right)}\left(\hat{s} - \bar{d}\right) - \frac{\omega_0}{R\gamma\left(\pi_d + n\pi_u\right)}. \tag{10.22}$$

In turn, the equilibrium asset allocation is, simply,

$$\hat{x}_I^i = \hat{x}_U^j = \omega_0, \quad i = 1, \cdots, n_I, \quad j = 1, \cdots, n_U.$$

The equilibrium price in (10.22) fully reveals the average information disseminated in the market, $\hat{s}$, just as in the Walrasian case (see Eq. (10.16)), although of course this REE differs from the Walrasian's. The REE collapses to Walrasian's only (i) in the absence of uninformed agents attempting to extract information about the price, $n_U = 0$, and (ii) when the informed agents have access to the same signal, $s^i = s^j$ for all i, j (in which case we would set $n = 1$ in (10.22)). Finally, note that, now, arbitrage is ruled out in a market with a large number of agents:

$$\operatorname*{plim}_{n\to\infty} p = p_* \equiv \frac{d}{R}.$$

The mechanism in this model is the following. While trading, the informed agents embed pieces of information into the equilibrium price—they "set" the information content of the price system so to speak. The uninformed agents free-ride on these prices. The price now performs two roles: one, classical, is to determine the agents' budget constraint; the second, less mechanical, is to inform the uninformed agents about the information other investors possess. Thus, when an uninformed agent observes a low price, he will increase his demand as his budget constraints soften. However, a low price reveals to the uninformed agents that the informed might have received bad information about the asset quality, which decreases their demand. The two effects compensate exactly with each other, leading to a price-inelastic demand, $\hat{x}_U^j = \omega_0$, $j = 1, \cdots, n_U$. Given this, the informed agents can only be allocated their initial asset endowment as well, $\hat{x}_I^i = \omega_0$, $i = 1, \cdots, n_I$. The mathematics of this price-inelasticity will be made more clear once we examine the NREE in the following section.

This model leads to a number of puzzling predictions. First, a feature of the model pointed out earlier, known as the *Grossman's paradox* (Grossman, 1976). By Eqs. (10.20) and (10.21), any informed agent i gives up his knowledge about his own signal, s^i, relying as he does on the average signal, $\hat{s}$. But no agent is using his own information while trading, we may now wonder how can the price end up aggregating information in the first place?

A second model's implication links to the equilibrium allocation. Because each investor is allocated his own endowment in equilibrium, informed and uninformed investors make the same expected profits and, hence, have the same welfare. This raises the following issue, known as the *Grossman-Stiglitz paradox* (Grossman and Stiglitz, 1980): why would we be willing to purchase private information (i.e. the signals s^i), if this information could be freely "read" from the equilibrium price? And, precisely because there are no incentives to purchase private information in this case, the price should then not reflect any.

We now explain these paradoxes could be solved by assuming the presence of "noise" in the equilibrium, leading to a partially revealing price: when the uninformed agents make inference on the private information investors have, based on the price, they can only extract part of this information: while part of their information is given for free to uninformed, uniformed investors can still have incentives to acquire information.

10.6 Noisy Rational Expectations Equilibrium

A natural way to cope with the issues arising in a REE is to assume that the risky asset supply is random. This assumption has an economic interpretation: small realizations of the asset

supply might be viewed as liquidity shocks. The mechanism through which this assumption allows to deal with the conceptual difficulties inherent the REE is the following. In equilibrium, the asset price embeds all of the private information *and* the realization of a liquidity shock. However, investors cannot extract both upon observing the price. The price system, therefore, conveys information, albeit in an imperfect way, which is "lucky" in a way—as Grossman and Stiglitz (1980, p. 393) put it—for were it to do it perfectly, an equilibrium would not exist. It is the role of "noisy" information famously pointed out by Black (1986), and discussed in the Introduction.

We consider two models. One, with asymmetric information, and another, with differential information.

In the model with asymmetric information, the uninformed investors trade at disadvantage compared to the informed. An equilibrium in the market for information obtains once the welfare of the two types of agents (informed and uninformed) are equalized controlling for the cost of information. An equilibrium in the asymmetric information market thus involves a fixed point. For a given number of informed agents, the price reveals some information to the uninformed who determine the asset demand on the basis of the probabilistic distribution of the price. But in equilibrium, prices depend on the uninformed demand. In the Grossman and Stiglitz (1980) market reviewed in Section 10.6.1, there is a simple solution to this *information transmission* problem. In equilibrium, prices are informative, to the extent that the incentives to purchase information decrease with the number of the already informed agents—information choices are strategic substitutes.

In the differential information market of Section 10.6.2, there are no agents with superior information. The models that analyze this market are introduced by Hellwig (1980) and Diamond and Verrecchia (1981). They aim to study issues regarding *information aggregation*: how does the market help aggregate information dispersed across investors? This question has very distant origins in economics. A competitive equilibrium leads to Pareto optimal allocations (first welfare theorem); conversely, a given (desired) Pareto optimal allocation can be decentralized through a dedicated re-distribution of wealth (second welfare theorem). Pushed to the extreme, the second welfare theorem may seem to suggest that any desired market outcome could be implemented through a centralized, socialist type planning system, as in the planning literature following the work of Lange (1936, 1942); that is, once the planner fixes a given outcome as an objective, this outcome could be achieved through a dedicated re-distribution of resources that leads agents to the desired objective under *laissez faire*. Hayek (1945) rejects these ideas: we cannot implement this mechanism while missing data regarding information that is local. After all, the second welfare theorem regards Walrasian equilibria, not the equilibria we are examining in this chapter.

Hellwig and Diamond & Verrecchia models formalize the process of information aggregation in the context of financial markets. It is a complex task because, now, compared to the previous REE markets, agents cannot extract all the information that others know from the price, and need to condition on their own signals and forecast both signals and forecasts of others, over a fixed point. An equilibrium in this market exists once all these forecasts are mutually and internally consistent and, the price is only partially revealing of all the information disseminated in the market.

10.6.1 Asymmetric information: information transmission

Grossman and Stiglitz (1980) consider a model with asymmetric information, in which investors can acquire information on the asset payoff after agreeing to pay a fee, C say. Once the investors agree (say at time t_0) to pay the fee (say once payoffs are revealed), they observe a signal on the asset payoff, and then they trade at $t_1 > t_0$ based on their private information. The uninformed investors trade while making inference on this private information based on the equilibrium price that they are seeing. In equilibrium, we shall see, uninformed investors can only infer part of the information possessed by the informed.

In a first step, we study the equilibrium in the asset market for a given number of informed investors. Due to asymmetric information, informed investors have higher welfare than the uninformed (before paying C). In a second step, we determine the number of informed investors by requiring that the welfare of the uninformed and the informed are the same once accounting for the cost C. It turns out that in this model, the higher the number of informed agents, the lower the incentives to acquire information.

10.6.1.1 Asset market equilibrium

Each informed investor agrees to pay the constant cost C, and observes the same signal on the asset payoff, $s^i = s^j \equiv s$ for all i, j in Eq. (10.10). That is, acquiring information only leads to a partial resolution of risk.[1] We assume, and later justify, that the equilibrium price in this market does not reveal more information than is already possessed by the informed agents. Therefore, the asset demand of the informed agents is the same as in the Walrasian market, Eq. (10.11), viz

$$\hat{x}_I \equiv \hat{x}_I^i = \gamma \pi_{\,d|s} \left(E\left(d \middle| \, s \right) - Rp \right), \quad i = 1, \cdots, n_I, \tag{10.23}$$

with agent's i conditional estimate and precision of the asset payoff being the same as in (10.12)-(10.13),

$$E\left(d \middle| \, s \right) = \bar{d} + \frac{\pi_u}{\pi_d + \pi_u} \left(s - \bar{d} \right), \quad \pi_{\,d|s} = \pi_d + \pi_u. \tag{10.24}$$

We know from the previous section that the assumptions formulated so far lead to an equilibrium in which the uninformed investors free-ride on the equilibrium price, as the latter is fully revealing. Obviously prices *do* reveal the information that informed investors impound on them. The issue is to ascertain whether *some* of this information could not be revealed for free to uninformed investors. Grossman and Stiglitz do indeed study an equilibrium with partial information revelation. One of their key assumption is that the asset supply is random, such that, now, markets reveal both fundamental and non-fundamental sources of information, which an uninformed investor cannot tell apart. Specifically, assume that the total asset supply is random, and equal to[2]

$$nZ \equiv n\omega_0 + \omega, \quad \omega \sim N\left(0, \sigma_z^2 \right). \tag{10.25}$$

We interepret ω as a "liquidity shock." For example, a positive and large realization of ω could be interpreted as an asset sell-off due to non-fundamental reasons (say, asset owners who monetize their investments due to previously unexpected consumption contingencies).

[1] In their original formulation of the model, Grossman and Stiglitz assume that the asset payoff is $d = \theta + \epsilon$, where both θ and ϵ are normally distributed, and that the informed investors observe θ for a fee. The formulation of the model in this section is equivalent. Moreover, we are assuming that the information fee is only paid once payoffs are revealed (rather than at t_0) to simplify the presentation.

[2] To study the model predictions as n increases, we would need to make assumptions on how σ_z^2 would change with n; see Footnote 3 for one alternative assumption. In this section, we take n as given.

The partial revelation mechanism operates as follows. First, note that the uninformed investors now formulate their asset demand conditional upon having learnt about the equilibrium price,

$$\hat{x}_U \equiv \hat{x}_U^j = \gamma \pi_{d|p} \left(E\left(d \middle| p \right) - Rp \right), \quad j = 1, \cdots, n_U. \tag{10.26}$$

Replacing Eqs. (10.23)-(10.26) into the market clearing condition,

$$Z = \lambda \hat{x}_I + (1 - \lambda) \hat{x}_U, \quad \lambda \equiv \frac{n_I}{n}, \tag{10.27}$$

reveals that the equilibrium price, p, is informationally equivalent to the "compound signal" θ, defined as

$$\theta \equiv \left(s - \bar{d} \right) - c \left(Z - \omega_0 \right), \quad c \equiv \left(\lambda \gamma \pi_u \right)^{-1}. \tag{10.28}$$

The compound signal θ (or, equivalently, the price) does reveal information about fundamental information, $s - \bar{d}$; however, it does so imperfectly, conveying as it does additional information regarding possible liquidity shocks, $Z - \omega_0$. An uninformed investor who observes the price, now, cannot tell how much of a price increase is due to good news (say high s) or a negative liquidity shock (say low Z). Therefore, we expect the equilibrium in this market to be one in which informed agents have higher welfare (before paying C) due to their superior information: they know s and of course, they know the price p, and then Z; that is, they know everything (but d).

We search for a linear equilibrium, that is, an equilibrium in which the asset price is a linear function of the compound signal, θ in Eq. (10.28),

$$p = a + b\theta, \tag{10.29}$$

for two constants a and b to be determined. We determine a and b by replacing the asset demands of the informed and uninformed investors into the market clearing condition, Eq. (10.27). We proceed as follows.

First, we determine the conditional expectation, $E\left(d \middle| p \right)$, and the conditional precision, $\pi_{d|p}$, in Eq. (10.26). Note that the equilibrium price p in Eq. (10.29) is affine in θ, and θ is normally distributed, such that $p \sim N\left(a, b^2 \sigma_\theta^2 \right)$. Therefore,

$$E\left(d \middle| p \right) = \bar{d} + b_p \left(p - a \right), \quad \pi_{d|p} = \pi_d + \ell \pi_u, \quad b_p \equiv \frac{cov\left(d, p \right)}{\sigma_p^2}, \tag{10.30}$$

where,

$$b_p = \frac{\pi_\theta}{b \pi_d}, \quad \ell = \frac{\pi_z}{\pi_z + c^2 \pi_u}, \quad \frac{1}{\pi_\theta} = \sigma_d^2 + \sigma_u^2 + c^2 \sigma_z^2, \tag{10.31}$$

such that, the uninformed asset demand in (10.26) is

$$\hat{x}_U = \gamma \pi_{d|p} (\bar{d} + \underbrace{b_p \left(p - a \right)}_{\text{info revelation}} - \underbrace{Rp}_{\text{budget constr.}}). \tag{10.32}$$

Regarding the informed asset demand, we have, by Eqs. (10.23)-(10.24), that:

$$\hat{x}_I = \gamma \pi_{d|s} \left(\bar{d} - Rp \right) + \gamma \pi_u \left(s - \bar{d} \right). \tag{10.33}$$

We shall develop the economic interpretation of both the uninformed and the informed asset demand schedules below.

Second, we plug (10.32) and (10.33) into Eq. (10.27), obtaining the equilibrium price conjectured in (10.29)

$$p = \underbrace{\frac{\bar{d}}{R} - \frac{\omega_0}{\gamma R\left(\lambda \pi_{d|s} + (1-\lambda)\,\pi_{d|p}\right)}}_{=a} + \underbrace{\frac{\lambda \pi_u + (1-\lambda)\,\pi_{d|p}\frac{\pi_\theta}{\pi_d}}{R\left(\lambda \pi_{d|s} + (1-\lambda)\,\pi_{d|p}\right)}}_{=b} \cdot \theta.$$

The model leads to a number of important predictions. First, in equilibrium, the price does not reveal all private information to the uninformed agents, but only a noisy version of it, θ: we say the price is *partially revealing*. The model thus provides a resolution to the Grossman-Stiglitz paradox: as the next subsection will show, there exist equilibria in which it is valuable to purchase private information.

Second, consider the uninformed demand, $\hat{x}_U$ in Eq. (10.32). One the one hand, it decreases with the price p as a mechanical implication of the agent's budget constraint. On the other hand, the uninformed demand increases with p, due to information. In the REE, the two effects exactly compensate with each other, as explained in the previous section. Straightforward calculations do, instead, reveal, that in the equilibrium of this model,

$$b_p - R \leq 0,$$

with an equality only holding once $\sigma_z^2 = 0$. That is, the uninformed asset demand in (10.32) "slopes down." The interpretation is simple: in the NREE, the effects of information revelation are weaker than in the REE case, as the price is partially revealing, as explained.

Third, the constant c determines the price impact of a *liquidity shock*. The higher c, the higher the impact of a liquidity shock, $Z-\omega_0$, compared to information, $s-\bar{d}$. Its inverse, $c^{-1} = \lambda\gamma\pi_u$, is a measure of the risk-bearing capacity of the informed agents vis-à-vis information. Note, indeed, that the informed agent's demand $\hat{x}_I$ in (10.33) has two components. The first relates to how high the price is, compared to the unconditional guess of the asset value; it increases (in absolute value) with the conditional precision of the value estimate given private information, $\pi_{d|s}$. The second component relates to how good news are, compared to unconditional guesses; it increases (in absolute value) with the conditional precision of the noise (i.e. with the informational advantage), reflecting a better quality of private information. Thus, and as anticipated, the term c^{-1} measures the total effect of the informed agent demand with respect to information. The higher this term, the easier becomes for the price system to absorb a liquidity shock, $Z-\omega_0$.

Finally, consider the conditional precision of the asset payoff upon observation of the price, i.e., $\pi_{d|p}$ in Eq. (10.30): it is a measure of *price informativeness*. In particular, the constant ℓ in (10.31) reflects asymmetric information: it increases with π_u, i.e., it decreases with the size of the noisy component of private information, σ_u^2. Moreover, it decreases with c, i.e. it increases with the risk-bearing capacity of the informed agents. In other words, price informativeness decreases as private information becomes more noisy and increases with the risk-bearing capacity of the informed agents. Note, finally, that because $\ell < 1$,

$$\pi_{d|p} = \pi_d + \ell\pi_u < \pi_d + \pi_u = \pi_{d|s}. \tag{10.34}$$

That is, prices provide less information than private signals. This property, while very intuitive, has important welfare implications, discussed below.

10.6.1.2 The value of information

The uninformed agents do not entirely learn about the private information other investors have acquired. Whilst observing the equilibrium price, they confuse fundamentals, $s - \bar{d}$, with liquidity shocks, $Z - \omega_0$ and sometimes, then, trade against the informed. Uninformed investors should therefore expect lower profits and welfare than the informed (before information costs). In Appendix 2, we show that,

$$\frac{E\left(-e^{-\tau(W_I - C)}\right)}{E\left(-e^{-\tau W_U}\right)} = e^{\tau C}\sqrt{\frac{\pi_{d|p}}{\pi_{d|s}}}, \tag{10.35}$$

where $W_I - C$ and W_U denote the terminal wealth of the informed and the uninformed agent. Define the ex-ante profit certainty equivalent of the informed and the uninformed agents as the two values $\mathcal{C}_I$ and $\mathcal{C}_U$ that solve:

$$-e^{-\tau \mathcal{C}_I} = E\left(-e^{-\tau(W_I - C)}\right) \quad \text{and} \quad -e^{-\tau \mathcal{C}_U} = E\left(-e^{-\tau W_U}\right).$$

We define the *value of information* as the net gain of becoming informed in terms of the previous certainty equivalents, which by Eq. (10.35) is,

$$\mathcal{C}_I - \mathcal{C}_U = \frac{1}{2\tau}\ln\left(\frac{\pi_{d|s}}{\pi_{d|p}}\right) - C. \tag{10.36}$$

Because prices provide less information than private signals, $\pi_{d|s} > \pi_{d|p}$ (see Eq. (10.34)), the informed agent is always better off compared to the uninformed, before accounting for the information cost C. Moreover, note that the conditional precision of the asset value estimate made by the informed agents, $\pi_{d|s}$, is independent of the number of informed agents, λ. Instead, that of the uninformed, $\pi_{d|p}$, is increasing in λ, as discussed in the previous section. Therefore, the value of information is strictly decreasing in λ: the higher the number of informed agents, the less valuable it is to acquire private information. An interior equilibrium in the market for information is given by the proportion of informed agents λ^* such that $\mathcal{C}_I - \mathcal{C}_U = 0$. Figure 10.1 depicts the value of information as a function of the proportion of informed agents for given constellations of parameter values.

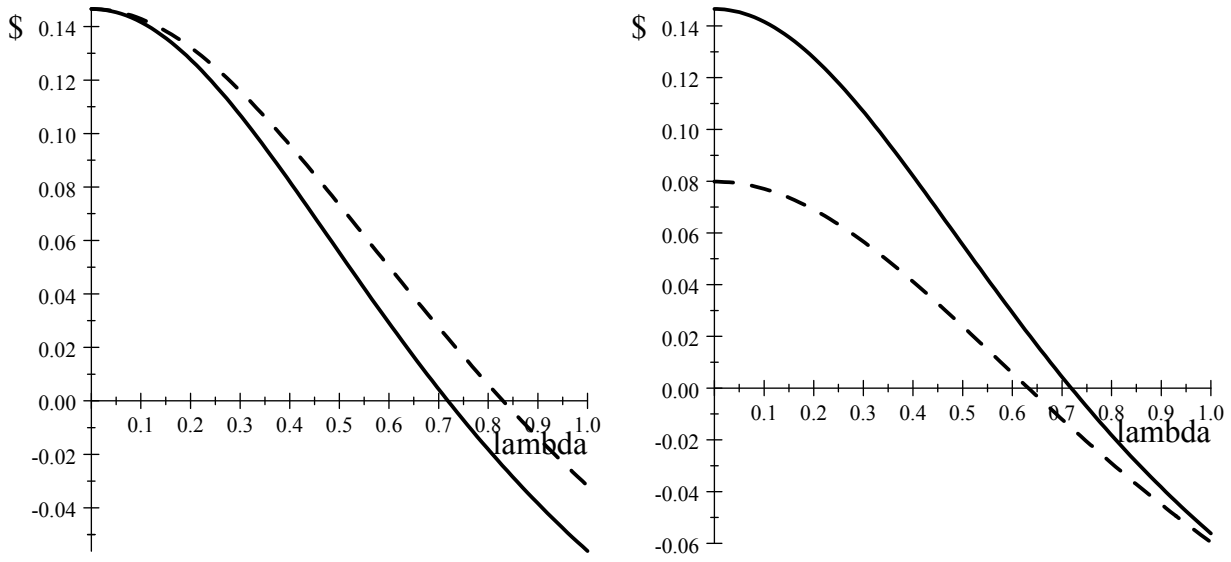


FIGURE 10.1. This figure depicts the value of information, $\mathcal{C}_I - \mathcal{C}_U$ in (10.36), when the cost of information is $C = 0.20$, and parameters are set equal to $\tau = \pi_d = \pi_u = \pi_z = 1$, which is the benchmark solid line in both panels. The left and right panels compare the benchmark with the value of information arising when the size of liquidity shocks increases so as to make $\pi_z = \frac{3}{4}$ (left panel, dashed line), and when the precision of information acquired by the informed agent decreases such that $\pi_u = \frac{3}{4}$ (right panel, dashed line).

In all cases depicted in Figure 10.1, there exist interior equilibria. For example, when the cost of information is $C = 0.20$ and all remaining parameters are set equal to one (the benchmark), the equilibrium obtains with $\lambda^* = 71\%$. This equilibrium is stable in the following sense. When the proportion of informed agents is less than this equilibrium, $\lambda < \lambda^*$, there are incentives to become informed ($\mathcal{C}_I - \mathcal{C}_U > 0$), such that more agents will become informed until λ^* is reached. The opposite occurs when $\lambda > \lambda^*$. There might be non-interior equilibria. For example, the cost of information could be so prohibitive, to make $\mathcal{C}_I - \mathcal{C}_U < 0$ in (10.36) for all λ and in this case the equilibrium would be $\lambda^* = 0$. In general, non-interior equilibria are defined as $\lambda^* = 0$ if $\mathcal{C}_I - \mathcal{C}_U < 0$ for all λ, and $\lambda^* = 1$ if $\mathcal{C}_I - \mathcal{C}_U > 0$.

The comparative statics in Figure 10.1 are easy to explain. In the left panel, we compare a market in which liquidity shocks have higher variability than in the benchmark. The value of information is higher than in the benchmark as prices reveal less information than in the benchmark, and the equilibrium proportion of informed agents increase. In the right panel, the benchmark is compared with a market in which the quality of private information deteriorates, as captured by a lower value of π_u. The value of information in this market now deteriorates and the equilibrium proportion of informed agents decreases as a result.

10.6.1.3 Information sales

Review the literature on information sales.

[In progress]

10.6.2 Differential information: information aggregation

Hellwig (1980) and Diamond and Verrecchia (1981) consider models in which, unlike in the previous section, there are no investors with superior information. The key feature of these models is that in equilibrium, investors learn about the asset fundamentals from both the price *and* the signals they observe. The reasons this is possible are similar to those explained in the asymmetric information market of the previous section: due to the asset being in random supply, the equilibrium price is partially revealing, such that the signals available to each investor are useful while formulating the inference process leading to asset demands.[3]

In detail, we assume that each investor observes a signal on the asset fundamental (thus, $n_U \equiv 0$ and $n_I \equiv n$), which is the same as s^i in Eq. (10.10), and that the total asset supply, nZ, is as in Eq. (10.25). The asset demand of each agent is

$$\hat{x}^i = \gamma \pi_{\,d|s,p} \left(E\left(d \middle| \, s^i, p \right) - Rp \right), \quad i = 1, \cdots, n, \tag{10.37}$$

[3]The two models are similar except that Diamond and Verrecchia consider endowment shocks as the channel leading to partially revealing prices. The inference process of the agents in Diamond and Verrecchia is also more elaborated than in Hellwig, as each agent information set includes the price, the signal on the dividend, and the endowment shock. In Hellwig, there are no endowment shocks as explained.

where, as usual, p denotes the equilibrium price, and where the conditional precision of the payoff estimate is assumed to be the same for each agent, $\pi_{d|s,p} \equiv \pi_{d|s^i,p}$. This conjecture on the conditional precision will be confirmed below.

We shall formulate conjectures regarding p below. For now, we assume that the price p is a random variable that helps each investor forecast the asset payoff, d. The rationale behind this conjecture is that in equilibrium,

$$nZ = \sum_{i=1}^{n} \hat{x}^i, \tag{10.38}$$

such that the clearing price aggregates the signals all investors observe, which are potentially useful in each investor's inference problem.

Furthermore, we conjecture that the conditional expectation,

$$E\left(d|\, s^i, p\right) = \bar{d} + \chi_s \left(s^i - \bar{d}\right) + \chi_p \left(p - a\right), \tag{10.39}$$

for two constants χ_s and χ_p. If p were exogenous, these two constants could be readily determined relying on the projection theorem. However, p is endogenous, and these coefficients could well depend on how the equilibrium price aggregates all available information. In other words, due to the endogeneity of the price, p, we need to check whether the projection theorem representation in (10.39) holds and whether it is unique.

There is a natural conjecture to make regarding the equilibrium price, which we shall justify in a moment, namely that p is an affine function of the average signal and the per-capita supply,

$$p - a = b_s \left(\hat{s} - \bar{d}\right) - b_z \left(Z - \omega_0\right), \tag{10.40}$$

where $\hat{s}$ denotes the average signal in Eq. (10.17). The rationale underlying this conjecture is the following. Note that by replacing Eq. (10.39) into the asset demand $\hat{x}^i$ in (10.37), and then plugging the result into the equilibrium condition (10.38), yields, after rearranging terms, that,

$$p - a = \underbrace{\frac{-Ra + \bar{d} - \frac{1}{\gamma \pi_{d|s,p}}\omega_0}{R - \chi_p}}_{\equiv 0} + \underbrace{\frac{\chi_s}{R - \chi_p}}_{\equiv b_s}\left(\hat{s} - \bar{d}\right) - \underbrace{\frac{1}{\gamma \pi_{d|s,p}\left(R - \chi_p\right)}}_{\equiv b_z}\left(Z - \omega_0\right), \tag{10.41}$$

where the identities under the brackets follow by the conjecture made in (10.40).

The first restriction in (10.41) delivers the value of a in (10.40),

$$a = \frac{\bar{d}}{R} - \frac{1}{\gamma R \pi_{d|s,p}}\omega_0. \tag{10.42}$$

Instead, the coefficients b_s and b_z depend on the coefficient updates in the conditional expectation of Eq. (10.39), χ_s and χ_p,

$$b_s = \frac{\chi_s}{R - \chi_p}, \quad b_z = \frac{1}{\gamma \pi_{d|s,p}\left(R - \chi_p\right)}, \tag{10.43}$$

where the very same χ_X, $X \in \{s, p\}$ depend on b_Y, $Y \in \{s, z\}$, over a a fixed point.

Intuitively, the statistical distribution of the equilibrium price depends on the agents' beliefs because the latter determines the asset demand; however, the agents' beliefs depend on the equilibrium distribution that generates the conditioning information the agents rely on while

formulating their very same beliefs—the price. It is a fixed point leading to endogenous beleifs, i.e., (noisy) rational expectations. More in detail, the price coefficients, $b_\cdot$, depend on the agents' coefficient updates, $\chi_\cdot$, as the agents' conditional expectation of the dividend (and, hence, $\chi_\cdot$) determine their asset demand, thereby contributing to the equilibrium price (through Eq. (10.38)). In turn, the agents' coefficient updates $\chi_\cdot$ depend on the statistical distribution of the equilibrium price (and, hence, $b_\cdot$), as the latter is a conditioning variable for the agents' expectations of the dividends. Appendix 2 shows that there is a unique solution to this fixed point, and that the price in Eq. (10.40) can be expressed as

$$p - a = \frac{\pi_u \left(\left(1 - \frac{1}{n}\right) \pi_z + \psi_*^2 \pi_u\right)}{R \left(\left(1 - \frac{1}{n}\right) \left(\pi_u + \frac{1}{n}\pi_d\right) \pi_z + \psi_*^2 \left(\pi_u + \pi_d\right) \pi_u\right)} \cdot \theta, \tag{10.44}$$

where θ is the compound signal, defined as

$$\theta = \left(\hat{s} - \bar{d}\right) - \psi_* \left(Z - \omega_0\right), \tag{10.45}$$

and ψ_* is a strictly positive and bounded constant defined in the Appendix (see Eq. (10A.18)).

The fact that $\psi_* > 0$ implies that the price only partially reveals the average signal, $\hat{s}$. Similarly as for the constant c in the Grossman and Stiglitz (1980) model (see Eq. (10.28)), the constant ψ_* in this model determines the price impact of a liquidity shock. Below, we shall see a striking similarity between these two constants in the limiting case of a large number of agents.

The compound signal in Eq. (10.45) is the differential information counterpart to θ in (10.28). That is, in the differential information market of this section, the price is partially revealing, as θ conveys noisy information about the asset fundamentals. The equilibrium in this market then addresses a number of issues that the REE cannot. First, Eqs. (10.44) and (10.45) reveal that b_s and b_z are both positive. Moreover, χ_s and χ_p are both positive, provided that the conditional precision $\pi_{d|s,p} > 0$,[4] a condition that can be shown to hold. That is, every agent now conditions on both his own signal and the price while formulating his demand function—a resolution to the Grossman (1976) paradox. The coefficient χ_s is positive because the agents' signals are obviously positively correlated with the dividend; and the coefficient χ_p is also positive, because the equilibrium price is positively correlated with the asset dividend.

The solution in Eq. (10.44) is not known in closed form because the constant ψ_* is not. Note that this constant determines the price sensitivity to the average signal and liquidity shocks, b_s and b_z and also the conditional precision, $\pi_{d|s,p}$ in (10.42). However, a solution is available in the limiting case in which the number of agents is large. This case provides intuition regarding the finite case.

So suppose that the number of agents is large, such by the Law of Large Numbers, the price in (10.44) satisfies[5]

$$\operatorname*{plim}_{n\to\infty} p = \bar{p} \equiv \underbrace{\frac{\bar{d}}{R} - \frac{1}{R\gamma\bar{\pi}_{d|s,p}}\omega_0}_{\equiv \bar{a}_s} + \underbrace{\frac{\pi_u \left(1 + \gamma^2 \pi_u \pi_z\right)}{R \left(\pi_d + \left(1 + \gamma^2 \pi_u \pi_z\right) \pi_u\right)}}_{\equiv \bar{b}_s} \cdot \upsilon, \tag{10.46}$$

[4] Assume momentarily that $\chi_p > 0$. Because $b_z > 0$, the second of Eqs. (10.43) implies that $\chi_p < R$. Hence, by the first of Eqs. (10.43), and $b_s > 0$, $\chi_s > 0$. The fact that $\chi_p > 0$ follows by Eq. (10A.12) in Appendix 2.

[5] By Eq. (10.25), we have that $Z - \omega_0 \sim N\left(0, \frac{1}{n^2}\sigma_z^2\right)$. We assume that the limit $\lim_{n\to\infty} \frac{1}{n^2}\sigma_z^2$ is bounded and bounded away from zero. For example, this case arises when the shock ω in (10.25) is the sum of the endowment shocks affecting each agent, say $\omega \equiv \sum_{i=1}^n w_i$, where $w_i \sim N\left(0, \sigma_z^2\right)$.

where υ is the compound signal,

$$\upsilon \equiv \left(d - \bar{d}\right) - \bar{\psi}\left(Z - \omega_0\right), \quad \bar{\psi} \equiv \left(\gamma \pi_u\right)^{-1},$$

and $\bar{\pi}_{d|s,p}$ denotes the limiting conditional precision,

$$\bar{\pi}_{d|s,p} = \lim_{n\to\infty} \pi_{d|s,p} = (\pi_d + \pi_u) + \gamma^2 \pi_z \pi_u^2. \tag{10.47}$$

The price now reveals better information about the asset fundamentals than in the finite agents case, as the signal υ embeds more precise information about the asset payoff than θ in (10.45). Naturally, each agent still finds it useful to condition his demand on both his own signal and the equilibrium price. Moreover, the price is partially revealing in this limiting case too. Note, now, that the limiting coefficient, $\bar{\psi}$, is the same as c in the Grossman and Stiglitz (1980) model when $\lambda = 1$ (see Eq. (10.28)). In other words, in this market, liquidity shocks are absorbed through a mechanism quite comparable to that in the asymmetric information market.

Finally, the conditional precision in (10.47) is the sum of the precision of the payoff given each agent's signal, $\pi_{d|s} = \pi_d + \pi_u$, plus a second component, $\gamma^2 \pi_z \pi_u^2$. This second component is increasing in γ: the agent's demand becomes more aggressive as risk-tolerance increases, leading to a more informative equilibrium price. The second component is also increasing in both π_z and π_u. The effects regarding an increase in π_z are explained by the price being more responsive to fundamentals than noise as σ_z^2 decreases. The effects regarding an increase in π_u link to each agent providing more information over equilibrium aggregation while trading more aggressively on less noisy signals.

10.7 Dealers markets: Introduction

Asset prices fluctuate even in the absence of important public news about the fundamentals, exhibiting a quite pronounced volatility at as such high frequency scale as intraday. Information problems could be a market driver. Indeed, the previous sections emphasize that asset prices can incorporate private information conveyed by investors, albeit in a partially revealing fashion. Intuitively, changes in private information can be incorporated into asset prices more or less quickly, but once prices are partially revealing, they can fluctuate even while asset demand is driven by reasons unrelated to fundamentals.

This and the following section formalize the previous reasoning and build up on the insights of the previous section to develop information-based explanations of asset price fluctuations. While the models in the previous sections provide the necessary foundations to the role information plays in asset markets, the organization of financial markets and the very trading systems have institutional details that go beyond those underlying the previous models. These details include the timing of the market clearing mechanism (when), the geography (where) and modalities (how).

The *timing* of the clearing mechanism can be best illustrated by two main organization protocols: (i) call auctions (batch markets), whereby trading occurs over prespecified windows of time, and (ii) continuous auctions, whereby trading occurs continuously. The *geography* of markets is varied. There are centralized markets such as those for products listed on an exchange. There are decentralized markets (e.g., OTC markets). There are dealership markets, in which end-users trade through a number of dealers; dealership markets are therefore decentralized

markets: an hypothetical end-user who is not satisfied with the quotes a dealer provides him with, can search for better opportunities offered by another dealer. Inter-dealer markets are markets in which only dealers operate who adjust their desired balances from those carried over while trading with end-users. While dealers typically operate in OTC markets, they can also operate on exchanges like the LSE. Finally, *modalities* of the market clearing mechanism, or order submissions, can broadly take on two forms: (i) market orders or (ii) limit orders. These two forms may well affect market conditions. For example, market orders (i.e. those most likely to be initiated by informed/"active" investors) "consume" limit orders, and decrease market liquidity, while limit orders (most likely to be initiated by passive investors) improve market liquidity.

We analyze the role of information in the price formation process while taking into account some of the previous "market microstructure" details. In this and in the following section, we study markets in which there are traders submitting market orders and dealers who operate as market makers and set prices. Market makers can be competitive or not. This section displays an important example of a market outcome in a setting where market makers are competitive, a famous model developed by Glosten and Milgrom (1985). The model is a cornerstone in the theory of market microstructure, and helped develop empirical measures of liquidity. We hinge upon this model to illustrate how information matters in dealers markets, and how the model could be enriched (in the next section).

10.7.1 Markets with symmetric information

Suppose that market makers are perfectly competitive and risk-neutral, and that they receive market orders from potentially informed traders. The latter are potentially informed, in that the market makers do not know whether the orders they receive emanate from an informed or a liquidity trader—an assumption that parallels those leading to partial revelation in the NREE models of the previous section. As usual, informed traders are those who observe a signal on the asset payoff, d. Denote with $\mathbb{F}_t$ the information set that is available to market makers at time t, and assume buy or sell orders arrive sequentially to them. Clearly, $\mathbb{F}_t = \mathbb{F}_{t-1} \bigcup o_t$, where o_t can be either a buy or a sell order, i.e. $o_t \in \{\text{buy}_t, \text{sell}_t\}$. For simplicity, we assume interest rates are zero.

Suppose, now, that $\mathbb{F}_t$ is the same as that available to all traders: that is, there are no "informed" traders. Then, if dealer i receives a buy order, and is risk-neutral, he will set the ask price at t, a_{it} say, such that,

$$0 = E\left(a_{it} - d \middle| \mathbb{F}_{t-1}, o_t = \text{buy}_t\right) = E\left(a_{it} - d \middle| \mathbb{F}_{t-1}\right), \tag{10.48}$$

where the second equality follows because the buy order is totally uninformative, that is, it has no predictive power on d.

Likewise, if dealer i receives a sell order, he will set the bid price at t, b_{it} say, such that,

$$0 = E\left(d - b_{it} \middle| \mathbb{F}_{t-1}, o_t = \text{sell}_t\right) = E\left(d - b_{it} \middle| \mathbb{F}_{t-1}\right). \tag{10.49}$$

The expected profits in (10.48) and (10.49) are set equal to zero, due to the assumption dealers are perfectly competitive, and the second equalities in both (10.48) and (10.49) follow by the assumption that trades arrive to dealers while not carrying any information about d. Therefore, the bid-ask spread is zero for each dealer, and the asset price p_t collapses to

$$p_t = a_{it} = b_{it} = E\left(d \middle| \mathbb{F}_{t-1}\right).$$

In absence of public news about d, the price p_t will not change. It is counterfactual, as explained. Glosten and Milgrom (1985) show how the bid-ask spread can be positive in a model in which dealers face informed investors. We present a simplified version of their model next.

10.7.2 With asymmetric information: bid-ask spreads

Suppose the asset payoff d has a binary distribution: it is high, d_h, with probability α, or low, d_l, with complementary probability. Denote with α_t the probability that $d = d_h$ conditionally upon the entire history of orders seen by the market makers up to t,

$$\alpha_t = \Pr\left(d = d_h | \, \mathbb{F}_t\right).$$

Trades arrive sequentially to the market makers. The latter believe there is (i) a probability π the order at t is information driven; (ii) a probability $1-\pi$ the order at t is liquidity driven, and in this case, 50% chances the liquidity motivated order is a buy or a sell. Note that the order, now, might contain private information regarding the asset payoff. While market makers are not sure about whether the order is informative, they post order-contingent rules that reflect updated information: a bid price if the order they receive is a sell, and a ask price if the order is a buy.

For example, and regarding a sell order, competition leads a representative market maker to post the following bid price

$$b_t = E\left(d | \, \mathbb{F}_{t-1}, o_t = \text{sell}_t\right) = \alpha_{t-1}\left(\text{sell}_t\right) d_h + \left(1 - \alpha_{t-1}\left(\text{sell}_t\right)\right) d_l, \tag{10.50}$$

where by Bayes' law,

$$\alpha_{t-1}\left(\text{sell}_t\right) \equiv \Pr\left(d = d_h | \, \mathbb{F}_{t-1}, o_t = \text{sell}_t\right) = \frac{\Pr\left(o_t = \text{sell}_t | \, d = d_h, \mathbb{F}_{t-1}\right)}{\Pr\left(o_t = \text{sell}_t | \, \mathbb{F}_{t-1}\right)} \alpha_{t-1}. \tag{10.51}$$

Once again, note that the market maker posts bid and ask prices before knowing whether he will receive a sell or buy order. As explained, his is a posting rule and actually a *regret-free* rule: once the order is filled, market makers will not regret having executed at the prices they had previously posted given the information of the time. This rule is, of course, entirely consistent with market practice.

To determine the probabilities of the ratio in (10.51), note that conditionally on $d = d_h$, informed traders would never sell, such that the probability to receive a sell order in this case is simply the probability that the order is liquidity motivated, $1-\pi$, times the probability that a liquidity trader sells, $\frac{1}{2}$,

$$\Pr\left(o_t = \text{sell}_t | \, d = d_h, \mathbb{F}_{t-1}\right) = \frac{1}{2}\left(1 - \pi\right). \tag{10.52}$$

Furthermore,

$$\Pr\left(o_t = \text{sell}_t | \, \mathbb{F}_{t-1}\right) = \alpha_{t-1} \cdot \frac{1}{2}\left(1 - \pi\right) + \left(1 - \alpha_{t-1}\right) \cdot \left(\pi + \frac{1}{2}\left(1 - \pi\right)\right). \tag{10.53}$$

The first term on the RHS of the previous equality arises because conditionally upon $d = d_h$ (which has α_{t-1} likelihood given $\mathbb{F}_{t-1}$), the only sell trades can be liquidity motivated, which occurs with probability $\frac{1}{2}\left(1 - \pi\right)$, as explained while deriving Eq. (10.52). The second term on

the RHS of (10.53) is due to the fact that conditionally on $d = d_l$ (which has $1-\alpha_{t-1}$ likelihood given $\mathbb{F}_{t-1}$), the probability of receiving a sell order is the sum of (i) the prob of an informed trade, π, plus (ii) the prob of a liquidity motivated order, $\frac{1}{2}(1-\pi)$.

Replacing Eqs. (10.52) and (10.53) into Eq. (10.51) leaves

$$\alpha_{t-1}(\text{sell}_t) = \frac{\frac{1}{2}(1-\pi)}{\frac{1}{2}(1-\pi)+(1-\alpha_{t-1})\pi}\alpha_{t-1}. \tag{10.54}$$

An analogous reasoning leads to the following expression for the ask price posted by a representatitive market maker

$$a_t = E(d|\,\mathbb{F}_{t-1}, o_t = \text{buy}_t) = \alpha_{t-1}(\text{buy}_t)\, d_h + (1-\alpha_{t-1}(\text{buy}_t))\, d_l, \tag{10.55}$$

where

$$\begin{aligned}\alpha_{t-1}(\text{buy}_t) &\equiv \Pr(d = d_h|\,\mathbb{F}_{t-1}, o_t = \text{buy}_t)\\ &= \frac{\Pr(o_t = \text{buy}_t|\,d = d_h, \mathbb{F}_{t-1})}{\Pr(o_t = \text{buy}_t|\,\mathbb{F}_{t-1})}\alpha_{t-1}\\ &= \frac{\frac{1}{2}(1+\pi)}{\frac{1}{2}(1-\pi)+\alpha_{t-1}\pi}\alpha_{t-1}.\end{aligned} \tag{10.56}$$

It is immediate to check that

$$\alpha_{t-1}(\text{buy}_t) \geq \alpha_{t-1}(\text{sell}_t).$$

That is, the order flow is informative.[6]

To summarize, Eqs. (10.50) and (10.55) form a stochastic difference system, where the conditional probability the asset is good, α_t, is

$$\alpha_t = \begin{cases} \alpha_{t-1}(\text{sell}_t), & \text{as in Eq. (10.54)}, \quad \text{with prob. } \bar{\pi}(d)\\ \alpha_{t-1}(\text{buy}_t), & \text{as in Eq. (10.56)}, \quad \text{with prob. } 1-\bar{\pi}(d)\end{cases} \tag{10.57}$$

where $\bar{\pi}(d)$ denotes the probability of a sell order when the true asset value is $d \in \{d_l, d_h\}$, notably[7]

$$\bar{\pi}(d_h) = \frac{1}{2}(1-\pi) \quad \text{and} \quad \bar{\pi}(d_l) = \pi + \frac{1}{2}(1-\pi).$$

The model implies that *adverse selection costs* are borne by liquidity motivated orders. To motivate the emphasis on adverse selection, note that we are describing a market where the agents who hold private information would selectively trade in order to benefit from their knowledge: they buy when the asset is good (and the price is low) and sell when the asset is bad (and the price is high), such that the market would actually collapse without noise, as in the lemon problem.[8]

[6]Suppose the order size is $\tilde{x} = \mp 1$, such that $E_t(\tilde{x}|\,d = d_h) = 1\cdot\pi + \left(1\frac{1}{2} - 1\frac{1}{2}\right)\cdot(1-\pi) = \pi$ and, similarly, $E_t(\tilde{x}|\,d = d_l) = -\pi$, where E_t is the expectation taken given information up to time t. Therefore, by the Law of Iterated Expectations and simple calculations, $E_t(d\tilde{x}) = E_t(E_t(d\tilde{x}|\,d)) = \alpha_t d_h E_t(\tilde{x}|\,d = d_h) + (1-\alpha_t)\, d_l E_t(\tilde{x}|\,d = d_l) = (\alpha_t d_h - (1-\alpha_t)\, d_l)\,\pi$, which is non-zero in general.

[7]The model can be simulated as follows. We take as given the true value of the asset payoff, say $d \in \{d_l, d_h\}$, and simulate the conditional probabilities in Eqs. (10.57), thereby generating bid and ask prices according to Eqs. (10.50) and (10.55).

[8]Part I provides a description of the celebrated lemon problem (see Chapter 4).

In particular, consider the fair value of the asset, i.e. that arising in the absence of any information frictions (i.e. prior to possibly informed trades arriving at t), defined as

$$p_{t-1} = E\left(d|\,\mathbb{F}_{t-1}\right) = \alpha_{t-1}d_h + \left(1-\alpha_{t-1}\right)d_l. \tag{10.58}$$

It is the price we would observe absent any adverse selection. We have, accordingly, and using Eq. (10.50), Eq. (10.54) and Eq. (10.58),

$$b_t - p_{t-1} = s_t^b \cdot \left(d_h - d_l\right), \tag{10.59}$$

where

$$s_t^b \equiv \alpha_{t-1}\left(\text{sell}_t\right) - \alpha_{t-1} = -\frac{\left(1-\alpha_{t-1}\right)\pi}{\frac{1}{2}\left(1-\pi\right)+\left(1-\alpha_{t-1}\right)\pi}\alpha_{t-1} \leq 0.$$

The bid price is less than the fair value because once a market maker receives a sell order, he will be unsure about whether the sell order is liquidity motivated: it could be information motivated. This correlation between the order flow and the asset value leads the market maker to evaluate the asset less than based on his information set prior to the received order. The spread, s_t^b (posted at $t-1$ in anticipation of a sell order possibly occurring at t) represents the updated beliefs of the market maker after receiving a sell order. Adverse selection costs are borne by liquidity traders because those affected by a liquidity shock, will sell the asset at a price b_t lower than the fundamental value according to the previous information set not including their own order, p_{t-1}.

Likewise, and using Eq. (10.55), Eq. (10.56) and Eq. (10.58),

$$a_t - p_{t-1} = s_t^a \cdot \left(d_h - d_l\right), \tag{10.60}$$

where

$$s_t^a \equiv \alpha_{t-1}\left(\text{buy}_t\right) - \alpha_{t-1} = \frac{\left(1-\alpha_{t-1}\right)\pi}{\frac{1}{2}\left(1-\pi\right)+\alpha_{t-1}\pi}\alpha_{t-1} \geq 0.$$

Now, those who want to buy an asset for liquidity reasons and, accordingly, place a buy order, will buy at a price a_t higher than p_t, bearing an adverse selection cost.

Note, also, that the fact that $s_t^b \leq 0$ and $s_t^a \geq 0$ imply that transaction prices (bid, when a sell order is placed; and ask, when a buy order is place) move in the direction of the order flow. Moreover, the two spreads, s_t^a and s_t^b, cannot be the same. In particular, we have

$$-s_t^b > s_t^a \Longleftrightarrow \alpha_{t-1} > \frac{1}{2}.$$

That is, if the conditional probability the asset payoff is high, α_t, is growing (say for example because there are more and more buy orders), adverse selection on the ask side will be less and less severe, as informed traders are "giving them away." Indeed, it is possible to show that the model predicts the price is asymptotically *strongly* efficient, in that $\alpha_t \to 0$ if $d = d_l$ and $\alpha_t \to 1$ if $d = d_h$. Intuitively, chances become higher and higher the asset payoff is truly low after a large number of sell orders arriving to the market maker's trading desk.

Finally, the bid-ask spread predicted by the model is

$$a_t - b_t = \left(s_t^a - s_t^b\right)\left(d_h - d_l\right).$$

It is increasing in the volatility of the asset payoff, consistent with empirical evidence, being zero only when $\pi = 0$ (no asymmetric information) or asymptotically, when either $\alpha_t \to 0$ or $\alpha_t \to 1$.

The insights of the model are very important although the model has limitations. First, the only source of the bid-ask spread relies on information asymmetries and the inherent adverse selection: the market maker may receive buy orders by investors who know the asset payoff is good, or sell orders by investors who know the asset is a lemon. That is, the order flow correlates with the asset payoff, which leads to "price impacts" of trades not related to fundamentals—the liquidity shocks; that is, whenever a liquidity trader trades, he will move the price above or below the fundamentals because the market maker anticipates that the order he observes could be information driven.[9] Furthermore, two assumptions of the model are that informed traders do not act strategically, and that the order size is fixed. The next section analyzes markets in which traders understand their trades can have a price impact and can, accordingly, optimize on their order size and also "distribute" their order sizes over time optimally.

10.7.3 Inventory risk and bid-ask spreads

Bid-ask spreads are an important determinant of liquidity but liquidity has many dimensions. Traders obviously would always like to "get in" when needed but, at the same time, "they would also like to get out" of a trade. Narrow bid-ask spreads "do not help" if trades are difficult to execute as books lack depth. One possible reason for scarce depth is that inventories are subject to market fluctuations. Not only, then, does market makers' risk-aversion lead to positive bid-ask spreads that compensate for this "inventory risk"; this risk-aversion may lead market makers to maintain finite inventories. This section develops a model that focusses on the bid-ask spread aspect of this complex problem. It is based on a variant of a formulation originally proposed by Stoll (1978).

Consider an asset market in which a liquidity supplier (a dealer, say) has concerns regarding his inventory risk. The dealer is ready to buy the asset at some ask price and sell the same asset at some bid price, and he does perform a purely intermediating role. It is possible that in performing this role, the dealer ends up with a non-zero net position, say some positive asset holding. Inventory risk is the risk the dealer's asset holding may change in value. Aversion to this risk leads to a bid-ask spread, even in the absence of asymmetric information.

Consider a two-period market in which such dealers are competitive. Let p_t the price of the asset at time-t; and let W_t be the dealer's wealth, ω_t the dealer's asset inventory, x_t the asset holdings and $n_t = \omega_t - x_t$ the asset units that are traded at time-t. The dealer is selling the risky asset if n_t is positive, and is buying if n_t is negative. Assume that investing in cash produces a zero net return, i.e., set $R = 1$. We know, since the beginning of Section 10.4, that the wealth constraint of any trader should satisfy $W = (W_0 - (x - \omega_0)\, p)\, R + xd$, where W_0 denotes the trader's initial wealth and ω_0 denote his initial endowment of the asset. We can rely on this budget constraint to conclude that the dealer's wealth at time-$t+1$ is

$$W_{t+1} = n_t p_t + (\omega_t - n_t)\, p_{t+1} + c_t,$$

where c_t denotes cash available at t.[10] Note that asset inventories at time-$t+1$ is simply the inventories at t minus the asset units sold at t, viz

$$\omega_{t+1} = \omega_t - n_t. \tag{10.61}$$

We assume that at time-$t+1$, the asset pays off d, which is normally distributed with mean $\bar{d}$ and variance σ_d^2, as usual. Obviously, then, in this two-period market, $p_{t+1} = d$. In this model,

[9] Section 10.9 provides additional explanations of bid-ask spreads, based on search frictions.

[10] Just set $c_t \equiv W_0$, $x_t \equiv x$, $\omega_t \equiv \omega_0$, $p_{t+1} \equiv d$.

then, we can interpret the difference $d - \bar{d}$ as information regarding the asset value that leads to market fluctuations. It is this risk that the dealer is averse to. Precisely, we assume that the dealer maximizes the expected utility of his final wealth with respect to n_t, and assume that this utility is CARA with coefficient $\tau > 0$. The first order conditions for the dealer lead to the following dealer's supply function

$$n_t = \omega_t + \gamma \pi_d \left(p_t - \bar{d} \right), \tag{10.62}$$

where we have denoted $\gamma \equiv \tau^{-1}$ (constant risk-tolerance) and $\pi_d \equiv \sigma_d^{-2}$ (the dividend precision), as usual.

The dealer is willing to sell more than allowed by his current inventory should the asset price p_t be above its unconditional expectation $\bar{d}$. That is, if $p_t > \bar{d}$, then $n_t > \omega_t$, such that $\omega_{t+1} < 0$ (see Eq. 10.61)): the dealer enters into a short position. Instead, should the price p_t be below $\bar{d}$, the dealer would then be willing to supply less than the inventory, such that $\omega_{t+1} > 0$, resulting in a long position. The dealer's willingness to serve the market in this way increase with his risk-bearing capacity $\gamma \pi_d$.

Finally, consider the equilibrium in this market. At time-t, the dealer receives an order, a random variable $o_t = +1$ (in case of a buy order) or $o_t = -1$ (in case of a sell order). In equilibrium, $n_t = o_t$, such that, by Eq. (10.62)

$$p_t = \begin{cases} m_t + \frac{1}{\gamma \pi_d} & \text{if } o_t = +1 \text{ (ask price)} \\ m_t - \frac{1}{\gamma \pi_d} & \text{if } o_t = -1 \text{ (bid price)} \end{cases}$$

where $m_t \equiv \bar{d} - \frac{\omega_t}{\gamma \pi_d}$ is the mid quote. The bid-ask spread is equal to $\frac{2}{\gamma \pi_d}$, and decreases with the dealer's risk-bearing capacity.

10.8 Markets with strategic players

Investors with superior information do affect asset prices in all the markets analyzed so far. The question arises as to how such investors trade while making the best use of their information. In the markets of the previous sections, agents do not understand they have a price impact. This section removes this assumption and considers dealer markets in which investors with superior information make a strategic use of their information while realizing they have a price impact. The models we study make sharp predictions regarding how liquidity conditions are affected by the extent of information asymmetry between traders and dealers.

How does an informed trader trade while realizing his trade has a price impact? On the one hand, an investor with superior information (an "insider trader," say) should trade aggressively on his information advantage against the market makers. On the other hand, prices likely move against the insider trader while he trades aggressively: they increase when the insider buys and decrease when he sells. So a large trade is not optimal because prices could move adversely, but then, a small trade could just be "under-utilized information."

In the celebrated Kyle (1985) model, there is a trading size that optimizes the previous trade-off given the market makers pricing rule. Regarding this pricing rule, we assume that the market makers fully understand that the "order flow" they observe comprises both informed trading and noise. But similarly as in the Glosten and Milgrom (1985) model, they cannot disentangle the two components. Therefore, liquidity traders bear an adverse selection cost: the order flow has a price impact, such that the liquidity traders will buy (sell) at higher (lower) price than the

fundamental value, controlling for the informed trading. A price impact actually exists because the market makers understand that there might be informed traders.

Kyle (1985) solves for the equilibrium of this game and shows that the price impact, and hence, liquidity conditions, tightly link to the extent of the information asymmetry between traders and dealers. In Sections 10.8.1 and 10.8.2, we analyze the static, baseline version of his model as well as a few variants of it. In Section 10.8.3, we analyze and discuss dynamic extensions of the baseline model, explaining the model's implications in terms of trading patterns and general market behavior.

10.8.1 The Kyle baseline model

Kyle (1985) considers a market in which prices are determined in a *sequential equilibrium*, similarly as in the Glosten and Milgrom (1985) market. An insider trader knows the value of the asset payoff, d, and submits his order, x. The market maker (many of them, actually), observes the aggregate order flow, defined as the sum of the insider trade and a liquidity shock, Z, i.e. $y \equiv x + Z$, where $d \sim N(\bar{d}, \sigma_d^2)$ and $Z \sim N(0, \sigma_z^2)$, and fundamentals (d) and noise (Z) are independent.

Given the order flow they observe, and perfect competition, the representative market maker sets the asset price according to semi-strong informational efficiency,

$$p = E(d|\, y)\,. \tag{10.63}$$

Thus, in Glosten and Milgrom (1985), market makers set prices in advance of the orders they receive (but they are conditional upon a buy or sell), and these orders can be either informed or uninformed. In Kyle (1985), market makers set price upon observing an order flow that comprises both uninformed and informed trades. Only the market maker observes the order flow.

We conjecture the price is linear in the aggregate order flow,

$$p = \bar{d} + \lambda y, \tag{10.64}$$

where the coefficient, λ, known as "Kyle's lambda," measures the price impact of the order flow. In other words, its inverse, $1/\lambda$, is a measure of *market depth*, i.e. the order flow that is needed to induce prices to change by one dollar. So because the order flow obviously contains liquidity shocks, the higher λ, the less liquid the market is. We shall determine the value of λ, below, as part of the equilibrium of the game between the informed trader and the market maker.

The insider trader chooses his order size to maximize his profits while anticipating he has price impact,

$$\max_x E\left((\tilde{d} - p)x \middle|\, \tilde{d} = d\right),$$

with the price p being as in Eq. (10.64). The first order conditions of the previous program lead to the following demand schedule,

$$x = \beta\left(d - \bar{d}\right), \quad \beta \equiv \frac{1}{2\lambda}. \tag{10.65}$$

That is, the insider trade is proportional to his informational advantage, $d - \bar{d}$, which is the difference between his information about the asset payoff, and the unconditional guess of the

asset payoff, $\bar{d}$—the "fair value" of the asset. Naturally, the coefficient of proportionality, β, has still to be determined.

Regarding the market maker's beliefs, we have, by the projection theorem, that,

$$p = E(d|\, y) = \bar{d} + \frac{cov(d,y)}{var(y)} y = \bar{d} + \underbrace{\frac{\beta\sigma_d^2}{\beta^2\sigma_d^2 + \sigma_z^2}}_{\equiv\lambda} \cdot y, \tag{10.66}$$

where the identity under the brackets follows by the conjecture made in (10.64).

To summarize, the equilibrium in this market is one in which the market maker sets the price p as in Eq. (10.64), and the optimal trading size of the insider is given by x in Eq. (10.65). The price impact, λ, in (10.64) and the trading aggressiveness coefficient, β, in (10.65), satisfy:

$$\begin{cases} \beta &= \dfrac{1}{2\lambda} \\ \dfrac{1}{\lambda} &= \beta + \dfrac{\sigma_z^2}{\sigma_d^2}\dfrac{1}{\beta} \end{cases} \tag{10.67}$$

The first condition says that when markets are deep (i.e. when the price impact λ is low), the price is obviously less likely to move against the insider, who will therefore trade aggressively on his informational advantage, $d - \bar{d}$. The second condition says that the market maker makes the markets deep both when the insider trades little on his information advantage (i.e. when β is low) and, when he trades so aggressively to reveal much of his information (i.e. when β is high). An equilibrium of this game is the solution to the system (10.67),

$$\beta = \frac{\sigma_z}{\sigma_d}, \quad \lambda = \frac{1}{2}\frac{\sigma_d}{\sigma_z}. \tag{10.68}$$

The insider trades more and more aggressively as liquidity trades become large (i.e. as σ_z becomes large), because it is easier to hide his information in this case. Alternatively, the probability the order flow contains information decreases with σ_z, making adverse selection less acute, leading the market maker to lower λ and, then, the informed traders to trade more aggressively. Likewise, the insider's informational advantage increases with σ_d, leading the market maker to raise the adverse selection costs λ, and, then, the insider to trade less aggressively.

Price discovery leads to halve the initial uncertainty about the asset payoff,

$$var(d|\, p) = \sigma_d^2 - \frac{cov^2(d,p)}{var(p)} = \frac{1}{2}\sigma_d^2, \tag{10.69}$$

where the first equality follows by the projection theorem, and the second by a direct calculation.

Finally, the expected profits to the trader are, unconditionally,

$$E((d-p)\, x) = \frac{1}{2}\sigma_z\sigma_d.$$

They increase with the insider's informational advantage, measured by σ_d, and with the insider's ability to hide his orders, measured by σ_z.

10.8.2 Markets with multiple traders and dealers

We consider three extensions of the Kyle baseline's model: one in which there are multiple insiders with the same information; a second, in which insiders have differential information; and a third, in which there are multiple dealers.

10.8.2.1 Identical private information

When N traders observe the asset payoff, the market maker will see an order flow equal to, $y \equiv Nx + Z$, and will set the price as in Eq. (10.63). We still conjecture that this price is linear in y, as in Eq. (10.64), and that each insider trades on his information advantage according to,

$$x = \beta \left(d - \bar{d} \right). \tag{10.70}$$

Note, now, that while formulating his trading rule, each insider knows he will have a price impact but that the other agents have price impact too. A Cournot-Nash equilibrium is one in which (i) each trader formulates his optimal trading decision whilst taking the trading rules adopted by his peers as being as in (10.70) and still, (ii) finds (10.70) being optimal to him. The reason this market is referred to as being Cournotian is because it parallels the Cournot (1838) model in Industrial Organization, in which a finite number of firms compete for the same product (and, hence, the same price),[11] and act by maximizing their profits based on "residual demands," just as in the model of this section. Naturally, traders formulate their strategies while also conjecturing that the market maker bases his inference on the aggregate order flow, as explained.

Each trader maximizes his expected profits, as follows:

$$\max_x E\left((\tilde{d} - p)x \middle| \tilde{d} = d \right), \quad p = \bar{d} + \lambda \cdot \left((N-1)\beta \left(d - \bar{d} \right) + x + Z \right),$$

leading to the following first order conditions,

$$x = \underbrace{\frac{1 - \lambda (N-1)\beta}{2\lambda}}_{\equiv \beta} \left(d - \bar{d} \right),$$

where the identity under the brackets follows by the conjecture made in Eq. (10.70). That is,

$$\beta = \frac{1}{(N+1)\lambda}, \tag{10.71}$$

and one determines λ similarly as in (10.66),

$$\lambda = \frac{cov\left(d, Nx + Z\right)}{var\left(Nx + Z\right)} = \frac{N\beta\sigma_d^2}{N^2\beta^2\sigma_d^2 + \sigma_z^2}. \tag{10.72}$$

The solution to Eqs. (10.71)-(10.72) is,

$$\beta = \frac{1}{\sqrt{N}} \frac{\sigma_z}{\sigma_d}, \quad \lambda = \frac{\sqrt{N}}{N+1} \frac{\sigma_d}{\sigma_z}. \tag{10.73}$$

The interpretation regarding σ_z and σ_d is the same as in the previous section. Regarding N, note that in the market of this section, traders trade less aggressively as their number increases. The reason is that the overall price impact of information is $\lambda N \beta = \frac{N}{N+1}$, which is increasing in N, making the probability of an informed trade increasing in N as well. In other

[11] Chamberlin's (1933) model of monopolistic competition deals with markets in which firms compete for products that have imperfect substitutes.

words, as λN increases, β decreases. Intuitively, as more and more traders trade on the same information, price move more and more against them, leading each trader to be less aggressive. Note, however, that markets become more liquid with N, as the expression for λ in (10.73) suggests.

Furthermore, note that while each trader trades less aggressively than in the baseline case where $N = 1$ (see Eqs. (10.68)), price discovery improves as N increases, as a simple calculation reveals,

$$var\left(d\middle|\, p\right) = \sigma_d^2 - \frac{cov^2\left(d,p\right)}{var\left(p\right)} = \frac{1}{N+1}\sigma_d^2.$$

Finally, competition amongst traders lead them to experience lower profits as their number increases,

$$E\left(\left(d-p\right)x\right) = \frac{1}{\sqrt{N}\left(N+1\right)}\sigma_d\sigma_z.$$

We now explain that some of these properties do not hold in markets where traders have heterogenous information.

10.8.2.2 Heterogeneity in private information

We now assume that each trader i observes a different signal s^i, and that each s^i is as in Eq. (10.10). The aggregate order flow is now $y \equiv \sum_{i=1}^{N} x_i + Z$, where each trader's demand x_i reflects his own informational advantage against the market maker. Below, we shall formulate conjectures about how this informational advantage is reflected into the trader's strategy (see Eq. (10.75)). We initially focus on explaining the gaming of this market with heterogeneous information.

First, and as usual, market making makes the price semi-strong informationally efficient, just as in Eq. (10.63). Second, we conjecture that this price is linear in y, as in Eq. (10.64). Third, we analyze this market as being a Cournot-Nash one, as in the previous section: each trader trades so as to maximize his expected profits given the trading strategies of the other players and, the market maker's inference. Accordingly, each trader maximizes his expected profits, as follows:

$$\max_{x_i} E\left(\left(d-p\right)x_i\middle|\, s^i\right), \quad p = \bar{d} + \lambda\cdot\left(\sum_{j\neq i} x_j + x_i + Z\right).$$

It is instructive to analyze the first order conditions of this problem,

$$0 = E\left(d-\bar{d}\middle|\, s^i\right) - \lambda\sum_{j\neq i} E\left(x_j\middle|\, s^i\right) - 2\lambda x_i. \tag{10.74}$$

Note that while determining his optimal asset demand x_i, each agent needs to forecast the demand of his peers given the information he has access to (that is, he is determining $E\left(x_j\middle|\, s^i\right)$), while at the same time fully understanding that his peers are doing exactly the same, in that x_j reflects agent's j expectation of agent i (plus all other agents) expectations given s^j. Forecasting the forecasts of others leads to an "infinite regress" problem, whereby agents end up making conjectures about the conjectures others are making about them, ad infinitum.

There is no guarantee a fixed point exists to this reasoning. However, in a linear equilibrium, a solution exists and is easy to describe. Conjecture that each agent trades on his informational advantage according to,

$$x_i = \beta\left(s^i - \bar{d}\right). \tag{10.75}$$

Then, each agent forecasts of the forecasts of others simplifies, collapsing as it does to forecasting the informational advantages of the peers, $E\left(x_j\middle|\, s^i\right) = \beta E\left(s^j - \bar{d}\middle|\, s^i\right)$. Therefore, in

this linear equilibrium, each agent i can determine his optimal demand from the first order conditions (10.74), as follows:

$$x_i = \frac{E\left(d-\bar{d}\middle|\, s^i\right) - \lambda\beta\left(N-1\right)E\left(s^1-\bar{d}\middle|\, s^i\right)}{2\lambda} = \underbrace{\frac{\rho\left(1-\lambda\beta\left(N-1\right)\right)}{2\lambda}}_{\equiv\beta}\left(s^i-\bar{d}\right) \tag{10.76}$$

where we have used the conjecture that each agent trades according to (10.75), the symmetry $E\left(x_j\middle|\, s^i\right) = \beta E\left(s^1-\bar{d}\middle|\, s^i\right)$ for each j, and the second equality follows because by the projection theorem,

$$E\left(d-\bar{d}\middle|\, s^i\right) = E\left(s^1-\bar{d}\middle|\, s^i\right) = \underbrace{\frac{\sigma_d^2}{\sigma_d^2+\sigma_u^2}}_{\equiv\rho}\left(s^i-\bar{d}\right), \tag{10.77}$$

and, finally, the identity under the brackets in (10.76) follows by the conjecture made in Eq. (10.75).

Note that Eq. (10.77) says that the slope regression estimates of dividends and signals are the same, and equal to ρ. Moreover, ρ is also the correlation coefficient between the signals traders have access to. The model collapses to the identical information market in the previous section once $\sigma_u^2 = 0$, i.e. $\rho = 1$.

By Eq. (10.76),

$$\beta = \frac{\rho}{\lambda\left(2+\rho\left(N-1\right)\right)}, \tag{10.78}$$

and regarding the price impact, the Appendix shows that,

$$\lambda = \frac{cov\left(d,y\right)}{var\left(y\right)} = \frac{N\beta\rho\sigma_d^2}{N\beta^2\sigma_d^2\left(1+\left(N-1\right)\rho\right)+\rho\sigma_z^2}, \tag{10.79}$$

such that,

$$\beta = \sqrt{\frac{\rho}{N}}\frac{\sigma_z}{\sigma_d}, \quad \lambda = \frac{\sqrt{\rho N}}{2+\left(N-1\right)\rho}\frac{\sigma_d}{\sigma_z}. \tag{10.80}$$

In equilibrium, trading aggressiveness, β, increases with the correlation of the traders' signals. Intuitively, in this model, the signals the traders observe become less and less correlated as their quality deteroriates,

$$\rho = \frac{n}{1+n},$$

where $n = \frac{\sigma_d^2}{\sigma_u^2}$ is a measure of the signal's quality. That is, as the noise component increases (i.e. σ_u^2 increases), each trader has access to more and more idiosyncratic, albeit noisy, pieces of information. Each trader then trades less aggressively as his signal becomes less precise.[12]

The price impact in (10.80), λ, is increasing in ρ, provided N is small enough. The mechanism is the following. The market maker anticipates that an increase in the quality of the traders' information (reflected by a higher ρ) results in a more informative trading, which makes adverse selection more severe—λ increases with ρ as a result. In other words, market makers have few traders (N is small), who in addition are equipped with high quality information, and λ increases with ρ as a result. Note that when N is large, λ can be decreasing in ρ provided the latter is

[12] In Section 10.8.3, we discuss alternative signals' structures such that a decrease in correlation does not necessarily imply less precise signals but an increase in monopolistic information power.

large enough. The reason is that as N increases, the order flow become more informative (as $N\beta$ is increasing in N)—informed traders are then "revealing their presence," and an increase in ρ would make their presence revealed even more so to speak: adverse selection costs lower as a result.

Price discovery deteriorates as the correlation decreases, as it can be easily verified that:

$$var\left(d|\,p\right)=\sigma_{d}^{2}-\frac{cov^{2}\left(d,p\right)}{var\left(p\right)}=\frac{2-\rho}{2+\left(N-1\right)\rho}\sigma_{d}^{2}.$$

Finally, the Appendix shows that the expected profits for each trader i are,

$$E\left(\left(d-p\right)x_{i}\right)=\sqrt{\frac{\rho}{N}}\frac{1}{2+\left(N-1\right)\rho}\sigma_{d}\sigma_{z}. \tag{10.81}$$

When N is small, expected profits increase as ρ increases, reflecting the fact that agents observe signals with better quality. As N increases, the expected profits might actually fall with ρ when ρ is sufficiently high, reflecting a lower informational advantage, as explained earlier while commenting on λ in (10.80).

10.8.2.3 Multiple dealers

Kyle (1989) considers a model in which uninformed and informed traders co-exist and act strategically. Foucault, Pagano and Röell (2013) consider a simplified version of this model, in which all traders are risk-neutral, and the "uninformed" traders are interpreted as "dealers." These dealers compete through a call auction mechanism which leads to an equilbrium price such that the dealers make profits on average. As in Kyle (1985), the order flow is $y=\sum_{i=1}^{N}x_{i}+Z$, but in the presence of M dealers acting in an imperfectly competitive market,

$$y=\sum_{j=1}^{M}y_{j}\left(p\right), \tag{10.82}$$

where $y_{j}\left(p\right)$ is the supply schedule submitted by dealer j, a function of the price. The auctioneer sets a price such that the demand schedules of the dealers sum up to the order flow.

Each dealer j maximizes his expected profits taking the other dealers (and the informed traders) actions as given, as in a Cournot-Nash market,

$$\max_{y_{j}}E\left(\left(p-d\right)y_{j}|\,y\right), \tag{10.83}$$

and one constraint that takes into account the nature of imperfect competition amongst dealers, as we now explain.

Assume that the other dealers supply schedules are, for each i,

$$y_{i}\left(p\right)=\gamma\left(p-\bar{d}\right),\quad i\neq j, \tag{10.84}$$

for some γ to be determined in the equilibrium of the game. The meaning of y_i is that dealer i sells if $y_i>0$, and buys if $y_i<0$. Replacing Eq. (10.84) into Eq. (10.82) leaves the residual demand function for any dealer j,

$$p=\bar{d}+\frac{y-y_{j}}{\gamma\left(M-1\right)}. \tag{10.85}$$

The maximization problem faced by each dealer j is therefore that in (10.83), under the constraint (10.85). The solution is,

$$y_j = \frac{1}{2}\left(y - \gamma\left(M-1\right)\left(E\left(d|\, y\right) - \bar{d}\right)\right) = \frac{1}{2}\left(1 - \gamma\left(M-1\right)\lambda\right) y, \tag{10.86}$$

where the last equality follows by the projection theorem and, as usual, and assuming for simplicity a single insider trader,

$$\lambda = \frac{cov\left(d, \beta\left(d-\bar{d}\right)+Z\right)}{var\left(\beta\left(d-\bar{d}\right)+Z\right)} = \frac{\beta\sigma_d^2}{\beta^2\sigma_d^2+\sigma_z^2}. \tag{10.87}$$

The previous expression is the same as the usual Kyle's lambda (see Eq. (10.66)), although it is not the price impact in this model unless M is large.

We determine γ in Eq. (10.84) and then, we search for an equilibrium price by solving Eq. (10.82). First, eliminate y from Eq. (10.85) and Eq. (10.86), such that,

$$y_j\left(p\right) \equiv \underbrace{\frac{\left(1-\gamma\left(M-1\right)\lambda\right)\gamma\left(M-1\right)}{1+\gamma\left(M-1\right)\lambda}}_{\equiv\gamma}\left(p-\bar{d}\right),$$

where the identity under the brackets follows by the conjecture made in Eq. (10.84). Therefore,

$$\gamma = \frac{M-2}{\lambda M\left(M-1\right)}. \tag{10.88}$$

Finally, Eq. (10.82), Eq. (10.84), and Eq. (10.88) imply that the equilibrium price is,

$$p = \bar{d} + \bar{\lambda} y, \quad \bar{\lambda} = \frac{M-1}{M-2}\lambda, \quad M > 2. \tag{10.89}$$

To determine the overall equilibrium, we need to determine β in (10.87), and then, $\bar{\lambda}$ in Eq. (10.89). The usual maximization problem of the insider trader leads to,

$$\beta = \frac{1}{2\bar{\lambda}}. \tag{10.90}$$

Eq. (10.87) and Eq. (10.90) imply that:

$$\beta = \sqrt{\frac{M-2}{M}}\frac{\sigma_z}{\sigma_d}, \quad \bar{\lambda} = \frac{1}{2}\sqrt{\frac{M}{M-2}}\frac{\sigma_d}{\sigma_z}. \tag{10.91}$$

Dealers' market power makes asset markets less deep than in the perfect competition case of Kyle (1985), because the price impact, $\bar{\lambda}$, is decreasing in M as (10.91) reveal. However, price discovery remains the same as in the Kyle (1985) model (see Eq. (10.69)),

$$var\left(d|\, p\right) = \frac{1}{2}\sigma_d^2.$$

Intuitively, the insider trades through half of the market depth both in the perfectly competitive (see (10.65)) and in the imperfectly competitive case (see (10.90)).

Finally, the expected profits to the insider trader are,

$$E\left(\left(d-p\right)x\right) = \frac{1}{2}\sqrt{\frac{M\left(M-2\right)}{\left(M-1\right)^2}}\sigma_z\sigma_d.$$

and are increasing in the number of dealers M, converging to the perfectly competitive case as M gets large.

10.8.3 Dynamic markets

10.8.3.1 Positions in discrete time

Consider an asset that pays off a dividend d at some point in time, N. The asset can be traded in N distinct periods. A trader's cumulative *position* at the end of the trading period is $\theta_N = \sum_{i=1}^{N} \Delta x_i$, where Δx_i denotes the position in the asset at the trading period t (these positions are not part of a self-financed strategy). His final profit is,

$$\tilde{J} \equiv \theta_N d - S,$$

where $S \equiv \sum_{i=1}^{N} \Delta x_i p_i$ denotes the sum of the values of all the positions over the trading period. Therefore,

$$\tilde{J} = \sum_{i=1}^{N} (d - p_i) \Delta x_i. \tag{10.92}$$

Below, we shall utilize the continuous time counterpart to this expression while reviewing a dynamic extension of the Kyle's model.

10.8.3.2 Monopolistic trader

How does the insider trader dilute his information when allowed to trade over different batch auctions? We analyze the continuous time version of Kyle's model, in which trading occurs over a finite horizon fixed to $[0, 1]$. As in the baseline market, the insider trades based on his available information, and the market maker (many of them, actually) updates his beliefs regarding the fundamental asset value d based on having observed the aggregate order flow.

The aggregate cumulative orders at any $t \in [0, 1]$, y_t say, comprise two components: (i) the insider's cumulative demand, $x_t \equiv \int_0^t \theta_\tau d\tau$, and (ii) the cumulative orders by the liquidity traders, $\sigma_z Z_t$, where σ_z is a constant, and Z_t is a standard Brownian motion. Therefore, $y_t = x_t + \sigma_z Z_t$, such that the aggregate order flow satisfies:

$$dy_t = \theta_t dt + \sigma_z dZ_t. \tag{10.93}$$

The market maker does not make any profits while he sets the price according to the semi-strong efficiency rule paralleling that in Eq. (10.63),

$$p_t = E\left(d | \left(y_\tau\right)_{\tau \in [0,t]}\right). \tag{10.94}$$

The insider trader maximizes his expected gains from trade, the continuous time version of $\tilde{J}$ in Eq. (10.92):

$$J\left(t, p_t\right)|_{t=0} \equiv \max_{(\theta_t)_{t\in[0,1]}} E\left(\int_0^1 (\tilde{d} - p_t)\theta_t dt \middle| \tilde{d} = d\right). \tag{10.95}$$

Motivated by the market behavior in the static setting of the previous sections, we conjecture that the optimal trading strategy in (10.95) is

$$\theta_t = \beta_t \left(d - p_t\right), \tag{10.96}$$

for some deterministic function β_t, such that the order flow in Eq. (10.93) can be written as

$$dy_t = \beta_t \left(d - p_t\right) dt + \sigma_z dZ_t. \tag{10.97}$$

To determine the dynamics of the price in Eq. (10.94), note, heuristically, that,

$$\begin{aligned} p_{t+dt} &= E\left(d|\left(y_\tau\right)_{\tau\in[0,t]}, y_{t+dt}-y_t\right) \\ &= E\left(d|\left(y_\tau\right)_{\tau\in[0,t]}\right) + \frac{cov(\,d, y_{t+dt}-y_t|\left(y_\tau\right)_{\tau\in[0,t]})}{var(\,y_{t+dt}-y_t|\left(y_\tau\right)_{\tau\in[0,t]})}\left(y_{t+dt}-y_t\right) \\ &= p_t + \frac{\beta_t\Sigma_t dt}{\beta_t^2\Sigma_t dt^2+\sigma_z^2 dt}\left(y_{t+dt}-y_t\right), \end{aligned} \tag{10.98}$$

where the second equality follows by the projection theorem, and the third by Eq. (10.94), Eq. (10.97) and the following definition of "residual variance,"

$$\Sigma_t \equiv E\left(\left(d-p_t\right)^2\Big|\left(y_\tau\right)_{\tau\in[0,t]}\right). \tag{10.99}$$

The residual variance, Σ_t, can be interpreted as a gauge of informational efficiency, or price discovery process, i.e. the market maker's inference about the asset payoff. We shall see that in equilibrium, price discovery will be complete, in that $\lim_{t\to 1}\Sigma_t = 0$.

In the limit, and disregarding dt^2 terms, the price in (10.98) satisfies,

$$dp_t = \lambda_t dy_t, \quad \lambda_t = \frac{\beta_t\Sigma_t}{\sigma_z^2}. \tag{10.100}$$

We now determine λ_t and Σ_t, and then use (10.100) to determine β_t, thereby having determined the trading strategy θ_t in Eq. (10.96). Later, we shall verify that the thusly determined strategy is indeed optimal. Regarding Σ_t, suppose that dy_t becomes available over the time interval dt. Then,

$$\begin{aligned} \Sigma_{t+dt} &= E\left(\left(d-p_t\right)^2\Big|\left(y_\tau\right)_{\tau\in[0,t+dt]}\right) \\ &= E\left(\left(d-p_t\right)^2\Big|\left(y_\tau\right)_{\tau\in[0,t]}\right) - \frac{cov^2(\,d, y_{t+dt}-y_t|\left(y_\tau\right)_{\tau\in[0,t]})}{var(\,y_{t+dt}-y_t|\left(y_\tau\right)_{\tau\in[0,t]})} \\ &= \Sigma_t - \frac{\beta_t^2\Sigma_t^2 dt^2}{\beta_t^2\Sigma_t dt^2+\sigma_z^2 dt} \\ &= \Sigma_t - \frac{\beta_t^2\Sigma_t^2}{\sigma_z^2}dt, \end{aligned}$$

such that in the limit, and by the expression for λ_t in (10.100),

$$d\Sigma_t = -\lambda_t^2\sigma_z^2 dt. \tag{10.101}$$

We conjecture that $\lambda_t = \lambda$, a constant, and that the equilibrium trading strategy in (10.96) leads to a complete price discovery process, $\lim_{t\to 1}\Sigma_t = 0$. By integrating Eq. (10.101), and using the condition $\Sigma_0 = \sigma_d^2$, leaves $\Sigma_t = \sigma_d^2 - \lambda^2\sigma_z^2 t$, and by the conjecture that $\Sigma_1 = 0$, we find that the price impact in (10.100) is,

$$\lambda = \frac{\sigma_d}{\sigma_z}, \tag{10.102}$$

and, accordingly,

$$\Sigma_t = \sigma_d^2\left(1-t\right). \tag{10.103}$$

That is, the pace of information revelation is constant in this market. Eqs. (10.102), (10.103) and (10.100) now imply that the trading aggressiveness in (10.96) is,

$$\beta_t = \frac{1}{\lambda\,(1-t)} = \frac{\sigma_z}{\sigma_d\,(1-t)}. \tag{10.104}$$

Eqs. (10.102), (10.103) and (10.96) complete the description of the market equilibrium, once we prove the trading strategy θ_t in Eq. (10.96) is optimal with β_t as in (10.96). The proof proceeds in two steps. First, we show that the proposed θ_t leads the price process to converge to the fundamental d ("no money left on the table"); and second, we show that any strategy is optimal when it satisfies this property.

Regarding convergence to fundamental value, note that by replacing (10.93) into (10.100), using the expressions for θ_t in (10.96), and the expressions for λ and β_t in (10.102) and (10.104),

$$dp_t = \frac{d-p_t}{1-t}dt + \sigma_d dZ_t. \tag{10.105}$$

Now, by Karatzas and Shreve (1991, Corollary 6.10, p. 359), the process, $W_t = \frac{p_t - \bar{d}}{\sigma_d}$, is a Brownian bridge from 0 to $\frac{d-\bar{d}}{\sigma_d}$, meaning that, $W_0 = 0$ and $W_1 = \frac{d-\bar{d}}{\sigma_d}$. The latter equality means implies that $p_1 = d$.[13] That is the price is a Brownian bridge under the insider's information although the end-point of it is unknown to the market maker.

Regarding the previous convergence property as being one of the optimal trading strategies, consider the following Hamilton-Jacobi-Bellman equation satisfied by $J\,(t, p_t)$ in Eq. (10.95),

$$0 = \max_{\theta_t} E_t \left((d-p_t)\,\theta_t + J_t\,(t,p_t) + J_p\,(t,p_t)\,\lambda\theta_t + \frac{1}{2}J_{pp}\,(t,p_t)\,\lambda^2\sigma_z^2 \right), \tag{10.106}$$

where the expectation is taken conditional upon the price, and where we have used $dp_t = \lambda dy_t$ and the expression of (10.93) for dy_t. As originally noted by Back (1992), the maximand in (10.106) is linear in θ_t such that the terms multiplying θ_t must sum up to zero, with the remaining terms summing up to zero as well, viz

$$\begin{cases} d - p + J_p\lambda = 0 \\ J_t\,(t,p) + \frac{1}{2}J_{pp}\,(t,p)\,\lambda^2\sigma_z^2 = 0 \end{cases} \tag{10.107}$$

Eqs. (10.107), and the boundary condition $J\,(1, p_1) = 0$, lead to the following expression for the value function in (10.95),

$$J\,(t,p_t) = \frac{1}{2\lambda}\,(d-p_t)^2 + \frac{1}{2}\lambda\sigma_z^2(1-t). \tag{10.108}$$

Next, consider the value function in (10.95) for *any* θ_t, say $J^\theta\,(t,p_t)$. By Itô's lemma,

$$dJ^\theta = \left(LJ^\theta\right)dt + J_p^\theta\lambda_t\sigma_z dZ_t,$$

where LJ^θ denotes the usual infinitesimal generator, which by Eq. (10.95) satisfies $0 = LJ^\theta + (d-p)\,\theta$. Therefore,

$$L_\theta \equiv E\left(\int_0^1 (\tilde{d} - p_t)\theta_t dt \middle| \tilde{d} = d \right) = J^\theta\,(0,p_0) - E\left(J^\theta\,(1,p_1)\right).$$

[13] Note that the insider's cumulative orders x_t are bounded in probability as a result of this convergence. Indeed, by the definition of W_t, we have that $\sigma_d W_t = p_t - \bar{d} = \lambda\,(x_t + \sigma_z Z_t)$: boundedness in probability of W_t and Z_t implies that of x_t.

Note that by evaluating Eq. (10.108) at $t = 1$, we conclude that $J(1, p_1) = \frac{1}{2\lambda}(d - p_1)^2 \geq 0$. This means that for an optimal strategy $\hat{\theta}$ we have that $L_{\hat{\theta}} = J^{\hat{\theta}}(0, p_0) - E(J^{\hat{\theta}}(1, p_1)) \leq J^{\hat{\theta}}(0, p_0)$, and therefore that $E(J^{\hat{\theta}}(1, p_1)) = 0$; because $J(1, p_1) \geq 0$, it follows that $J(1, p_1) = 0$, i.e. $p_1 = d$.

10.8.3.3 Imperfectly competitive traders

Foster and Viswanathan (1996) generalize the dynamic market of Kyle (1985), by relying on a discrete time setting with multiple traders (N say), and a general correlation structure of the signals (still assumed to be Gaussian). They assume that each trader observes a signal correlated with the asset payoff, which has variance Λ_0, and that any two signals have the same correlation equal to $\rho \equiv \frac{\Omega_0}{\Lambda_0}$.[14]

One leading example of signals in this setting is the "exhaustive information structure," arising when the sum of all signals is the truth, i.e. all traders would form a "mega" Kyle trader upon information collusion, $d = \sum_{i=1}^{N} s^i$, where s^i is the signal available to trader i. It is immediate to verify that in this case, for each i,

$$cov(s_i, f) = \Lambda_0 + (M - 1)\Omega_0,$$

and that the variance of the asset payoff is:

$$\sigma_d^2 = N\Lambda_0 + N(N - 1)\rho\Lambda_0. \tag{10.109}$$

Eq. (10.109) can be inverted for ρ, leaving, for all i, j,

$$cov(s_i, s_j) \equiv \rho = \frac{1}{N(N - 1)}\left(\frac{\sigma_d^2}{\Lambda_0} - N\right).$$

The correlation coefficient ρ is a measure of monopolistic information power amongst traders. The lower ρ, the more "unique" a signal is to each trader—and the higher is his monopolistic power. Foster and Viswanathan (1996) show, numerically, that in a dynamic context, traders engage in a *rat race* once ρ is high: because traders have comparable information, they trade very aggressively since the beginning of the trading period to preempt being anticipated by others, but then, revealing virtually all their information over the first trading rounds. In fact, Back, Cao and Willard (2000) consider a continuous time model along the same lines, and conclude that an equilibrium fails to exist when $\rho = 1$ due to this behavior.

However, when ρ is small, and the degree of monopolistic information power is high as a result, traders engage in a *waiting game*, trading little at the beginning of trading period, and aggressively towards the conclusion. These trading patterns lead to a price discovery process, which occurs at a slow pace at the beginning of the trading period, and at a high pace at the end. These patterns of price discovery process cannot be generated by the constant information flow predicted by the single insider trader's model of Kyle (1985), as summarized by Eq. (10.103).

10.8.4 Mandatory disclosure

In practice, U.S. securities laws require that traders disclose their trades. How does this institutional detail affect traders' behavior? Naturally, the Kyle's baseline model cannot address this

[14]The variance-covariance matrix of the signals available to all traders is, then, invertible, provided $\Lambda_0 > -(N - 1)\Omega_0$.

issue because its key feature is that traders don't disclose their trades. Huddart, Hughes and Levine (2001) consider a model with disclosure, in which the informed traders' strategy is to add some noise (i.e., a randomly drawn buy or sell) to the asset demand related to the information they possess, a *mixed strategy*. A mixed strategy allows the insider trader to dissimulate his trade vis-à-vis the market maker, thereby still earning an information related profit.

Consider a three-period market (the model may be extended to a multiperiod setting), in which the insider trades at time-1 and time-2 and must disclose his trade at the close of the first trading round. Disclosure of time-2 trades is economically irrelevant, because the asset pays off at time-3; as usual, we assume the payoff is some $\tilde{d} \sim N\left(\bar{d}, \sigma_d^2\right)$. Noise trading at time $i = 1, 2$ is $Z_i \sim N\left(0, \sigma_z^2\right)$. Prices in all periods are set by the competitive market makers through the usual semi-strong efficiency rule: they are expectations of the asset payoff taken conditionally upon the relevant available information.

The trading timing is the following. At time-1, the insider trades x_1 and the market maker sets a price equal to $p_1 = E(\tilde{d}\,|y_1)$, where $y_1 = x_1 + Z_1$ denotes the time-1 order flow. After trading x_1, the trader discloses x_1, and the market maker updates his estimate of the asset payoff from p_1 to $p_1^* = E(\tilde{d}\,|x_1, y_1) = E(\tilde{d}\,|x_1)$, where the second equality follows because y_1 obviously provides coarser information on $\tilde{d}$ than x_1. In a linear equilibrium, provided it exists,

$$p_1 = \bar{d} + \lambda_1 y_1, \quad p_1^* = \bar{d} + \gamma x_1, \tag{10.110}$$

where λ_1 and γ are determined endogenously. At time-2, the insider trader trades x_2 and the order flow is $y_2 = x_2 + Z_2$. Now, the time-2 price is $p_2 = E(\tilde{d}\,|x_1, y_2)$. In a linear equilibrium, provided it exists,

$$p_2 = p_1^* + \lambda_2 y_2,$$

where λ_2 is determined endogenously.

The informed trader solves his intertemporal problem by backward induction. First, he takes his trade x_1 as given and, then, determines his time-2 trade in order to maximize the second period profits, viz

$$\hat{x}_2 = \arg\max_{x_2} E\left((\tilde{d} - p_2)x_2 \middle| \tilde{d} = d\right) = \beta_2\left(d - p_1^*\right), \quad \beta_2 = \frac{1}{2\lambda_2}. \tag{10.111}$$

Accordingly, his expected profits at time-2 are

$$\pi\left(p_1^*\right) \equiv E\left((\tilde{d} - p_2)\hat{x}_2 \middle| \tilde{d} = d\right) = \frac{\left(d - p_1^*\right)^2}{4\lambda_2},$$

where p_1^* is as in Eq. (10.110) and, is, thus, a function of x_1.

At time-1, the insider chooses his trade at time-1, x_1, while maximizing the expected profits at time-1 and, also, taking into account the "information leakage" onto the second period generated by mandatory disclosure,

$$\hat{x}_1 = \arg\max_{x_1} E\left((\tilde{d} - p_1)x_1 + \pi\left(p_1^*\right) \middle| \tilde{d} = d\right).$$

We conjecture, and later verify, that there exists an equilibrium in which,

$$\hat{x}_1 = \beta_1\left(d - \bar{d}\right) + \eta, \quad \eta \sim N\left(0, \sigma_\eta^2\right), \tag{10.112}$$

where β_1 and σ_η^2 are determined endogenously. The first order conditions yield $0 = (1 - \frac{1}{2}\frac{\gamma}{\lambda_2})\left(d - \bar{d}\right) + (\frac{1}{2}\frac{\gamma^2}{\lambda_2} - 2\lambda_1)x_1$. Substituting Eq. (10.112) into this equation reveals that it must be that

$$\lambda_1 = \lambda_2 \equiv \lambda, \quad \gamma = 2\lambda.$$

To determine λ_2, note that

$$\lambda_2 = \frac{cov(\tilde{d}, y_2 \,|x_1, y_1\,)}{var\,(\,y_2|\,x_1, y_1)} = \frac{cov(\tilde{d}, y_2\,|x_1\,)}{var\,(\,y_2|\,x_1)} = \frac{\beta_2 \sigma^2_{d|x_1}}{\beta_2^2 \sigma^2_{d|x_1} + \sigma_z^2}, \tag{10.113}$$

where $\sigma^2_{d|x_1} = var(\tilde{d}\,|x_1\,) = \frac{1}{2}\sigma_d^2$ (see Appendix 2).

Therefore, $\beta_2 = \frac{1}{2\lambda_2}$ in (10.111) and (10.113) can be solved for explicitly, just as in the baseline Kyle's model (see Eqs. (10.67)), leaving

$$\beta_2 = \frac{\sigma_z}{\sigma_{\,d|x_1}}, \quad \lambda_2 = \frac{1}{2}\frac{\sigma_{\,d|x_1}}{\sigma_z}.$$

Finally, the Appendix proves that $\beta_1 = \frac{1}{2}\beta_2$ and that the variance of the noise in the mixed strategy is $\sigma_\eta^2 = \beta_1^2 \sigma_d^2$.

Note that the insider trader sets the total variance of his trade equal to the noise variance,

$$var(\hat{x}_1) = \sigma_z^2.$$

Moreover, it is possible to show that price discovery and other market quality variables improve with mandatory disclosure, if compared with the Kyle case.

10.9 Limits to arbitrage and further frictions

[In progress]

10.9.1 Liquidity trading

The assumption of noise traders. Discussion

10.9.2 Price impacts and derivatives

Price pressures and demands for derivatives: Gârleanu, Pedersen and Poteshman (2009).

10.9.3 Arbitrage imperfections

Irrational traders: DeLong, Shleifer, Summers and Waldman (1990). Risky arbitrage. Constrained arbitrage: equity (Gromb and Vayanos, 2002). Preferred habitat and the yield curve. Greenwood and Vayanos (2014). Capital immobility.

10.10 Over-the-counter markets

10.10.1 Background

Over-the-counter (OTC) trading venues are pervasive in financial markets, and cover products such as corporate, government and municipal bonds; interest rate derivatives (e.g., interest rate swaps, swaptions, caps, floors); currencies, credit products such as collateralized debt obligations, credit default swaps; or large blocks of equities. Many of these products are actually reviewed, and priced, in great detail in Part III (see, e.g., Chapters 13 and 14). Duffie (2012) is a very important introduction to much of the economics underlying OTC markets.

OTC markets differ from centralized exchanges as they operate through peculiar pricing mechanisms. In these markets, for example, traders may only be aware of "local" information: some traders might be negotiating a deal for a product at a time and geographical venue while ignoring information regarding a similar product in other areas. Once a trader receives a price quote (bid or ask) from a dealer (i.e., an entity that maintains a two-way market), he may well continue to search for better conditions and other counterparties. The dealer may revise his quotes over time, based on public news (e.g., news on broad macroeconomic developments) and the evidence gathered while maintaining his own market segment. Quotes may occur by phone or electronic communication, including screen-query systems. At the time of writing, brokerage is also important in some of these markets (e.g., interest rate swap markets)—brokers intermediate while not trading on their own accounts. Interdealer brokers are institutions that specialize in connecting dealers.

Why do OTC markets exist? Some financial products are particularly complex and, sometimes, tailored to the need of clients requesting for them. Only a few specialists may be able to understand risks, and trade, products originated through such bespoke deals. Centralizing these products on an exchange would not necessarily attract a more widespread market: widespread markets necessitate standardization and simplicity. By construction, bespoke products are not standardized. Moreover, they may be too complex.

Products that are initially bespoke may, however, gain the interest of other market participants, and trigger volumes and liquidity that may justify standardization and the introduction to an exchange. The typical example that is often discussed is a product that comes from the interest rate derivative markets. At the end of June 2015, the notional amount of outstanding interest rate derivatives amounted to $435 trillion, approximately 80% of the global OTC derivatives market. Interest rate swaps make $320 trillion of these figures, the largest share in this space.

There are many reasons that justify these enormous figures. One is certainly related to widespread hedging motives, discussed in detail in Chapter 13. For example, pension funds would like to hold assets paying a flow of fixed payments, payments that are needed to honour their liability sides. An interest rate swap contract would allow a pension fund to receive fixed payments from a counterparty (say an investment bank) against floating rates (obtained, perhaps, by investing current fund contributions), and over long horizon spans. The pension fund is, then, a "swap receiver." Chapter 13 explains reasons that may attract institutions to be on the other side of the deal (being, thus, long a "swap payer"). Now, a bank that gathers a deal from a client (say a swap "payer" denominated in U.S. dollars), may not want to carry this deal on its book. Our bank, may, then, search for counterparties willing to take on this product; one simple solution would be to offset this position while initiating a new swap "receiver" with another counterparty. In turn, the latter may dislike to hold a payer position, and

may successfully search for counterparties and initiate new receivers. The process may continue, contributing to a large market.

An exchange would probably eliminate some of this business, due to its "netting" ability. It would also eliminate counterparty risk, that is, the risk of default by players in these markets. In OTC markets, counterparty risk can become systemic risk when the affected players have a core role in the trading network. To illustrate, the Lehman Brothers failure in 2008 spread concerns on such networks effects so dark to literally stop the financial system for several months, contributing to business cycle developments of the time. At the time of writing, such interest rate derivatives are still traded OTC. It is not clear how resilient a centralized venue would be to large shocks, even accounting for its netting abilities—an exchange is subject to default risk anyway. Moreover, OTC markets are not necessarily illiquid. In fact, swap markets are an example of a very liquid market, and a standardized one. We have swaps for any reasonable maturity, and derivatives on these swaps (called swaptions) that are equally liquid and standardized. But, again, this liquidity may suddenly disappear in case of systemic events.

10.10.2 Search

The next section analyzes price mechanics in these markets that differ from an exchange. We focus on what arguably is the most important characteristics of OTC markets: search. The previous discussion suggests that in the absence of centralized information, searching for trading counterparts is vital in OTC markets. We describe a model of search based on work by Duffie, Gârleanu and Pedersen (2007), in which traders need to search for counterparts. Search is a friction that leads to quite novel predictions regarding price formation in market microstructure: notably, search leads to price discounts (compared to an otherwise purely Walrasian market) and make prices recover slowly after an aggregate liquidity shock. The price discount is a liquidity premium and increases with searching frictions.

In a related model with market makers, Duffie, Gârleanu and Pedersen (2005) do in fact show that bid-ask spreads decrease with the ability of traders to find trading counterparties. In other words, a sophisticated trader has access to good trading conditions, which is somehow the opposite to conclusions achieved by information-based explanations of market microstructure. The next section presents the simplest version of Duffie, Gârleanu and Pedersen (2007), a model without market making.

10.10.3 A model with symmetric information

Many products that are traded in OTC markets are reviewed in Part III (e.g., Chapters 13 and 14), as explained, and will be priced while partially ignoring the market mechanisms inherent in OTC markets. One embarassing reason for this treatment is that academic research and market practice have only relatively recently begun dialoguing regarding the details of these markets. However, Part III will illustrate the main efforts made in the market practice to incorporate the main imperfections that plague these markets (e.g., counterparty risk, or "multiple curves" in interest rate swap markets). This section only aims at providing a simple framework to understand the impact of search frictions on prices.

We consider an infinite horizon, continuous time market for a consol, an asset that yields an instantaneous payoff equal to one unit of consumption. Traders are risk neutral and we take their discounting rate equal to the riskless interest rate, r. There are four types of traders: traders with a low valuation of the asset who own (lo) or don't own (ln) the asset, and traders with a high valuation of the asset who own (ho) or don't own (hn) the asset. Agents with

low valuations incur an instantaneous holding cost equal to δ. The model does not incorporate counterparty risk.

Traders may switch their types from low to high, with intensity λ_u; and from high to low, with intensity λ_d. Finally, any agent is matched to any other with intensity λ. Whether these agents will trade depends on their preferences and asset ownerwship, as elaborated below. Traders' masses are denoted with $\mu_{\mathcal{T}}$, $\mathcal{T} \in \{lo, ln, ho, hn\}$, with obvious notation, and they sum up to one. Per capita supply is denoted with s, $s = \mu_{lo} + \mu_{ho}$. A transaction may only occur if a lo (resp., hn) trader meets a hn (resp., lo) trader, which occurs with intensity $\lambda\mu_{hn}$ (resp., $\lambda\mu_{lo}$). The objective of the model is to determine the transaction price, P say.

The law of movement for $\mu_{\mathcal{T}}$ is

$$\begin{cases} \dot{\mu}_{lo} = -\lambda\mu_{hn}\mu_{lo} - \lambda_u\mu_{lo} + \lambda_d\mu_{ho} \\ \dot{\mu}_{ln} = \lambda\mu_{hn}\mu_{lo} - \lambda_u\mu_{ln} + \lambda_d\mu_{hn} \\ \dot{\mu}_{ho} = \lambda\mu_{hn}\mu_{lo} - \lambda_d\mu_{ho} + \lambda_u\mu_{lo} \\ \dot{\mu}_{hn} = -\lambda\mu_{hn}\mu_{lo} - \lambda_d\mu_{hn} + \lambda_u\mu_{ln} \end{cases} \tag{10.114}$$

Consider, for example, the second equation of (10.114). The mass of ln traders, (i) increases by $\lambda\mu_{hn}\mu_{lo}$, the aggregate meeting rate of lo and hn traders (the first term on the R.H.S.), (ii) decreases by $\lambda_u\mu_{ln}$, the mass of ln traders switching to hn type, (iii) increases by $\lambda_d\mu_{hn}$, the mass of hn traders switching to ln type. A steady state of this system in defined as $\mu_{\mathcal{T}} : \dot{\mu}_{\mathcal{T}} = 0$ for all $\mathcal{T}$.

Each trader's continuation utility equals current wealth plus a component $V_{\mathcal{T}}$ that depends on his type $\mathcal{T}$. To determine these components in steady state, note that, given risk neutrality, and under regularity conditions, the indirect utility grows at at r, such that[15]

$$\begin{cases} rV_{lo} = \lambda_u (V_{ho} - V_{lo}) + \lambda\mu_{hn} (P - \Delta V_l) + 1 - \delta \\ rV_{ln} = \lambda_u (V_{hn} - V_{ln}) \\ rV_{ho} = \lambda_d (V_{lo} - V_{ho}) + 1 \\ rV_{hn} = \lambda_d (V_{ln} - V_{hn}) + \lambda\mu_{lo} (\Delta V_h - P) \end{cases} \tag{10.115}$$

where

$$\Delta V_l \equiv V_{lo} - V_{ln}, \quad \Delta V_h \equiv V_{ho} - V_{hn}.$$

Consider the first equation of (10.115). The L.H.S. is the expected growth of the lo trader's continuation utility. The first term on the R.H.S. accounts for the jump in the utility the lo trader would experience in case he switched from low to high type; the second term accounts for the jump in his utility, should he meet a trading counterparty (i.e., a hn trader): the jump in utility would equal the price received from the transaction, P, plus the (negative) change in utility deriving from switching from lo to ln, i.e., $-\Delta V_l$; finally, the term $1 - \delta$ is the consol instantaneous payoff net of its holding costs. The remaining equations are interpreted similarly.

Note that ΔV_h is the asset reservation value for a hn trader (the buyer) and ΔV_l is the minimum price required by a lo trader (the buyer). Therefore, any transaction price $P \in [\Delta V_l, \Delta V_h]$. It is assumed that, once two traders with opposite needs meet, they agree on a price achieved through bilateral bargaining,

$$P = (1 - q)\,\Delta V_l + q\Delta V_h, \tag{10.116}$$

[15] Eqs. (10.115) are Bellman equations evaluated in steady state when any trader type has utility equal to current wealth plus $E_t(\int_t^\infty e^{-r(\tau-t)} c_\tau d\tau)$ subject to budget constraint, where c_t denotes instantaneous consumption; randomness is due to types being subject to switching and matching processes.

where q is an exogenous parameter: the higher q, the higher the bargaining power of the seller.

In the Appendix, we show that Eqs. (10.115) and (10.116) imply that

$$P = \frac{1}{r} - \frac{\delta}{r} \frac{\lambda_d + (r + \lambda \mu_{lo})(1 - q)}{r + \lambda_u + \lambda_d + \lambda ((1 - q)\mu_{lo} + q\mu_{hn})}. \tag{10.117}$$

The model predicts that the asset price equals its console value (the price in a Walrasian market) minus an illiquidity premium. The illiquidity premium is increasing in δ, the holding costs and, for given μ_{lo} and μ_{hn}, decreasing in λ. We do indeed expect the illiquidity premium to increase with searching costs. Moreover, the price is decreasing in λ_d and increasing in λ_u, *ceteris paribus*. These conclusions hold while taking μ_{lo} and μ_{hn} fixed. However, the masses of *lo* and *hn* traders are endogenous, and evolve through Eqs. (10.114). The Appendix shows that, in steady state μ_{lo} is solution to the quadratic equation,

$$0 = \lambda \mu_{lo}^2 + \left(\lambda \left(\frac{\lambda_u}{\lambda_u + \lambda_d} - s \right) + (\lambda_u + \lambda_d) \right) \mu_{lo} - \lambda_d s \tag{10.118}$$

and that (see Eq. (10A.36) in the Appendix) μ_{hn} is given by

$$\mu_{hn} = \mu_{lo} + \frac{\lambda_u}{\lambda_u + \lambda_d} - s. \tag{10.119}$$

Figure 10.2 depicts the equilibrium price in Eq. (10.117) as a function of the average time needed to sell (in days), determined below, with μ_{lo} and μ_{hn} given by (10.118)-(10.119). We set $\delta = 0.3$, and use the intensity parameters values of Duffie, Gârleanu and Pedersen (2007),[16] the reference period taking equal to a year. Thus, for example, a value of $\lambda = 1250$ means that there is an expected number of meetings equal to 1250 per year, or 5 per day;[17] on average, then, a *lo* trader expects to meet $\lambda \mu_{hn} = 139.75$ traders every year, implying an average time to sell equal to almost two days. These values are obtained with $\mu_{hn} = 0.1118$ (the steady state resulting in this example), and Figure 10.2 extends price calculations to other average waiting times for selling.

[16] This section follows the convention of Duffie (2010), whereby the values of λ are twice those in Duffie, Gârleanu and Pedersen (2007).

[17] For any Poisson model with intensity λ, the expected number of events on a time interval $T - t$ is equal to $\lambda (T - t)$ (see Section 5.10 in Chapter 5). Fixing $T - t = 1$ is tantamount to "annualize" λ.

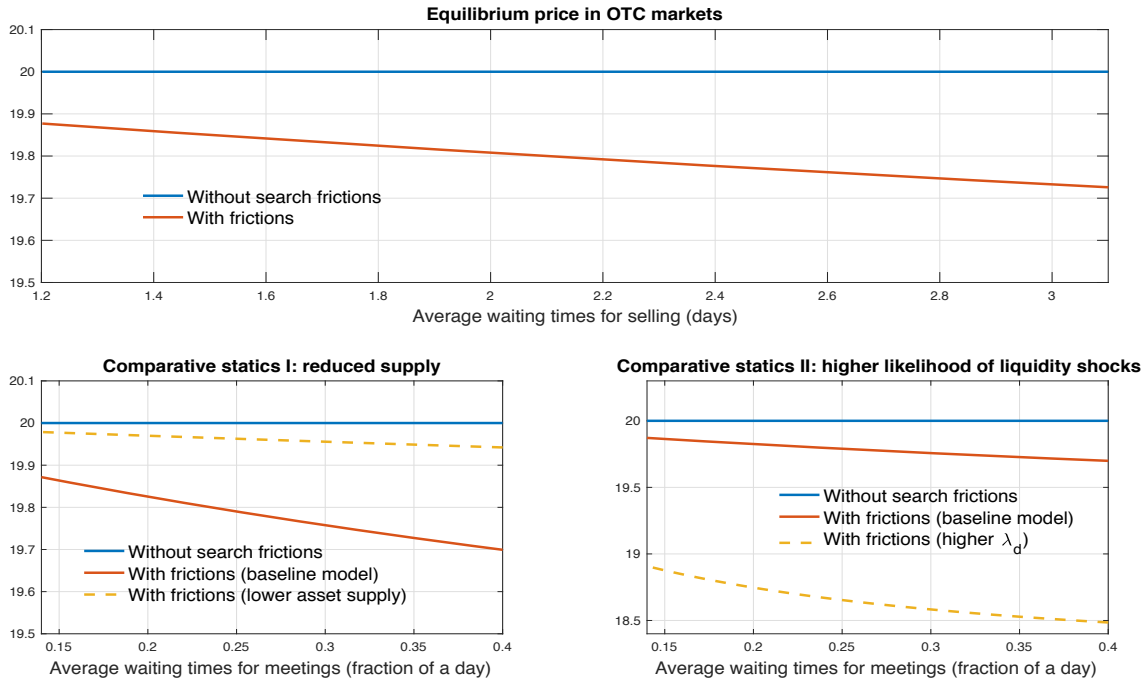


FIGURE 10.2. The top panel depicts the equilibrium price in Eq. (10.117), as a function of the average waiting times for selling (days), defined as $(\mu_{hn}\lambda\Delta t)^{-1}$, where $\Delta t = \frac{1}{250}$. Parameter values are $r = 5\%$, $\lambda_d = 0.5$, $\lambda_u = 5$, $s = 0.8$, $q = 0.5$. The bottom panels depict comparative statics resulting from a lower asset supply ($s = 0.6$, left) and a higher likelihood of liquidity shock ($\lambda_d = 1$, right) than in the baseline model; in both panels, the equilibrium prices are expressed as a function of the average waiting times for a meeting, expressed as a fraction of a day, $(\lambda\Delta t)^{-1}$.

The model predicts that the price is increasing in λ, and confirms the intuition that the illiquidity premium increases with the intensity of liquidity shocks, λ_d. Furthermore, the model predicts that the price increases as the asset supply s lowers. One of the most remarkable predictions is that illiquidity discounts decrease with search frictions: the higher λ, the closer the OTC price is to its Walrasian counterpart. Finally, the price is increasing in q, the bargaining power of seller. People who want to sell will get price closer to their own valuations provided they have enough bargaining power. Duffie, Gârleanu and Pedersen (2007) further examine the effects of an aggregate liquidity shock, obtained while tilting the ergodic distribution of the traders towards one with many of the low-type, finding that prices may recover quite slowly following a liquidity shock.

[Discuss empirical evidence]

10.11 Coordination failures in financial markets

[In progress]

10.11.1 Banking crises

Related to sunspots and related to business cycle.

10.11.2 Higher order beliefs and beauty contests

Common knowledge and determinacy of equilibrium.

10.12 Appendix 1: The projection theorem

Consider the following Gaussian environment, in which a random vector is normally distributed:

$$\left[\begin{array}{c} y \\ x \end{array}\right] \sim N\left(\left[\begin{array}{c} \mu_y \\ \mu_x \end{array}\right], \left[\begin{array}{cc} \Sigma_y & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_x \end{array}\right]\right).$$

We have,

$$E\left(y|\, x\right) = \mu_y + \Sigma_{yx}\Sigma_x^{-1}\left(x - \mu_x\right) \tag{10A.1}$$

and

$$var\left(y|\, x\right) = \Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}. \tag{10A.2}$$

A proof of this result can be obtained as follows. Consider the following "regression"

$$\left(y - \mu_y\right) = B\left(x - \mu_x\right) + \epsilon, \tag{10A.3}$$

where ϵ is zero-mean, and orthogonal to x. By taking covariances in Eq. (10A.3), $\Sigma_{yx} = B\Sigma_x$, which yields Eq. (10A.1), and by taking variances in Eq. (10A.3),

$$-B\Sigma_x B^\top + \Sigma_y = var\left(\epsilon\right) = var\left(y|\, x\right).$$

Plugging $B = \Sigma_{yx}\Sigma_x^{-1}$ into the previous equation leaves Eq. (10A.2).

We apply the projection theorem to study inference problems in which the conditioning variable x is a vector of signals, $x \equiv \left[s^1, \cdots, s^n\right]$, where,

$$s^i = y + u^i, \quad u \sim N\left(0, \sigma_u^2\right), \quad i = 1, \cdots, n. \tag{10A.4}$$

When $n = 1$,

$$E\left(y|\, s\right) = \mu_y + \frac{\sigma_y^2}{\sigma_y^2 + \sigma_u^2}\left(s - \mu_y\right), \quad var\left(y|\, s\right) = \sigma_y^2\frac{\sigma_u^2}{\sigma_y^2 + \sigma_u^2}.$$

We can express the previous relations in terms of precision of the signals,

$$E\left(y|\, s\right) = \mu_y + \frac{\pi_u}{\pi_y + \pi_u}\left(s - \mu_y\right), \quad var\left(y|\, s\right) = \frac{1}{\pi_y + \pi_u}, \tag{10A.5}$$

where $\pi_\cdot$ denotes the precision of a random variable, for example, $\pi_u = \frac{1}{\sigma_u^2}$. Eq. (10A.5) is easily generalized to the case of multiple signals in (10A.4),

$$E\left(y|\, s^1, \cdots, s^n\right) = \mu_y + \sum_{i=1}^n \frac{\pi_{u_i}}{\pi_y + \sum_{j=1}^n \pi_{u_j}}\left(s^i - \mu_y\right), \tag{10A.6a}$$

$$var\left(y|\, s^1, \cdots, s^n\right) = \frac{1}{\pi_y + \sum_{i=1}^n \pi_{u_i}}. \tag{10A.6b}$$

Eqs. (10A.6a)-(10A.6b) simplify once the precisions of the signals are all the same, $\pi_{u_i} = \pi_u$ for all i, leaving,

$$E\left(y|\, s^1, \cdots, s^n\right) = \mu_y + \frac{n\pi_u}{\pi_y + n\pi_u}\left(\hat{s} - \mu_y\right), \quad var\left(y|\, s^1, \cdots, s^n\right) = \frac{1}{\pi_y + n\pi_u},$$

where $\hat{s}$ denotes the average signal,

$$\hat{s} \equiv \frac{1}{n}\sum_{i=1}^n s^i.$$

10.13 Appendix 2: Details regarding solutions of selected models

Proof of Eq. (10.35) (Grossman and Stiglitz (1980) model). We, first, determine the expected utility of the would-be uninformed investors. By the Law of Iterated Expectations, it is:

$$\mathcal{U}_U = E\left(E\left(-e^{-\tau W_U}\middle|\, p\right)\right) = E\left(-e^{-\tau E(W_U|p)+\frac{1}{2}\tau^2 Var(W_U|p)}\right),$$

where, by the expression of $\hat{x}_U$ in (10.26),

$$E\left(W_U|\, p\right) = \left(W_0+\omega_0 p\right)R + \gamma\pi_{\,d|p}\left(E\left(d|\, p\right)-Rp\right)^2, \quad Var\left(W_U|\, p\right) = \gamma^2\pi_{\,d|p}\left(E\left(d|\, p\right)-Rp\right)^2,$$

such that,

$$\mathcal{U}_U = E\left(-e^{-\tau(W_0+\omega_0 p)R}\cdot e^{-\frac{1}{2}\pi_{\,d|p}(E(d|p)-Rp)^2}\right). \tag{10A.7}$$

Instead, by the Law of Iterated Expectations, the expected utility of the would-be informed investors before accounting for information costs is

$$\begin{aligned}\mathcal{U}_I &= E\left(E\left(-e^{-\tau W_I}\middle|\, p,s\right)\right)\\ &= E\left(-e^{-\tau(W_0+\omega_0 p)R}\cdot e^{-\frac{1}{2}\pi_{\,d|s}(E(d|s)-Rp)^2}\right)\\ &= E\left(-e^{-\tau(W_0+\omega_0 p)R}\cdot E\left(e^{-\frac{1}{2}\pi_{\,d|s}(E(d|s)-Rp)^2}\middle|\, p\right)\right).\end{aligned} \tag{10A.8}$$

To determine the inner expectation in the last line of (10A.8), we rely on the following distributional result, shown below,

$$\begin{aligned}\sqrt{\pi_{\,d|s}}\left(E\left(d|\, s\right)-Rp\right)\Big|_p &\sim N\left(E,V\right),\\ E \equiv \sqrt{\pi_{\,d|s}}\left(E\left(d|\, p\right)-Rp\right), &\quad V \equiv \frac{\pi_{\,d|s}}{\pi_{\,d|p}}-1.\end{aligned} \tag{10A.9}$$

Now, it is well-known that if $x \sim N\left(E,V\right)$, then

$$E\left(e^{-\frac{1}{2}x^2}\right) = \frac{1}{\sqrt{1+V}}e^{-\frac{1}{2}\frac{E^2}{1+V}}.$$

Applying this to determine the inner expectation in (10A.8) yields

$$\mathcal{U}_I = \sqrt{\frac{\pi_{\,d|p}}{\pi_{\,d|s}}}E\left(-e^{-\tau(W_0+\omega_0 p)R}\cdot e^{-\frac{1}{2}\pi_{\,d|p}((E(d|p)-Rp))^2}\right) = \sqrt{\frac{\pi_{\,d|p}}{\pi_{\,d|s}}}\cdot\mathcal{U}_U,$$

where the second equality follows by the expression of $\mathcal{U}_U$ in Eq. (10A.7). Eq. (10.35) follows by the previous expression of $\mathcal{U}_I$.

We are left to show that (10A.9) holds true. Regarding the conditional expectation E, note that $E\left(d|\, s\right) = E\left(d|\, s,Z\right)$, such that, by the Law of Iterated Expectations and the fact that the information content of p is coarser than that of (s,Z),

$$E\left(E\left(d|\, s\right)|\, p\right) = E\left(E\left(d|\, s,Z\right)|\, p\right) = E\left(d|\, p\right).$$

Regarding the expression for V in (10A.9), note that due to the normality of d,s and p, we have that:

$$\begin{aligned}\sigma^2_{d|p} &\equiv var\left(d|\, p\right)\\ &= var\left(E\left(d|\, s\right)|\, p\right) + var\left(d-E\left(d|\, s\right)|\, p\right)\\ &= var\left(E\left(d|\, s\right)|\, p\right) + \sigma^2_{d|s},\end{aligned}$$

where the last line follows because $\sigma^2_{d|s}$ is independent of p. That is,

$$V = \pi_{d|s} \cdot var\left(E\left(d|\, s\right)|\, p\right) = \pi_{d|s} \cdot \left(\sigma^2_{d|p} - \sigma^2_{d|s}\right),$$

whence, the expression in (10A.9).

The equilibrium price in Eq. (10.41) regarding the differential information model. By the projection theorem, the two coefficients appearing in the conditional expectation of (10.39) are,

$$\begin{aligned}\begin{bmatrix} \chi_s \\ \chi_p \end{bmatrix}^\top &= \begin{bmatrix} cov\left(d, s^i\right) & cov\left(d, p\right) \end{bmatrix} \begin{bmatrix} var\left(s^i\right) & cov\left(s^i, p\right) \\ cov\left(s^i, p\right) & var\left(p\right) \end{bmatrix}^{-1} \\ &= \begin{bmatrix} cov\left(d, s^i\right) & cov\left(d, p\right) \end{bmatrix} \frac{1}{var\left(s^i\right) var\left(p\right) - cov^2\left(s^i, p\right)} \begin{bmatrix} var\left(p\right) & -cov\left(s^i, p\right) \\ -cov\left(s^i, p\right) & var\left(s^i\right) \end{bmatrix},\end{aligned} \tag{10A.10}$$

where, given the assumption on s^i in (10.10) and the conjectured equilibrium price in (10.40),

$$\begin{array}{lcl} cov\left(d, s^i\right) & = & \sigma^2_d \\ cov\left(d, p\right) & = & b_s \sigma^2_d \\ var\left(s^i\right) & = & \sigma^2_d + \sigma^2_u \\ cov\left(s^i, p\right) & = & b_s\left(\sigma^2_d + \frac{1}{n}\sigma^2_u\right) \\ var\left(p\right) & = & b_s^2\left(\sigma^2_d + \frac{1}{n}\sigma^2_u\right) + b_z^2\sigma^2_z \end{array} \tag{10A.11}$$

and the fourth equality follows because,

$$cov\left(s^i, p\right) = b_s\left(\frac{1}{n}\sum_{i=1}^{n} cov\left(s^1, s^i\right)\right) = b_s \frac{var\left(s^1\right) + (n-1)\, cov\left(s^1, s^2\right)}{n} = b_s \frac{\sigma^2_d + \sigma^2_u + (n-1)\,\sigma^2_d}{n}.$$

Plugging (10A.11) into Eq. (10A.10), leaves, after tedious but straightforward calculations,

$$\begin{cases} \chi_s &= \dfrac{1}{\Delta} b_z^2 \sigma_z^2 \sigma_d^2 \\ \chi_p &= \dfrac{1}{\Delta} b_s \sigma_d^2 \left(1 - \dfrac{1}{n}\right) \sigma_u^2 \end{cases} \tag{10A.12}$$

where

$$\Delta \equiv var\left(s^i\right) var\left(p\right) - cov^2\left(s^i, p\right) = b_s^2\left(1 - \frac{1}{n}\right)\sigma_u^2\left(\sigma_d^2 + \frac{1}{n}\sigma_u^2\right) + b_z^2\sigma_z^2\left(\sigma_d^2 + \sigma_u^2\right). \tag{10A.13}$$

Next, replace (10A.12) into the expressions of the price coefficients, Eqs. (10.43), which leaves the following expressions for b_s and b_z:

$$b_s = \frac{b_z^2\sigma_z^2\sigma_d^2}{R\Delta - b_s\sigma_d^2\left(1 - \frac{1}{n}\right)\sigma_u^2}, \quad b_z = \frac{\Delta}{\gamma \pi_{d|s,p}\left(R\Delta - b_s\sigma_d^2\left(1 - \frac{1}{n}\right)\sigma_u^2\right)}. \tag{10A.14}$$

We determine the conditional precision, $\pi_{d|s,p}$, prior to determining b_s and b_z in (10A.14). By the projection theorem,

$$\frac{1}{\pi_{d|s,p}} = \sigma_d^2 - \begin{bmatrix} \chi_s & \chi_p \end{bmatrix} \begin{bmatrix} cov\left(d, s^i\right) \\ cov\left(d, p\right) \end{bmatrix} = \frac{1}{\Delta}\left(b_s^2 \frac{1}{n}\left(1 - \frac{1}{n}\right)\sigma_u^2 + b_z^2\sigma_z^2\right)\sigma_u^2\sigma_d^2, \tag{10A.15}$$

where the second equality follows by (10A.11), the expressions for χ_s and χ_p in (10A.12), and that for Δ in (10A.13). Replacing Eq. (10A.15) into the expression for b_z in (10A.14), and using (10A.13), leaves the following expressions for the asset price coefficients and the conditional precision,

$$\begin{cases} b_s &= \dfrac{b_z^2\sigma_z^2\sigma_d^2}{R\Delta - b_s\sigma_d^2\left(1-\frac{1}{n}\right)\sigma_u^2} \\ b_z &= \dfrac{b_s^2\frac{1}{n}\left(1-\frac{1}{n}\right)\sigma_u^2 + b_z^2\sigma_z^2}{\gamma\left(R\Delta - b_s\sigma_d^2\left(1-\frac{1}{n}\right)\sigma_u^2\right)}\sigma_u^2\sigma_d^2 \end{cases} \tag{10A.16}$$

and

$$\pi_{d|s,p} = \frac{b_s^2\left(1-\frac{1}{n}\right)\sigma_u^2\left(\sigma_d^2+\frac{1}{n}\sigma_u^2\right) + b_z^2\sigma_z^2\left(\sigma_d^2+\sigma_u^2\right)}{\left(b_s^2\frac{1}{n}\left(1-\frac{1}{n}\right)\sigma_u^2 + b_z^2\sigma_z^2\right)\sigma_u^2\sigma_d^2}. \tag{10A.17}$$

To show that there exists a unique solution for b_s and b_z, define the constant ψ through the following relation,

$$\psi \equiv \frac{b_z}{b_s} = \frac{b_s^2\frac{1}{n}\left(1-\frac{1}{n}\right)\sigma_u^2 + b_z^2\sigma_z^2}{\gamma b_z^2\sigma_z^2}\sigma_u^2 = \frac{\frac{1}{n}\left(1-\frac{1}{n}\right)\sigma_u^2 + \sigma_z^2\psi^2}{\gamma\sigma_z^2\psi^2}\sigma_u^2. \tag{10A.18}$$

where the first equality follows by Eqs. (10A.16) and the second by the very same definition of ψ. It is easy to see that there exists a unique solution for ψ to Eq. (10A.18), ψ_* say. Given ψ_*, we determine b_s in (10A.16) by expressing it as a function of ψ^*, and relying on the definition of Δ in (10A.13), as follows:

$$b_s = \frac{b_z^2\sigma_z^2\sigma_d^2}{R\Delta - b_s\sigma_d^2\left(1-\frac{1}{n}\right)\sigma_u^2} = \frac{\psi_*^2\sigma_z^2\sigma_d^2}{R\kappa - \frac{1}{b_s}\sigma_d^2\left(1-\frac{1}{n}\right)\sigma_u^2}, \tag{10A.19}$$

where

$$\kappa \equiv \left(1-\frac{1}{n}\right)\sigma_u^2\left(\sigma_d^2+\frac{1}{n}\sigma_u^2\right) + \psi_*^2\sigma_z^2\left(\sigma_d^2+\sigma_u^2\right).$$

The solution for b_s, and then b_z, are obtained after rearranging terms in (10A.19), and using the definition of ψ,

$$b_s \equiv \frac{\sigma_d^2\left(1-\frac{1}{n}\right)\sigma_u^2 + \psi_*^2\sigma_z^2\sigma_d^2}{R\kappa}, \quad b_z = b_s\psi_*.$$

Eq. (10.44) follows after using the definitions of precisions in the previous expression, and after rearranging terms.

The limiting case in Eq. (10.46) is obtained after taking the limits in Eq. (10.44) for n large, and noting that,

$$\lim_{n\to\infty}\psi_* = \bar{\psi} \equiv \frac{\sigma_u^2}{\gamma}.$$

Finally, the limiting conditional precision in (10.47) is obtained after taking the limit in Eq. (10A.17) for n large, and using the previous definition of $\bar{\psi}$.

Proof of Eq. (10.79). By Eq. (10.75), the order flow is,

$$y = \sum_{i=1}^{N} x_i + Z = N\beta\left(\hat{s} - \bar{d}\right) + Z, \tag{10A.20}$$

where $\hat{s}$ denotes the average signal,

$$\hat{s} \equiv \frac{1}{N}\sum_{i=1}^{N} s^i = d + \frac{1}{N}\sum_{i=1}^{N} u^i. \tag{10A.21}$$

We have,

$$cov\left(y,d\right)=N\beta\sigma_{d}^{2}, \tag{10A.22}$$

and,

$$var\left(y\right)=N^{2}\beta^{2}var(\hat{s})+\sigma_{z}^{2}=N^{2}\beta^{2}\frac{N\sigma_{s}^{2}+N\left(N-1\right)\rho\sigma_{s}^{2}}{N^{2}}+\sigma_{z}^{2}. \tag{10A.23}$$

Eq. (10.79) follows by Eqs. (10A.22)-(10A.23), and by,

$$\sigma_{s}^{2}\equiv\sigma_{d}^{2}+\sigma_{u}^{2}=\rho^{-1}\sigma_{d}^{2}. \tag{10A.24}$$

Proof of Eq. (10.81). Note that by Eq. (10A.20), the equilibrium price is,

$$p=\bar{d}+\lambda\left(N\beta\left(\hat{s}-\bar{d}\right)+Z\right).$$

Moreover, the trade of agent 1 is $x_1=\beta\left(s^1-\bar{d}\right)$. Therefore, the expected profits for each trader are,

$$\begin{aligned}E\left(\left(d-p\right)x\right)&=\beta E\left(\left(d-\bar{d}\right)\left(s^{1}-\bar{d}\right)\right)-\lambda N\beta^{2}E\left(\left(\hat{s}-\bar{d}\right)\left(s^{1}-\bar{d}\right)\right)\\&=\beta\left(\left(1-\lambda\beta N\right)\sigma_{d}^{2}-\lambda\beta\sigma_{u}^{2}\right)\\&=\beta\frac{\left(2-\rho\right)\sigma_{d}^{2}-\rho\sigma_{u}^{2}}{2+\left(N-1\right)\rho},\end{aligned}$$

where the second line follows by the expression for the average signal in Eq. (10A.21) and by rearranging terms, and the third by using the expression for $\lambda\beta$ in Eq. (10.78). Eq. (10.81) follows by the definition of ρ in (10A.24) and the expression for β in (10.80).

Proofs regarding the model with mandatory disclosure. We provide proofs based on arguments made in Mele (2017). By the projection theorem,

$$\lambda_{1}=\frac{cov(\tilde{d},y_{1})}{var\left(y_{1}\right)}=\frac{\beta_{1}\sigma_{d}^{2}}{\beta_{1}^{2}\sigma_{d}^{2}+\sigma_{\eta}^{2}+\sigma_{z}^{2}},\quad\gamma=\frac{cov(\tilde{d},x_{1})}{var\left(x_{1}\right)}=\frac{\beta_{1}\sigma_{d}^{2}}{\beta_{1}^{2}\sigma_{d}^{2}+\sigma_{\eta}^{2}}, \tag{10A.25}$$

and

$$\sigma_{d|x_{1}}^{2}=\sigma_{d}^{2}-\frac{cov^{2}(\tilde{d},x_{1})}{var\left(x_{1}\right)}=\left(1-\gamma\beta_{1}\right)\sigma_{d}^{2}. \tag{10A.26}$$

Let us conjecture that

$$\sigma_{\eta}^{2}=\ell\beta_{1}^{2}\sigma_{d}^{2}, \tag{10A.27}$$

for some constant ℓ. Then, by the second of Eqs. (10A.25), $\gamma\beta_1=\left(1+\ell\right)^{-1}$, which plugged into Eq. (10A.26), delivers

$$\sigma_{d|x_{1}}^{2}=\frac{\ell}{1+\ell}\sigma_{d}^{2}. \tag{10A.28}$$

Note, also, that $\gamma=2\lambda_2$ and $2\lambda_2=\frac{1}{\beta_2}$, such that $\gamma\beta_2=1$. Using this results, and $\gamma\beta_1=\left(1+\ell\right)^{-1}$, yield

$$\beta_{2}=\left(1+\ell\right)\beta_{1}. \tag{10A.29}$$

Next, use the expressions for λ_1 in the first of Eqs. (10A.25)) and for λ_2 (see Eq. (10.113)) into the equality, $\lambda_1=\lambda_2$,

$$\frac{\beta_{1}\sigma_{d}^{2}}{\beta_{1}^{2}\sigma_{d}^{2}+\sigma_{\eta}^{2}+\sigma_{z}^{2}}=\frac{\beta_{2}\sigma_{d|x_{1}}^{2}}{\beta_{2}^{2}\sigma_{d|x_{1}}^{2}+\sigma_{z}^{2}}. \tag{10A.30}$$

By plugging (10A.27)-(10A.28)-(10A.29) into (10A.30) leaves

$$\frac{\beta_{1}\sigma_{d}^{2}}{\left(1+\ell\right)\beta_{1}^{2}\sigma_{d}^{2}+\sigma_{z}^{2}}=\frac{\ell\beta_{1}\sigma_{d}^{2}}{\ell\left(1+\ell\right)\beta_{1}^{2}\sigma_{d}^{2}+\sigma_{z}^{2}}.$$

This equation is satisfied by $\ell=1$. Replacing this solution into (10A.27), (10A.28) and (10A.29) leads to the results stated oin the main text.

Appendix 3: Proofs regarding search models

Proof of Eq. (10.117). Consider Eqs. (10.115). Subtracting the second from the first equation and the fourth from the third leaves

$$\begin{cases} r\Delta V_l = \lambda_u (\Delta V_h - \Delta V_l) + \lambda \mu_{hn} (P - \Delta V_l) + 1 - \delta \\ r\Delta V_h = \lambda_d (\Delta V_l - \Delta V_h) + 1 - \lambda \mu_{lo} (\Delta V_h - P) \end{cases}$$

Replacing the price P in (10.116) into the previous system, and simplifying, leads to

$$\begin{cases} r\Delta V_l = (\lambda_u + \lambda \mu_{hn} q) (\Delta V_h - \Delta V_l) + 1 - \delta \\ r\Delta V_h = -(\lambda_d + \lambda \mu_{lo} (1 - q)) (\Delta V_h - \Delta V_l) + 1 \end{cases} \tag{10A.31}$$

which is solved by

$$\Delta V_h - \Delta V_l = \frac{\delta}{r + \lambda_u + \lambda_d + \lambda \mu_{lo} (1 - q) + \lambda \mu_{hn} q}.$$

Replacing the previous expression for $\Delta V_h - \Delta V_l$ into the first of Eqs. (10A.31) leads to an expression for ΔV_l. Replacing this expression for ΔV_l and the previous expression for $\Delta V_h - \Delta V_l$ into Eq. (10.116) leaves the expression for the price in (10.117).

Determination of the steady state. Evaluating Eqs. (10.114) at the steady state leaves

$$\begin{cases} 0 = -\lambda \mu_{hn} \mu_{lo} - \lambda_u \mu_{lo} + \lambda_d \mu_{ho} \\ 0 = \lambda \mu_{hn} \mu_{lo} - \lambda_u \mu_{ln} + \lambda_d \mu_{hn} \\ 0 = \lambda \mu_{hn} \mu_{lo} - \lambda_d \mu_{ho} + \lambda_u \mu_{lo} \\ 0 = -\lambda \mu_{hn} \mu_{lo} - \lambda_d \mu_{hn} + \lambda_u \mu_{ln} \end{cases} \tag{10A.32}$$

where all masses sum to one, and asset supply is equal to s,

$$1 = \mu_{lo} + \mu_{ln} + \mu_{ho} + \mu_{hn}, \quad s = \mu_{lo} + \mu_{ho}. \tag{10A.33}$$

Add the first and second (or, indifferently, the third and fourth) equations of (10A.32) and conclude that

$$\mu_{lo} + \mu_{ln} = (\mu_{ho} + \mu_{hn}) \frac{\lambda_d}{\lambda_u}. \tag{10A.34}$$

Combining the first of Eqs. (10A.33) and Eq. (10A.34) leaves

$$\mu_{lo} + \mu_{ln} = \frac{\lambda_d}{\lambda_u + \lambda_d}, \quad \mu_{ho} + \mu_{hn} = \frac{\lambda_u}{\lambda_u + \lambda_d}. \tag{10A.35}$$

Combining the second of Eqs. (10A.33) and the second of Eq. (10A.35) leaves

$$\mu_{hn} = \mu_{lo} + \frac{\lambda_u}{\lambda_u + \lambda_d} - s. \tag{10A.36}$$

Replacing the second of Eqs. (10A.33) and Eq. (10A.36) into the first equation of (10A.32) leaves the quadratic equation for μ_{lo} in the main text (see Eq. (10.118).

10.14 Appendix 4: Some foundations to pricing behavior in macroeconomics

We provide a few arguments to justify the functional form for the supply and demand equations of Section 10.2 (see Eq. (10.1) and Eq. (10.3)). We consider a variant of a model presented by Blanchard and Fischer (1989), which itself is a simplified version of Blanchard and Kiyotaki (1987). The model in this appendix differs from those above as it considers (i) a continuum of goods and (ii) taste shocks affecting each good. While the material in this appendix goes somehow beyond the scope of these lectures, it still helps maintaining the reasoning in this chapter self-contained.

Consumers-producers tastes and technology, and markets. So consider a continuum of consumers-producers, each with a utility function equal to

$$U_i \equiv C_i^g \left(\frac{M_i}{P}\right)^{1-g} - d\frac{L_i^\beta}{\beta}, \quad i \in [0,1]\,, \tag{10A.37}$$

where M_i is the nominal money holdings, P is a general price index (to be determined below; see Eq. (10A.44)), C_i is consumption of a basket of goods (to be defined in a moment), and L_i is labor supplied; the parameters satisfy $g \in (0,1)$, $d > 0$, $\beta \geq 1$. In words, each producer enjoys consuming a basket a goods but suffers a disutility while working.

We assume that the basket of goods contains CES ("Constant Elasticity of Substitution") substitutes as in Dixit and Stiglitz (1977), that is,

$$C_i = \left(\int_0^1 Z_j^{1/\theta} C_{ij}^{(\theta-1)/\theta} dj\right)^{\frac{\theta}{\theta-1}}, \tag{10A.38}$$

where C_{ij} is consumer-producer i consumption of the good j, Z_j is a taste shock for good j, assumed to be the same for all consumers-producers, and θ is the CES between any two consumption goods C_{ij} and $C_{i\ell}$. We assume that $\theta > 1$. We shall specify the properties of Z_j later.

Finally, we assume that (i) each producer specializes in the production of his own good according to the simple technology: $Y_i = L_i$, and that (ii) the market is one with monopolistic competition; this market collapses to the perfectly competitive one when θ is large.

Each consumer-producer i maximizes the utility in (10A.37) subject to the budget constraint,

$$\int_0^1 P_j C_{ij} dj + M_i = P_i Y_i + \bar{M}_i \equiv I_i, \tag{10A.39}$$

where P_i is the price of his produced good and $\bar{M}_i$ is his money endowment. It is easy to see that the solution to this problem is symmetric, in that each consumer-producer chooses the same basket *composition* (albeit acheiving different consumption quantities of this same basket, depending on the income I_i). We can, then, define the price index in (10A.37) as the minimal expenditure needed to purchase a unit of the composite good C_i in Eq. (10A.38),

$$P : PC_i = \int_0^1 P_j C_{ij} dj. \tag{10A.40}$$

We now determine demand for both the single and composite goods.

Goods demand and price index. We determine C_i first. Replacing Eq. (10A.40) into the budget constraint (10A.39) and maximizing (10A.37) subject to the resulting constraint, $PC_i + M_i = I_i$, leads to a standard result:

$$C_i = g\frac{I_i}{P}, \quad M_i = (1-g)\, I_i. \tag{10A.41}$$

Next, we identify the optimal composition of each single good, C_{ij}. It is

$$\hat{C}_{ij} = \arg\max_{C_{ij}} C_i \quad \text{s.t.} \quad \int_0^1 P_j C_{ij} dj = I_i - M_i = gI_i, \tag{10A.42}$$

where the second equality of the constraint follows by Eqs. (10A.41). Note that this program is simply maximization of (10A.37) under the constraint (10A.39). The first order conditions lead to

$$\left(\frac{Z_\ell}{Z_j}\right)^{-\frac{1}{\theta}} \left(\frac{C_{ij}}{C_{i\ell}}\right)^{1-\frac{1}{\theta}} = \frac{P_j C_{ij}}{P_\ell C_{i\ell}},$$

whence, the definition of θ as the CES between any two consumption goods. Note that to simplify notation, we are now reverting to writing $C_{ij} \equiv \hat{C}_{ij}$. Replacing the previous optimality condition into the constraint of (10A.42), and using the definition of the composite good in Eq. (10A.38), leads to

$$C_{i\ell} = Z_\ell \left(\frac{P_\ell}{P}\right)^{-\theta} C_i = gZ_\ell \left(\frac{P_\ell}{P}\right)^{-\theta} \frac{I_i}{P}, \tag{10A.43}$$

where the second equality follows by the second equality of the constraint in (10A.42).

The price index is obtained by solving for P in Eq. (10A.40) while relying on the first equality in Eq. (10A.43), leaving

$$P = \left(\int_0^1 Z_j P_j^{1-\theta} dj\right)^{\frac{1}{1-\theta}}. \tag{10A.44}$$

Finally, we determine the indirect utility of consumer-producer i, by replacing Eqs. (10A.41) into Eq. (10A.37),

$$U_i \equiv \gamma \frac{I_i}{P} - d\frac{Y_i^\beta}{\beta} = \gamma \left(\frac{P_i}{P} Y_i + \frac{\bar{M}_i}{P}\right) - d\frac{Y_i^\beta}{\beta}, \tag{10A.45}$$

where γ is a constant and where we have used the agent budget constraint and the production technology $Y_i = L_i$. Note that this expression resembles that of a maximizing firm. One issue is whether the consumer-producer has pricing power, that is, whether P_i in $P_i Y_i$ is affected by his decisions regarding Y_i. We assume it is the case below, although then this is not crucial for the interpretation of the pricing equations in Section 10.2.

Production and equilibrium. The aggregate demand for the good produced by the consumer-producer ℓ is obtained by aggregating the individual demands for this good,

$$Y_\ell \equiv \int_0^1 C_{i\ell} di = gZ_\ell \left(\frac{P_\ell}{P}\right)^{-\theta} \int_0^1 \frac{I_i}{P} di = Z_\ell \left(\frac{P_\ell}{P}\right)^{-\theta} Y, \tag{10A.46}$$

where we have used the second equality in (10A.43), and where the last equality holds by the following definition of aggregate demand, Y,

$$PY \equiv \int_0^1 \left(\int_0^1 P_j C_{ij} dj\right) di = \int_0^1 PC_i di = g \int_0^1 I_i di. \tag{10A.47}$$

Let $\bar{M}$ denote the aggregate money endowment. We can solve for Y, by noticing that

$$Y = g \int_0^1 \frac{I_i}{P} di = g \left(\int_0^1 \frac{P_i Y_i + \bar{M}_i}{P} di\right) = g \left(Y + \frac{\bar{M}}{P}\right), \tag{10A.48}$$

and the fact that, in equilibrium, $\bar{M} = \int_0^1 M_i di$, such that,

$$\int_0^1 P_i Y_i di = \int_0^1 PC_i di = PY,$$

where the first equality follows by (i) aggregating the agents budget constraint and (ii) the equilibrium condition equilibrium, $\bar{M} = \int_0^1 M_i di$; and the second by (10A.47). By Eq. (10A.48), the solution for Y is then

$$Y = \frac{g}{1-g}\frac{\bar{M}}{P},$$

which replaced into Eq. (10A.46) leaves the aggregate demand facing consumer-producer i

$$Y_i = Z_i \left(\frac{P_i}{P}\right)^{-\theta} \frac{g}{1-g}\frac{\bar{M}}{P}. \tag{10A.49}$$

In the context of monopolistic competition of this appendix, one could solve for the optimal pricing rule. The latter is obtained by replacing Eq. (10A.49) into Eq. (10A.45) and maximizing with respect to P_i, leaving

$$\frac{P_i}{P} = \left[\frac{d\theta}{\gamma(\theta-1)}\left(Z_i \frac{g}{1-g}\frac{\bar{M}}{P}\right)^{\beta-1}\right]^{\frac{1}{1+\theta(\beta-1)}}. \tag{10A.50}$$

We now proceed with the interpretation of some basic assumptions in Section 10.2 in light of the framework developed so far.

Relations with model in Section 10.2. Eq. (10A.49) provides foundations to the demand for each produced good in the model of Section 10.2 (see Eq. (10.3)). Regarding the supply equation (see Eq. (10.1)), note that assuming market power and that P is common knowledge leads to Eq. (10A.50), which could be replaced back into Eq. (10A.49) to determine the equilibrium production in each market. Alternatively, one could simplify and remove the assumption of market power. Assume, further, that P is not common knowledge, in which case producers maximize their expected utility in Eq. (10A.45),

$$\hat{Y}_i \equiv \arg\max_{Y_i} E\left(\gamma\frac{P_i Y_i}{P} - d\frac{Y_i^\beta}{\beta}\middle| P_i\right) = \left[\frac{\gamma}{d}E\left(\frac{P_i}{P}\middle| P_i\right)\right]^{\frac{1}{\beta-1}}.$$

Assume that $\ln\frac{P_i}{P}$ is normally distributed conditionally on P_i. It is easy to verify that in this case,

$$y_i^s \equiv \ln\hat{Y}_i = \kappa_0 + \kappa_1 E\left(p_i - p\middle| p_i\right), \tag{10A.51}$$

for two constant κ_0 and κ_1. Naturally, the assumption that $p_i - p$ is conditionally normally distributed needs to be verified to hold in equilibrium. However, the price index in Eq. (10A.44) reveals this to be problematic, unless a number of conditions hold true. For example, assume that the agents distribution is such that there is a mass $\mu(i)$ of agents i, and suppose that taste shocks are centered around one, or $\int Z_i d\mu(i) = 1$, and that $\theta = 1$, such that

$$p = \int Z_i p_i d\mu(i).$$

Even in this case, the price index is not exactly the average log-price due to the presence of the taste shocks, such that Eq. (10A.51) is only an approximation to Eq. (10.1).

References

Back, K. (1992): "Insider Trading in Continuous Time." *Review of Financial Studies* 5, 387-409.

Back, K., C.H. Cao and G.A. Willard (2000): "Imperfect Competition Among Informed Traders." *Journal of Finance* 55, 2117-2155.

Black, F. (1986): "Noise." *Journal of Finance* 41, 529-543.

Blanchard, O.J. and N. Kiyotaki (1987): "Monopolistic Competition and the Effects of Aggregate Demand." *American Economic Review* 77, 647-666.

Blanchard, O. and S. Fischer (1989): *Lectures on Macroeconomics*. Cambridge, MIT Press.

Bolton, P. and D.S. Scharfstein (1990): "A Theory of Predation Based on Agency Problems in Financial Contracting." *American Economic Review* 80, 93-106.

Chamberlin, E.H. (1933): *The Theory of Monopolistic Competition: A Re-orientation of the Theory of Value*. Harvard: Harvard University Press.

Cournot, A.A. (1838): *Recherches sur les Principes Mathématiques de la Théorie des Richesses*. Paris: Hachette.

DeLong, J.B., A. Shleifer, L.H. Summers and R.J. Waldman (1990): "Noise Trader Risk in Financial Markets." *Journal of Political Economy* 98, 703-738.

Diamond, D.W. (1984): "Financial Intermediation and Delegated Monitoring." *Review of Economic Studies* 51, 393-414.

Diamond, D.W. and R.E. Verrecchia (1981): "Information Aggregation in a Noisy Rational Expectations Economy." *Journal of Financial Economics* 9, 221-235.

Dixit, A.K. and J.E. Stiglitz (1977): "Monopolistic Competition and Optimum Product Diversity." *American Economic Review* 67, 297-308.

Duffie, D. (2012): *Dark Markets: Asset Pricing and Information Transmission in Over-the-Counter Markets* (Princeton Lectures in Finance). Princeton: Princeton University Press.

Duffie, D., N. Gârleanu and L.H. Pedersen (2005): "Over-the-Counter Markets." *Econometrica* 73, 1815-1847.

Duffie, D., N. Gârleanu and L.H. Pedersen (2007): "Valuation in Over-the-Counter Markets." *Review of Financial Studies* 20, 1865-1900.

Fama, E. (1970): "Efficient Capital Markets: A Review of Theory and Empirical Work." *Journal of Finance* 25, 383-417.

Foster, F.D., and S. Viswanathan (1996): "Strategic Trading When Agents Forecast the Forecasts of Others." *Journal of Finance* 51, 1437-78.

Foucault T., M. Pagano and A. Röell (2013): *Market Liquidity: Theory, Evidence and Policy*. Oxford: Oxford University Press.

Gârleanu, N., L.H. Pedersen and A.M. Poteshman (2009): "Demand-Based Option Pricing." *Review of Financial Studies* 22, 4259-4299.

Glosten, L.R. and P.R. Milgrom (1985): "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders." *Journal of Financial Economics* 14, 71-100.

Greenwood, R. and D. Vayanos (2014): "Bond Supply and Excess Bond Returns." *Review of Financial Studies* 27, 663-713.

Gromb, D. and D. Vayanos (2002): "Equilibrium and Welfare in Markets with Financially Constrained Arbitrageurs." *Journal of Financial Economics* 66, 361-407.

Grossman, S.J. (1976): "On the Efficiency of Competitive Stock Markets where Traders Have Diverse Information." *Journal of Finance* 31, 573-585.

Grossman, S.J. and J.E. Stiglitz (1980): "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70, 393-408.

Hayek, F.A. (1945): "The Use of Knowledge in Society." *American Economic Review* 35, 519-530.

Huddart. S., J.S. Hughes and C.B. Levine (2001): "Public Disclosure and Dissimulation of Insider Trades." *Econometrica* 69, 665-681.

Karatzas, I. and S.E. Shreve (1991): *Brownian Motion and Stochastic Calculus*. New York: Springer Verlag.

Kyle, A.S. (1985): "Continuous Auctions and Insider Trading." *Econometrica* 53, 1335-1355.

Kyle, A.S. (1989): "Informed Speculation with Imperfect Competition." *Review of Economic Studies* 56, 317-356.

Hellwig, M.F. (1980): "On the Aggregation of Information in Competitive Markets." *Journal of Economic Theory* 22, 477-498.

Lange, O. (1936): "On the Economic Theory of Socialism: Part I." *Review of Economic Studies* 4: 53-71.

Lange, O. (1942): "The Foundations of Welfare Economics." *Econometrica* 10: 215-228.

Lucas, R.E. (1972): "Expectations and the Neutrality of Money." *Journal of Economic Theory* 4, 103-124.

Lucas, R.E. (1973): "Some International Evidence on Output-Inflation Tradeoffs." *American Economic Review* 63, 326-334.

Lucas, R.E. (1977): "Econometric Policy Evaluation: A Critique." *Carnegie-Rochester Conference Series on Public Policy* 1, 19-46.

Lucas, R.E. (1981): *Studies in Business-Cycle Theory*. Boston, MIT Press.

Mele, A. (2017): "Insider Trading and the Tobin Tax." Unpublished Manuscript, Swiss Finance Institute at USI (Lugano).

Phelps, E.S. (1970): "Introduction." In: Phelps, E. S. (Editor): *Microeconomic Foundations of Employment and Inflation Theory*, New York: W. W. Norton.

Stoll, H.R. (1978): "The Supply of Dealer Services in Securities Markets." *Journal of Finance* 33, 1133-1151.

Part III

Asset pricing and reality

11
Options and volatility

11.1 Introduction

This is the first of four chapters that aim to illustrate how financial theory can be applied to deal with the evaluation of derivatives and related instruments. We already know that much of the theory in Part I is motivated as an attempt to rationalize the breakthrough made by Black and Scholes (1973) and Merton (1973) to price European options. How come we could even price an asset without making reference to any risk-aversion correction? The theory in Part I explains the rationale behind this and related results. We now apply the theory to explain how to price assets in markets more realistic than those originally idealized by Black, Scholes and Merton.

We face a paradox known since at least Hakansson (1979). If the Black & Scholes formula is true, we should acknowledge that markets are complete, as market completeness is the assumption needed to argue about the "redundancy" of the option and, then, the whole Black & Scholes theoretical construct. But if the option is redundant, why would we be willing to trade it in the first place? Alternatively, the option is not redundant—and options are massively traded indeed—but then the Black & Scholes formula is wrong, in that it relies on the counterfactual assumption that markets are complete. Indeed, in practice, many derivatives are traded over-the-counter, with financial intermediaries specializing in providing counterparties with payoffs. Financial intermediation is also about matching the clients' needs regarding the obtention of dedicated payoffs against a fee, on top of the "fair value" of the derivative. The fee might be justified by the specialization required to cope with the sources of market incompleteness, as well as the risks the intermediary will incur losses due to its obligation to honour the payoffs promised to clients.

This chapter analyzes a form of market incompleteness, arising when the volatility of the assets underlying these derivatives is random, and cannot be hedged through the underlying assets.

[Introduction in progress]

[Plan of the chapter]

11.2 Forwards and futures

A forward contract is an agreement to pay for the realization of a given risk at some maturity date, with the price being agreed at the inception of the contract. The realized risk can be anything: an asset price, a commodity price, or an abstract index. The agreed price is called *forward price*, and is such that it makes the value of the contract equal to zero at inception. A standard situation is when the risk is an asset price, and when the two parties agree to trade the asset at a price determined at inception, as in the examples below. In this case, we say that the party who has the obligation to buy the asset at the expiration has a long position, and the party who has the obligation to sell this asset has a short position.

Forward contracts are a means to lock-in uncertainty relating to the realization of the future risk. Consider the following examples. A chocolate producer starts his production process at time T, where he will need cocoa as an input. The output, chocolate, will be sold at time $S > T$, as depicted below.

$$0 \quad \longmapsto \quad T \text{ (cocoa)} \quad \longmapsto \quad S \text{ (chocolate)} \quad \longmapsto$$

Let p_T^{co} and p_S^{ch} be the price of cocoa and chocolate at time T and S, and F^{co} and F^{ch} be the corresponding forward prices. To insure against the vagaries of cocoa prices, the producer can go *long* a forward contract. This contract guarantees a payoff equal to $p_T^{\text{co}} - F^{\text{co}}$ at time T. This payoff, minus the unit input cost p_T^{co} to incur at time T, leaves exactly $-F^{\text{co}}$. That is, a forward contract allows the producer to freeze the sure amount F^{co} to be paid at T. Likewise, the producer may wish to *short* a forward on the price of his own chocolate; indeed, this position would allow him to receive a payoff equal to $F^{\text{ch}} - p_S^{\text{ch}}$ at S, which added to the price of chocolate sold at S, leaves the sure amount F^{ch}.

11.2.1 Forwards: definition and pricing in frictionless markets

If markets are frictionless, and the underlying asset is traded (a stock, say), forward contracts can be synthesized as follows. Let $P(t,T)$ be the price of a bond expiring at time T and S_t the price of a stock. Assuming the short-term rate r is constant, we have $P(t,T) = e^{-r(T-t)}$, where r denotes the short-term rate, which is the same for borrowing and lending. So at time t, borrow $P(t,T)F^*$ and buy the stock, choosing $F^* : P(t,T)F^* - S_t = 0$. The value of this portfolio at time T is $S_T - F^*$. But the portfolio is worthless at time t, so this trade is the same as a forward. Therefore, we have $F^* = F_t(T)$, where:

$$F_t(T) \equiv S_t e^{r(T-t)}. \tag{11.1}$$

Therefore, forwards are insensitive to volatility in this introductory example. They can actually be volatility-sensitive under circumstances clarified below (see the discussion after Eq. (11.2)) and in the following sections.

The pricing of a forward on a stock can be extended to cases where the underlying is not traded. Suppose that a risk materializes at time T, say the level of average temperature over a certain period preceding T, and over a pre-specified geographical area, or say the realized volatility experienced by the S&P 500 Index over the month preceding T. Denote this risk with $\tilde{X}_T$. The payoff of a forward contract from the perspective of its buyer is given by $\tilde{X}_T - F_t(T)$, such that by risk-neutral evaluation,

$$F_t(T) = \mathbb{E}_t(\tilde{X}_T), \tag{11.2}$$

where $\mathbb{E}_t$ denotes the risk-neutral expectation. Naturally, Eq. (11.2) collapses to Eq. (11.1), should $\tilde{X}_T$ be the price of a traded stock.

In general, though, Eq. (11.2) reveals that is unlikely we could come up with a preference-free evaluation of a non-traded risk, as expected from the theory developed in Part I. In these cases, the volatility of $\tilde{X}_T$ may well affect the forward price $F_t(T)$ due to risk-premiums.

There are exceptions in which the pricing in Eq. (11.2) is model-free even if $\tilde{X}_T$ is not traded. An important relatively recent advance is to have shown how to proceed with the model-free evaluation of non-traded risks, provided a sufficiently high number of additional derivatives are written on this risk. The CBOE-VIX index of expected volatility does rely on this idea, as explained in Section 11.8.

11.2.2 Forwards as a means to borrow money

Forward contracts can be used to borrow money. We can do the following: (i) go long a forward, which at time T, delivers the payoff $-F+S_T$; (ii) short-sell the underlying asset, which at time T, will give rise to a payoff of $-S_T$. So, (i) and (ii) are such that now, we access to S_t dollars, due to (ii), and at time T, we pay $-F$, i.e. the sum of the two payoffs resulting from (i) and (ii). By Eq. (11.1), this is tantamount to borrowing money at the interest rate r.

11.2.3 Marking to market

Consider a derivative we go long at time $t=0$, when it is worthless. As time unfolds, its value will change, which calls for marking to market it. Suppose the derivative pays off $\psi(S_T)-K_0$ at time T, where S_T is the price of some asset as of time T, and K_0 is set so as to make the derivative worthless at time zero. Assuming that interest rates are constant, we have that $K_0 : e^{-rT}\mathbb{E}_0[\psi(S_T)-K_0]=0$, where $\mathbb{E}_0$ is the expectation at time $t=0$, taken under the risk-neutral probability. That is, $K_0=\mathbb{E}_0[\psi(S_T)]$. The market value of the derivative at time t, say MtM_t, is simply the present value of the expected payoff at T, under the risk-neutral probability, $e^{-r(T-t)}\mathbb{E}_t[\psi(S_T)-K_0]$, or

$$\mathrm{MtM}_t = e^{-r(T-t)}(K_t-K_0). \tag{11.3}$$

For more elaborated payoffs, such as those depending on the realizations of the underlying risks over the life of the contract, marking to market updates may be more intricate than that in Eq. (11.3) (see, for example, the variance contract case in Section 11.7.3).

11.2.4 Futures

Forward contracts are typically OTC, and not standardized, and might not be traded after their inception. Futures are, instead, standardized. The cost of entering into a future contract is zero, as with a forward. However, on of the central features of future contracts is *marking-to-market*, which forces their value to be zero at any time we wish to enter into them after their inception. In contrast, note that we should generally have to pay (or be payed) to enter into a forward contract after its inception. Indeed, the payoff at T pertaining to a forward that we enter at time t is $S_T-F_t(T)$, such that, and similarly as for Eq. (11.3), the value of the forward at τ (and originated at t) is as in the following mark-to-market update,

$$V_\tau^F(T) \equiv e^{-r(T-\tau)}\mathbb{E}_\tau(S_T-F_t(T)) = e^{-r(T-\tau)}(F_\tau(T)-F_t(T)).$$

By construction, forwards are not "standardized," in that their value (the cost of entering into them) depends on when we enter into them! Moreover, in OTC markets, it is often the case that the appropriate marking notion is that of a mark-to-model, rather than mark-to-market.

Futures work differently, as the cost of entering into them is always zero, as mentioned. Precisely, let $\mathcal{F}_\tau(T)$ be the price process of the future at time τ. Intuitively, while we hold a future position, we only pay or receive differences up to the maturity T, $\mathcal{F}_{\tau_i}(T) - \mathcal{F}_{\tau_{i-1}}(T)$ over pre-specified periods, $\tau_i - \tau_{i-1}$.[1] Finally, we could "close" any position at any time before T by just initiating an opposite position.

For reasons developed below, let us assume that the short-term rate r_τ is random. The two defining properties of $\mathcal{F}_\tau(T)$ are that (i) at maturity, the futures "settle" at S_T, i.e. $\mathcal{F}_T(T) = S_T$ (a boundary condition that is part of the futures security design), and (ii) the gains generated by $d\mathcal{F}_\tau(T)$ are continuously credited from/debited to the future holder's account, such that in absence of arbitrage,

$$0 = \mathbb{E}_t\left(\int_t^T e^{-\int_t^\tau r_u du} d\mathcal{F}_\tau(T)\right).$$

The previous equality actually holds for any investment horizon $\ell \leq T$. Therefore, we have, heuristically, that $X_\ell \equiv \int_t^\ell e^{-\int_t^\tau r_u du} d\mathcal{F}_\tau(T)$ is a martingale under the risk-neutral probability, or $0 = \mathbb{E}_\ell(dX_\ell) = e^{-\int_t^\ell r_u du}\mathbb{E}_\ell(d\mathcal{F}_\ell(T))$, implying that $\mathcal{F}_\tau(T)$ is a martingale too. Therefore,

$$\mathcal{F}_t(T) = \mathbb{E}_t(S_T).$$

In contrast, it is easy to see that forward prices satisfy,

$$F_t(T) = \frac{1}{P_t(T)}\mathbb{E}_t\left(e^{-\int_t^T r_u du} S_T\right),$$

where $P_t(T)$ is the price of a zero coupon bond. In other words, the future price is a martingale under the risk-neutral probability, while the forward is not. In Chapter 13, we show that forward prices are martingales under another probability, referred to as the forward probability (see also Chapter 5). Naturally, both futures and forward prices are martingales under the risk-neutral probability, once interest rates are constant or deterministic, although this case is obviously not relevant whilst dealing with the evaluation of fixed income securities.

11.2.5 Backwardation and Contango

Let us assume interest rates are constant, such that forwards and futures are valued the same. A natural question arises: do markets price forwards or futures at a "discount" or "premium"? We have two notions of dicounts: one, regarding expectations of future prices; and a second, more operational, regarding current prices.

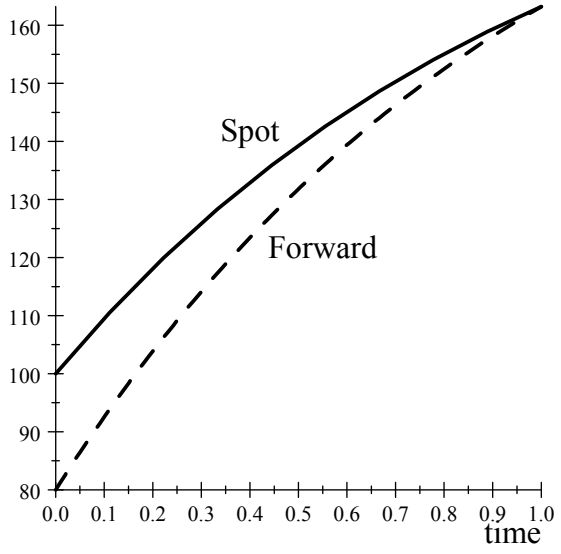
As for the first notion, we say markets are in *Backwardation* if the forward price is lower than the expected spot price, $F_t(T) = S_t e^{r(T-t)} \leq E_t(S_T)$, and in *Contango* otherwise, i.e. $F_t(T) \geq E_t(S_T)$, where E_t denotes the expectation taken under the physical probability given the infotmation at t. Clearly, markets are in Backwardation if prices follow Geometric Brownian motions, as $E_t(S_T) = S_t e^{\mu(T-t)}$, where $\mu > r$ is the drift coefficient under the physical probability. Keynes (1930) and Hicks (1939) would refer to markets in Backwardation as being

[1] In addition to these margins, one is also typically due to provide an initial margin, which aims to mitigate concerns regarding the solvability of the trading parties.

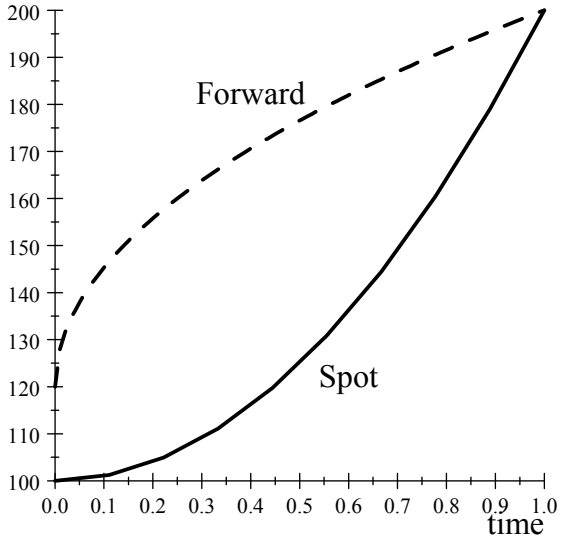
the standard situation, one of *Normal Backwardation*. According to the Keynes-Hicks Normal Backwardation hypothesis, commodity producers are more incentivized to hedge their risk than consumers, which leads forward prices to be cheaper than their current expectations. In terms of the introductory example in this section, risk-averse chocolate producers would like to short forwards, and their risk-aversion would make them likely to accept low forward prices for the chocolate they produce—lower than the chocolate price at which they expect to sell.

A second notion relies of discounts relies on spot prices, rather than the difficult to estimate expected prices. The criterion Keynes and Hicks would actually rely on links to the forward price being lower than the current spot price, not the expected spot. Accordingly, markets are in Backwardation if $F_t(T) \leq S_t$, and in Contango otherwise. It is easy to see that under this criterion, markets are now typically in Contango, as $F_t(T) = S_t e^{r(T-t)} > S_t$, as predicted by Eq. (11.1)!

To restore Backwardation, we need a mechanism by which spot prices are "rich." One possibility is to assume that the no-arbitrage arguments underlying Eq. (11.1) break down. For example, suppose that $F_t(T) < S_t e^{r(T-t)}$. A standard argument suggests to go long a forward contract, short the asset, and invest the short-sale proceeds into a money market account. Come time T, the money market account would deliver $S_t e^{r(T-t)}$, and by honouring the forward at T, which costs $F_t(T)$, we would buy the asset and, finally, close the short-sale position. The net payoff at T would be equal to $S_t e^{r(T-t)} - F_t(T) > 0$. An informal argument at this juncture would be that in the presence of a selling pressure aiming to exploit the arbitrage, the spot price would go down, thereby restoring Eq. (11.1) and, then, Contango. But in the presence of frictions such as short-sale constraints, the downward price pressure would not arise in the first place, leading to Backwardation.



Backwardation



Contango

The presence of a *convenience yield* is an alternative assumption to short-sale constraints, which we can make to restore Backwardation. A convenience yield represents the flow of services accruing to the owner of the asset (not the owner of the forward). In the presence of a constant

convenience yield, denoted with ψ, the forward price satisfies:[2]

$$F_t(T) = S_t e^{(r-\psi)(T-t)}, \tag{11.4}$$

such that we have Backwardation again, for ψ large enough.[3] Note, however, that Contango would be a normal situation in storable commodities markets characterized by costs of "carry" such as warehousing fees or foregone interests—a cost of carry would imply a negative ψ in Eq. (11.4). Thus, and according to the sign and magnitude of ψ, we can either have Backwardation or Contango.

Some commodity markets can actually switch from being in Backwardation to Contango, and in Backwardation again, through cycles. To account for these cycles, we may wish to consider a model with a stochastic convenience yield; an instance of such a model is one where under the risk-neutral probability, the spot price is solution to

$$\begin{cases} \dfrac{dS_t}{S_t} = (r - \psi_t)\, dt + \sigma_S d\hat{W}_{1t} \\ d\psi_t = \kappa(\alpha - \psi_t - \lambda\sigma_\psi)\, dt + \sigma_\psi d\hat{W}_{2t} \end{cases} \tag{11.5}$$

where $\hat{W}_{it}$ are standard Brownian motions under the risk-neutral probability, κ, α and σ_ψ are constant parameters, and λ is a risk-premium parameter arising because the stochastic convenience yield ψ_t is not tradable and, hence, $e^{-rt}\psi_t$ is not a martingale under the risk-neutral probability.

The expression for the forward price would collapse to that in Eq. (11.4), once we assume the convenience yield is constant. In the general case,

$$F_t(T) = \mathbb{E}_t(S_T) = S_t e^{r(T-t)} \mathbb{E}_t\left(e^{-\int_t^T \psi_u du}\right) \equiv S_t e^{r(T-t)} \varphi(\psi_t, T-t). \tag{11.6}$$

That is, a stochastic convenience yield might drive the correlation between the forward and the spot to values lower than one.

The expectation in Eq. (11.6), $\varphi(\psi, T-t)$, is actually known in closed-form. It is the same as the price of a zero coupon bond predicted by a model where the short-term rate is ψ_t, and is solution to the second equation in (11.5). This model, developed by Vasicek (1977), is discussed in Chapter 13. Naturally, the important implication of this model is that according to the specific values taken by ψ_t, markets can either be in Backwardation or in Contango—both in terms of the criterion comparing the forward price to the current spot and the expected spot.

Finally, the nature of the relation between the fictitious "bond price" $\varphi(\psi, T-t)$ and the volatility parameter σ_ψ is indeterminate. As discussed in Chapter 13, there are two effects that explain how the price of any zero coupon bond links to the volatility of the short-term rate, a "convexity" effect and a "risk-premium" effect. The former acts so as to lead $\varphi(\psi, T-t)$ to be increasing in σ_ψ, and the latter acts in the same direction if $\lambda > 0$, and in the opposite

[2]To prove Eq. (11.4), we generalize the reasoning leading to Eq. (11.1), as follows. Suppose to the contrary that $F_T(t) < S_t e^{(r-\psi)(T-t)}$. Then we go long a forward contract, short $\theta_t \equiv e^{-\psi(T-t)}$ shares at t, and invest the proceeds in a money market account. The short-sale at t implies that we need to pay out the convenience yield ψS_τ at any time $\tau \in (t,T)$, which we finance by shorting additional ψ shares per share already held, such that over any infinitesimal amount of time, we short $d\theta_\tau = \psi\theta_\tau d\tau$ additional shares. In total, then, at time T, we would have shorted just one share, as $\theta_T = \theta_t e^{\psi(T-t)} = 1$, which we buy back by honouring the forward, which costs $F_T(t)$, as usual. Finally, at time T, the money market account would yield $S_t e^{(r-\psi)(T-t)}$. Therefore, the net payoff at T would be equal to $S_t e^{(r-\psi)(T-t)} - F_T(t) > 0$, an arbitrage.

[3]To derive Eq. (11.4), note that by no-arbitrage, $\mathbb{E}_t\left(\frac{dS}{S}\right) + \delta dt = r dt$, where δ is the constant dividend yield. A convenience yield is simply $\psi \equiv \delta$. A storage cost is a negative convenience yield.

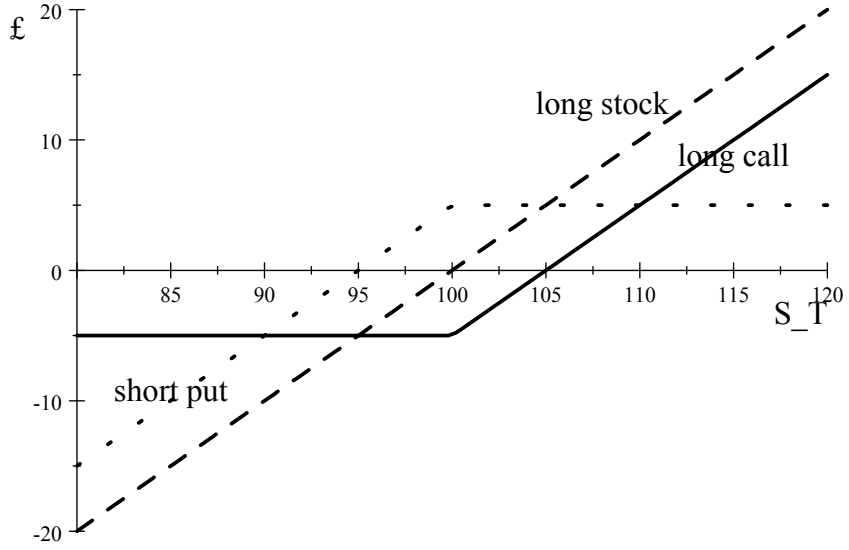
direction if $\lambda < 0$. For short-maturities, the risk-premium effect dominates over convexity, such that $\varphi(\psi, T-t)$ and, hence, the forward price, is decreasing in the volatility parameter σ_ψ, provided $\lambda < 0$. However, if $\lambda > 0$, the forward price, is increasing in σ_ψ, for any maturity.

11.3 Optionality and no-arb bounds

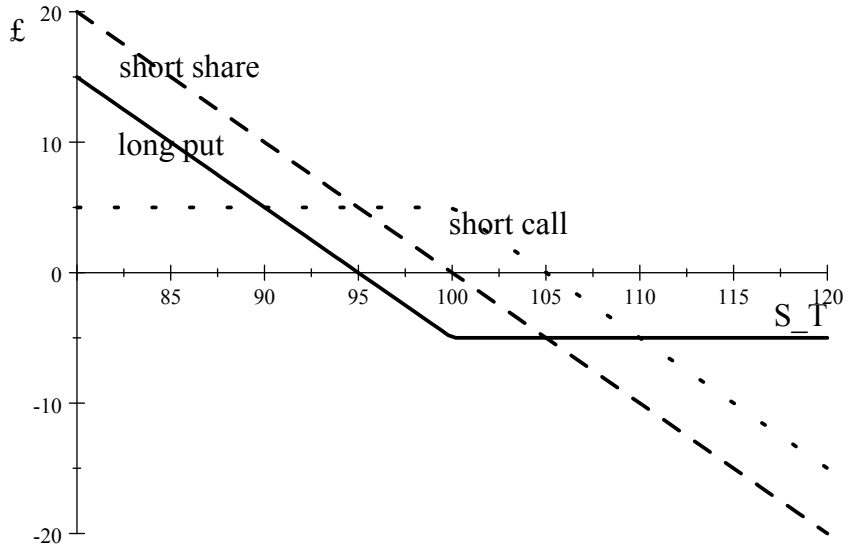
We analyze the well-known fundamental optionalities displayed by options: the right, but not the obligation, to buy or sell a given asset at a given price (the strike, or exercise price) at some future date. While Chapter 5 illustrates foundational issues regarding the evaluation of these contracts, this chapter describes additional properties of these contracts that help explain their use in the market practice, including those at the heart of the recent waves of financial innovation and underlying volatility trading. This is the first section that illustrates a few fundamental properties of European options, with a few examples.

11.3.1 Model-free properties

Let C and p the prices of the call and the put option, with S the price of the asset underlying these contracts, and with K and T the exercise prices and the expiration date. Let t be the current time. As we know, at the expiration, $C_T = (S_T - K)^+$ and $p_T = (K - S_T)^+$. Figure 11.1 depicts the net profits generated by holding or short-selling one asset and one option written on the asset. We take the short-term rate $r = 0$ to simplify the presentation. The exposure to losses generated by going long an asset drops by trading the appropriate option, i.e by a long position in a call, as illustrated by the top panel, at least provided $C_t < S_t$, which is indeed a no-arbitrage condition, as shown below. It is this insurance feature that makes the option economically valuable. Likewise, the exposure to losses relating to a short position in the asset is mitigated by trading the appropriate option, i.e by a long position in the put option, as illustrated by the bottom panel, provided again that $p_t < S_t$, another no-arbitrage condition. Once again, this feature makes the option economically valuable.



Bullish view



Bearish view

FIGURE 11.1. Top panel: The solid line depicts the P&L of an at-the-money European call option when the interest rate is zero, $(S_T - K)^+ - c_t$, where S_T is the stock price at expiration, $K = 100$ is the strike price, and $c_t = 5$ is the price of a call. The dashed line is the P&L from holding the stock, $S_T - S_t$, with $S_t = 100$. The dotted line is the payoff from the sale of a put option $p_t - (K - S_T)^+$, where by the put-call parity, $p_t = c_t - S_t + K = 5$. The bottom panel depicts P&Ls from going long a put, $(K - S_T)^+ - p_t$ (the solid line), shorting the stock, $S_T - S_t$ (the dashed line), and shorting a call $c_t - (S_T - K)^+$ (the dotted line).

The prices of the call and put options are related by the *put-call parity*. Let $P(t, T)$ be the time t price of a zero maturing at time $T > t$. Then, the prices of a put and a call option with

the same exercise price K and the same expiration date T satisfy,

$$p_t = C_t - S_t + KP(t,T). \tag{11.7}$$

To show Eq. (11.7), consider two portfolios: (a) long one call, short one underlying asset, and invest $KP(t,T)$; (b) long one put. The table below gives the value of the two portfolios at time t and at time T.

		Value at T	
	Value at t	$S_T \leq K$	$S_T > K$
(a)	$C_t - S_t + KP(t,T)$	$-S_T + K$	$S_T - K - S_T + K$
(b)	p_t	$K - S_T$	0

The two portfolios have the same value in each state of nature at time T. Therefore, their values at time t must be identical to rule out arbitrage. Alternatively, Eq. (11.7) follows by taking conditional expectations of the identity: $e^{-r(T-t)}(S_T - K)^+ \equiv e^{-r(T-t)}(K - S_T)^+ + e^{-r(T-t)}(S_T - K)$.

By the put-call parity, the properties of European puts are easily found to follow from those of calls. Therefore, we only focus on calls, whenever possible. The price of a European call option satisfies the following bounds:

$$\max\{0, S_t - KP(t,T)\} \leq C(S_t; K; T-t) \leq S_t. \tag{11.8}$$

Indeed, consider two portfolios: (a) long one call; (b) long the asset underlying the call and issue debt for an amount equal to $KP(t,T)$. The table below gives the value of the two portfolios at time t and at time T.

		Value at T	
	Value at t	$S_T \leq K$	$S_T > K$
(a)	C_t	0	$S_T - K$
(b)	$S_t - KP(t,T)$	$S_T - K$	$S_T - K$

The value of portfolio (a) dominates that of portfolio (b) at T, and the same must be true at time t. Moreover, the price C is positive because the payoff of the option is positive. Therefore, the first inequality in (11.8) is true. As for the second inequality, suppose the contrary, i.e. $C_t > S_t$. Then, at time t, we could sell one call and buy the underlying asset, thus making a profit equal to $C_t - S_t$. Come time T, the option will be exercized if $S_T > K$, in which case we shall sell the underlying assets and obtain K. If $S_T < K$, the option will not be exercized, and we will still hold the asset or sell it and make a profit equal to S_T. Eq. (11.8) implies the following asymptotic behavior of the call price: (i) $\lim_{S\to 0} C(S; K; T-t) \to 0$, (ii) $\lim_{K\to 0} C(S; K; T-t) \to S$, and (iii) $\lim_{T\to\infty} C(S; K; T-t) \to S$.

The top panel of Figure 11.2 illustrates the basic arbitrage bounds in (11.8), as well as the limiting behavior of the price for S_t small and for S_t large. First, the price C_t must be in the region within the AA and BB lines. Moreover, C_t is small when S_t is small, and large when S_t is large. However, C_t cannot lie outside the region within the AA line and BB lines, which implies that C gets large, by "sliding up" on the BB line.

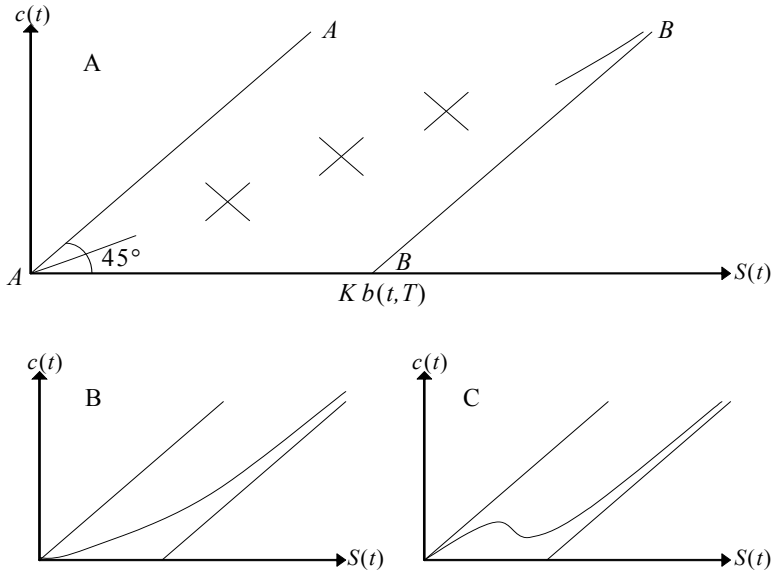


FIGURE 11.2.

How does the option price behave in the region within the AA and BB lines? We cannot tell. We may simply say that given the boundary behavior of C_t, if C_t is convex in S_t, it is also increasing in S_t. Convexity of C_t is a reasonable property, which holds for basic diffusive models, as originally noted by Bergman, Grundy and Wiener (1996). In this case, C_t would behave as in the left-hand side of the bottom panel of Figure 11.2. This case seems to be relevant, empirically, and consistent with the predictions of the celebrated Black and Scholes (1973) formula, and some of its extensions. However, it is not a general property of option prices. Bergman, Grundy and Wiener (1996) provide several counter-examples where C_t can be decreasing over some range of S_t, arising in models with jumps, or with stochastic volatility. Theoretically, we cannot rule out that the option price behaves as in the right-hand side of the bottom panel of Figure 11.2, as further developed in Section 11.5 [in progress].

The economic meaning of convexity is that the option is unlikely to be exercized when S_t is small. Therefore, changes in S_t have little effect on C_t. However, the option is likely to be exercized when S_t is large. A percentage increase in S is then to be followed by an even higher percentage increase in C. In other terms, the elasticity of the option price with respect to the asset price is larger than one, $\epsilon \equiv \frac{dC}{dS} \cdot \frac{S}{C} > 1$.[4] Therefore, option returns likely are more volatile than those on the underlying asset.

How does time-to-maturity affect the call price? Calls are known to be "wasting assets," meaning that their value decreases over time, as illustrated by an hypothetical example in Figure 11.3, which plots the option price function relating to three maturity dates, $T_1 > T_2 > T_3$, as predicted by the Black & Scholes model.

[4]For any increasing and convex function, which is zero at the origin, the tangent is higher than the secant.

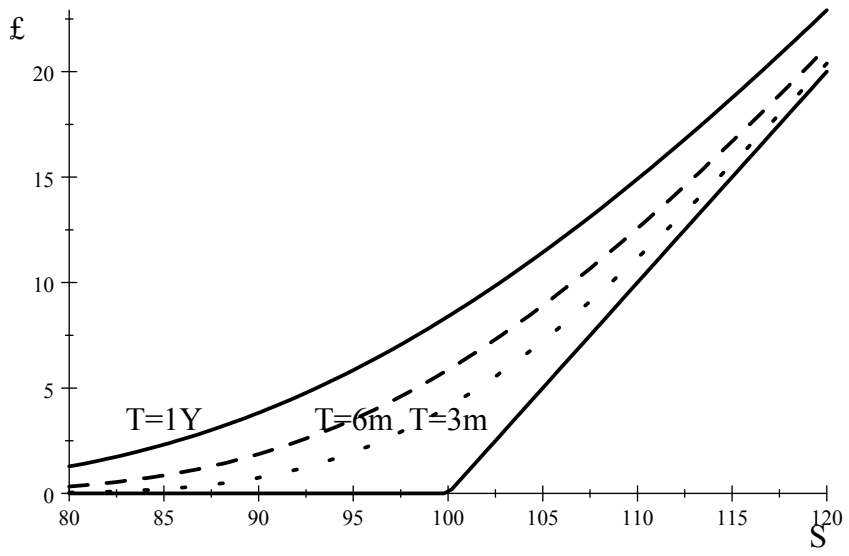


FIGURE 11.3. The value of a call option struckable at $K = 100$, as time to maturity shrinks, as predicted by the Black & Scholes model, with volatility parameter equal to 20% and short-term rate $r = 1\%$. The leftmost solid line is the price corresponding to time to maturity $T =$ one year, the dashed line is the price corresponding to time to maturity $T =$ six months, and the dotted line is the price corresponding to time to maturity $T =$ three months. The rightmost solid line is the no-arbitrage bound $(S_T - K)^+$.

11.3.2 Hedging

Financial intermediaries such as investment banks sell options that they want to hedge against, to avoid the exposure to losses illustrated in Figure 11.1. Hedging is important when the only objective is to receive fees from the sale of derivatives. The portfolio that "mimics" the option price must display the properties discussed in Section 11.3.1. For example, we need to ensure that it behaves as the call price behaves in left-hand side of the bottom panel of Figure 11.2, which is the most relevant, empirically.

We require this portfolio to exhibit a number of properties: (i) its value, V, should be increasing in S, which is ensured by including the asset underlying the option into the portfolio; (ii) the sensitivity of V with respect to S must be positive and bounded by one, $0 < \frac{dV}{dS} < 1$, which we can make, once the number of underlying assets is less than one; (iii) the elasticity of V with respect to S must be greater than one, $\frac{dV}{dS}\frac{S}{V} > 1$, a condition that could be met, by issuing debt. Mathematically, the value of the replicating portfolio should be $V = \theta S - D$, where θ denotes the number of the underlying assets, with $\theta \in (0, 1)$, and D is debt. In principle, this portfolio might lead these three properties to be satisfied.

In fact, hedging is dynamic in nature, because option prices obviously change over time. Therefore, we expect θ to be a function of the underlying asset price, S, and time to expiration of the option, in accordance with theory set forth in Part I (see Chapter 5). The portfolio needs to satisfy additional properties: (iv) the number of the underlying assets must increase with S, and the value of the portfolio should be virtually insensitive to changes in S when S is low, and "slide up" through the BB line in Figure 11.3 when S is large. These conditions are met if θ increases with S, with $\lim_{S\to 0} \theta(S) \to 0$ and $\lim_{S\to\infty} \theta(S) \to 1$. Finally, the portfolio needs to be self-financed, as the long position in the option does not entail additional inflows or outflows,

until time to expiration—any additional purchases of the underlying asset has to be financed by issuing new debt, and any additional sale of the underlying asset must be used to shrink existing. Section 11.4 aims to derive this hedging strategies within the Black & Scholes market.

11.3.3 A case study: accumulators, decumulators

Options can be used to build up dedicated structured products, such as those relying on baskets of options. Consider, for example, an "accumulator." Anecdotal evidence suggests that at times, accumulators might be quite popular amongst private bankers in Hong Kong. An accumulator is a portfolio which is long one call and short two or more puts. (A "decumulator" is short one accumulator.) The strike of the call is higher than the strike of the put. The rationale behind going long the call is to ensure profits are made once the market is up. Instead, puts are sold to finance the long position in the call, so as to make the value of the accumulator equal to zero at its inception. Consider, for example, Figure 11.4, where the strike of the call is $K_C = 100$. Note that since the value of the accumulator is zero at inception, the picture actually depicts net profits.[5]

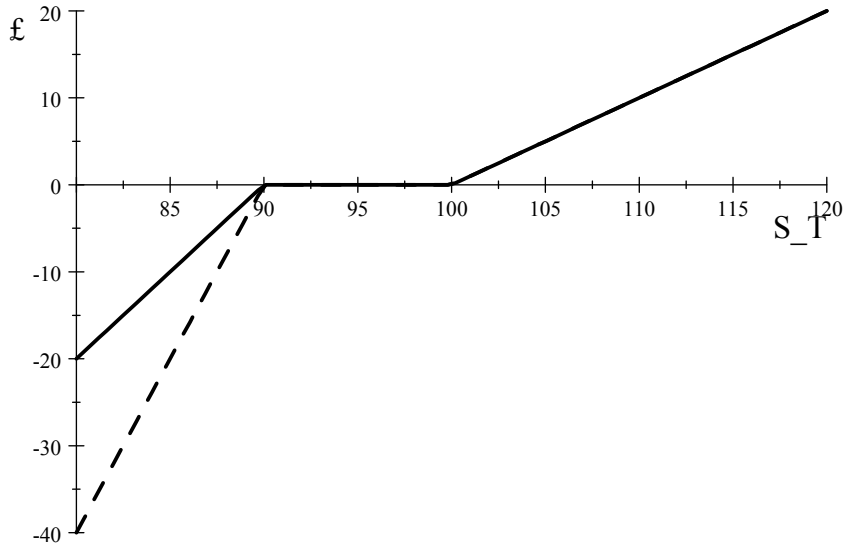


FIGURE 11.4. The solid line depicts the payoff guaranteed by an "accumulator," a structured product that is long one call option with strike price $K_C = 100$, and short n puts with strike price $K_p = 90$, and $n = 2$ (solid line), and $n = 4$ (dashed line).

If the current market level is $S = 102$, profits are likely to be made, at least provided the market does not fall below the strike at which the put is struck, which in this example is $K_p = 90$. However, accumulators are quite risky—the losses they might lead to during market downturns can be quite severe cmpared to possible gains in good times.

The size of the losses obviously depends on the number of puts to sell, n, and their strike, K_p. As the previous picture reveals, losses widen as we increase the number of puts we go short. Therefore, we can decrease the probability of experiencing any losses, by just fixing $K_p = 90$, and decreasing n, although this might entail less resources left over to go long a call.

[5]Therefore, underlying Figure 10.4 is the assumption that the price of the put is half that of the call. Below, we discuss this assumption, which we now make for illustrative purposes only, and which we shall remove in Section 10.5.3.

A possibility might then be to go long a less expensive call, by adding a knock-out feature into the call contract (one that says that the option becomes worthless once the market reaches a certain level such as, say, 105 within the investment horizon), or by purchasing a call with a lower strike price. Alternatively, a product can be designed that carries more risk, but also more upside, by choosing an appropriate strike for the puts. Obviously puts become more valuable as we increase the strike price. Therefore, we could increase K_p, from 90 to 95 say, whilst keeping n constant. While selling put options with higher strikes increases the probability to have losses, it also allows us to purchase more expensive calls, those with lower strikes, which leads to a higher probability to achieve positive returns.

Naturally, the previous reasoning hinges upom the assumption that the accumulator is self-financed at inception. This condition defines the type of options we can afford. Some call options can be too expensive and might require a large exposure relating to the short position in puts. For example, some calculations show that in a market with stochastic volatility such as that of Heston (1993b), we may need to sale short approximately six puts when the current index level is $S = 102$ and $K_p = 95$. Section 11.5.3 develops a case study where these risks are quantified under a variety of alternative assumptions about strikes and market volatility.

11.4 Evaluation

This section provides pricing formulae and discusses hedging in the special case of the Black and Scholes (1973) and Merton (1973) markets. It also examines issues relating to how hedging might possible spillover to the volatility of the underlying asset, thereby dealing with instances of "feedback effects," whereby the presence of derivatives (and trading activities on them) might affect the dynamics of the very same underlying—a phenomenon referred to as *endogenous risk* in Part II (see Chapter 9).

The next section provides a general evaluation formula to price futures and options, which goes beyond the assumptions underlying Black & Scholes and Merton market. Section 11.4.2 provides a derivation of Black-Scholes formula relying on this general framework, and also relying on a replication argument. Section 11.4.3 [...]

[In progress]

11.4.1 A pricing formula

Forward and options obviously differ, as they rely on the two distinct notions of obligation and optionality. However, we can encompass their payoffs into a single one, the value of which will be used a few times in this chapter. Consider, then, again, a contract similar to that in Section 11.2.1, where at time T, the payoff is given by $S_T - K$, for some constant value of K. We know that the current value of this payoff is:

$$e^{-r(T-t)}\mathbb{E}_t\left(S_T - K\right) = S_t - e^{-r(T-t)}K, \tag{11.9}$$

which is zero, for $K = F^*$. We want to further analyze a situation where this current value is not necessarily zero, and show that the previous expression is a special case of a quite important pricing formula. Consider a payoff at time T, equal to $S_T - K$, provided the stock price at T is at least as large as some positive constant $\ell \geq 0$,

$$\left(S_T - K\right)\mathbb{I}_{S_T \geq \ell}.$$

For $\ell = 0$, this payoff is just that of a forward, and for $\ell = K$, the payoff is that of a European call. To price this payoff, we proceed as follows:

$$\begin{aligned} e^{-r(T-t)}\mathbb{E}_t\left[(S_T - K)\,\mathbb{I}_{S_T \geq \ell}\right] &= S_t\mathbb{E}_t\left(\eta_t(T)\,\mathbb{I}_{S_T \geq \ell}\right) - e^{-r(T-t)}K\mathbb{E}_t\left(\mathbb{I}_{S_T \geq \ell}\right) \\ &= S_t\hat{\mathbb{E}}_t\left(\mathbb{I}_{S_T \geq \ell}\right) - e^{-r(T-t)}K\mathbb{E}_t\left(\mathbb{I}_{S_T \geq \ell}\right) \\ &= S_t \cdot \hat{Q}_t\left(S_T \geq \ell\right) - e^{-r(T-t)}K \cdot Q\left(S_T \geq \ell\right), \end{aligned} \tag{11.10}$$

where $\eta_t(T) \equiv \frac{e^{-r(T-t)}S_T}{S_t}$, Q_t is the risk-neutral probability given the information at time t, $\hat{Q}_t$ is a new probability, with Radon-Nikodym derivative given by

$$\frac{d\hat{Q}_t}{dQ_t} = \eta_t(T), \tag{11.11}$$

and, finally, $\mathbb{E}_t$ denotes the expectation under Q_t, and $\hat{\mathbb{E}}_t$ the expectation under $\hat{Q}_t$. Naturally, Eq. (11.10) collapses to Eq. (11.9), once $\ell = 0$, and to the celebrated Black and Scholes (1973) formula, once we take S_t to be a geometric Brownian Motion, and $\ell = K$, as explained in the following section. It is a general formula, and a quite useful one whilst dealing with difficult models such as those where the volatility of the underlying asset return is not constant, as illustrated in Section 11.5.

11.4.2 Black & Scholes

We assume that the stock price is a Geometric Brownian motion with parameters (μ, σ),

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t,$$

where W_t is a Brownian motion under the physical probability.

11.4.2.1 By replication: I

Replicating the value of an asset while trading in other assets is a theme framed in many junctures of Part I. For sake of completeness, we develop the arguments, which ultimately lead to the Black & Scholes formula.

We invest θ units in the asset underlying a call option, and θ_M in the money market account, and the value of this portfolio is $V = \theta S + \theta_M M$ with obvious notation. Once self-financed, the value of this portfolio satisfies

$$dV = \theta dS + \left(\theta_M \frac{dM}{dt}\right) dt, \tag{11.12}$$

whereas, by Itô's lemma, the call option price (C, say), satisfies

$$dC = \frac{\partial C}{\partial S}dS + \left(\frac{\partial C}{\partial t} + \frac{1}{2}\frac{\partial^2 C}{\partial S^2}\sigma^2 S^2\right) dt. \tag{11.13}$$

We conjecture that the value of this portfolio does exactly replicate the option value, a conjecture verified in the subsection below. By comparing Eq. (11.12) and Eq. (11.13) leaves

(i) the matching condition for dS, $\theta = \frac{\partial C}{\partial S}$, and (ii) the drift matching condition,

$$\begin{aligned}\frac{\partial C}{\partial t} + \frac{1}{2}\frac{\partial^2 C}{\partial S^2}\sigma^2 S^2 &= \theta_M \frac{dM}{dt} \\ &= r\theta_M M \\ &= r\left(V - \theta S\right) \\ &= r\left(C - \frac{\partial C}{\partial S}S\right),\end{aligned}$$

where the second equality follows by $\frac{dM}{M} = rdt$, the third by the expression for the value of the replicating portfolio, and the fourth, by the equality $C = V$ (arising due to the conjecture that the portfolio value is replicating the option's) and the expression for θ in (i). Rearranging terms leaves,

$$\frac{\partial C}{\partial t} + \frac{\partial C}{\partial S}rS + \frac{1}{2}\frac{\partial^2 C}{\partial S^2}\sigma^2 S^2 = rC, \tag{11.14}$$

subject to the boundary condition, $C\left(s, 0; K, \sigma\right) = \left(s - K\right)^+$.

The solution to Eq. (11.14) is the celebrated Black and Scholes (1973) formula,

$$C_{\text{BS}}\left(S, T-t; K, \sigma\right) = S\Phi(d) - Ke^{-r(T-t)}\Phi(d - \sigma\sqrt{T-t}), \quad d = \frac{\ln(\frac{S}{K}) + (r + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}, \tag{11.15}$$

where Φ denotes the cumulative standard normal distribution.

Finally, Appendix 1 derives Eq. (11.14) hinging upon the original arguments developed by Black and Scholes (1973), and Merton (1973), where one considers a portfolio comprising the option, the underlying asset, and a portfolio strategy aiming to make the portfolio locally riskless.

11.4.2.2 By replication: II

Eq. (11.15) holds even without requiring that a market for the option exists during the option's life, or that the pricing function $C\left(t, S\right)$ is differentiable. That the option price is differentiable is a result, not an assumption. Let us define the function $C\left(t, S\right)$ that solves Eq. (11.14), with boundary condition $C\left(T, S\right) = \left(S - K\right)^+$. Note, we are not assuming this function is the option price. Rather, we shall show this is the option price. Consider a self-financed portfolio of bonds and stocks, with $\pi = C_S S$. Its value satisfies,

$$dV = \left(C_S S(\mu - r) + rV\right)dt + C_S \sigma S dW.$$

Moreover, by Itô's lemma, $C\left(t, S\right)$ is solution to

$$dC = \left(C_t + \mu S C_S + \frac{1}{2}\sigma^2 S^2 C_{SS}\right)dt + C_S \sigma S dW.$$

By subtracting the previous two equations, leaves:

$$dV - dC = \underbrace{(-C_t - rSC_S - \frac{1}{2}\sigma^2 S^2 C_{SS} + rV)}_{=-rC}dt = r\left(V - C\right)dt.$$

Hence, we have that $V_\tau - C\left(\tau, S_\tau\right) = \left(V_0 - C\left(0, S_0\right)\right)e^{r\tau}$, for all $\tau \in [0, T]$. Next, assume that $V_0 = C\left(0, S_0\right)$. Then, $V_\tau = C\left(\tau, S_\tau\right)$ and $V_T = C\left(T, S_T\right) = \left(S_T - K\right)^+$. That is, the portfolio $\pi = C_S S$ replicates the payoff underlying the option contract. Therefore, V_0 is the value of the option at time zero, even when a market for the option does not exist over its life.

11.4.2.3 By probabilistic arguments

Alternatively, we can use the general framework in Section 11.2.3 and arrive at Eq. (11.15). We simply set $\ell \equiv K$ in Eq. (11.10), and determine the dynamics of the stock price under $\hat{Q}_t$ and under Q_t, and then $\hat{Q}(S_T \geq K)$ and $Q(S_T \geq K)$ in Eq. (11.10). Under Q, the stock price is the usual geometric Brownian motion with drift equal to rS, and then, $Q(S_T \geq K) = \Phi(d - \sigma\sqrt{T-t})$, which explains the second term in Eq. (11.15). As for the first term, we can show that the Radon-Nikodym derivative, $\left.\frac{d\hat{Q}_t}{dQ_t}\right|_{\mathbb{F}_T} = \eta_t(T) \equiv \frac{e^{-r(T-t)}S_T}{S_t}$, is such that $\eta_\tau(T)$ is solution to

$$\frac{d\eta_\tau(T)}{\eta_\tau(T)} = -(-\sigma)\, dW_\tau,$$

such that by the Girsanov theorem (see Chapter 5), the stock price is solution to

$$d\ln S_t = \left(r + \frac{1}{2}\sigma^2\right) dt + \sigma d\hat{W}_t,$$

where $\hat{W}$ is a Brownian motion under $\hat{Q}$. It easily follows that $\hat{Q}(S_T \geq K) = \Phi(d)$, thereby completing the proof of the Black-Scholes formula.

11.4.3 Surprising cancellations and "preference-free" formulae

Chapter 5 explains that due to what Heston (1993a) (p. 933) quite aptly terms "a surprising cancellation," the constant μ doesn't show up in the final formula. Heston (1993a) shows that this property is not robust to modifications in the assumptions for the underlying asset price process. [In progress, Gamma processes, incomplete markets.]

However, even within a diffusion setting, the expected return on the option does of course depend on μ. By Eq. (11.14),

$$E\left(\frac{dC}{C}\right) = r + \underbrace{\frac{C_S}{C}\sigma S}_{\equiv \beta_S} \cdot \lambda, \tag{11.16}$$

where β_S is, simply, by Itô's lemma, the instantaneous volatility of the option returns, and λ is the unit risk-premium related to the fluctuations of the asset price, $\lambda = (\mu - r)/\sigma$.

11.4.4 Future options and Black's formula

Consider a *future option*, one that gives the buyer the right, not the obligation, to enter into a future contract for a specified price K, at time T, such that the payoff at time T is, $(F_T(S) - K)^+$, where $F_\tau(S)$ denotes the future price in Eq. (11.1), $F_\tau(S) = S_\tau e^{r(S-\tau)}$. It is easy to see that $F_\tau(S)$ is martingale under the risk-neutral probability Q. For example, assuming that S is a Geometric Brownian motion with volatility σ, we have that,

$$\frac{dF_\tau(S)}{F_\tau(S)} = \sigma d\tilde{W}_\tau, \quad \tau \in (t, S), \tag{11.17}$$

where $\tilde{W}$ is a Brownian motion under Q. Therefore, the price of a future option as of time t is,

$$C_{\mathrm{F}}(F_t(S), T-t; K) \equiv e^{-r(T-t)}\mathbb{E}_t(F_T(S) - K)^+ = e^{-r(T-t)}\left(F_t(S)\Phi(d) - K\Phi(d - \sigma\sqrt{T-t})\right), \tag{11.18}$$

where

$$d = \frac{\ln\left(\frac{F_t(S)}{K}\right) + \frac{1}{2}\sigma^2(T-t)}{\sigma\sqrt{T-t}},$$

and the second equality of Eq. (11.18) follows by the Black & Scholes formula, Eq. (11.15). Eq. (11.18) is the celebrated Black's (1976a) formula, which turns out to be very useful in the context of fixed income security pricing, as explained in Chapter 13. Appendix 2 provides an alternative derivation of Eq. (11.18), based on the pricing approach of Section 11.2, and the slightly more general assumption that the volatility of the future price in Eq. (11.17) is time-varying—but deterministic.

Chapter 13 explains that the property that future prices are martingales under the risk-neutral probability generalizes to one holding when interest rates are time-varying, under a certain probability called *forward probability* (see Chapter 13, Section 13.2).

11.4.5 Hedging

The "cloning" arguments in Section 11.4.2 suggest how to replicate a call option in a Black and Scholes market. We set up a portfolio with an amount θ^* in the underlying asset and the remaining in the money market account, where

$$\theta_t^* \equiv \Delta_t^{\text{BS}} \equiv \frac{\partial C_{\text{BS}}\left(S_t, T-t; K, \sigma\right)}{\partial S}, \tag{11.19}$$

and $C_{\text{BS}}\left(\cdot\right)$ denotes the Black & Scholes formula.

Eq. (11.19) can actually be simplified. Note that Black & Scholes formula is homogenous of degree one in S and K, that is, $C_{\text{BS}}\left(\lambda S, T-t; \lambda K, \sigma\right) = \lambda C_{\text{BS}}\left(S, T-t; K, \sigma\right)$ for any constant λ. Therefore, by Euler's theorem,

$$C_{\text{BS}}\left(S, T-t; K, \sigma\right) = \frac{\partial C_{\text{BS}}\left(S, T-t; K, \sigma\right)}{\partial S} S + \frac{\partial C_{\text{BS}}\left(S, T-t; K, \sigma\right)}{\partial K} K. \tag{11.20}$$

Comparing Eq. (11.20) with the Black-Scholes formula in Eq. (11.15), produces,

$$\Delta_t^{\text{BS}} = \Phi(d). \tag{11.21}$$

But why do we need to replicate derivatives, in practice? Because most of them are dealt with by investment banks, which simply act as financial intermediaries, trading derivatives on behalf of third parties, and being compensated through fees. Suppose an investment bank receives an order to sell a put. The bank would like to hedge against this put by creating a replicating portfolio such that the value of this portfolio be the same as the final payoff to be paid off to honour the sale. So hedging is needed to replicate the final payoffs required to honour the contracts giving rise to these payoffs. Standard market practice is to use the "Black-Scholes delta" in Eq. (11.21).

Note that at the same time, investment banks, not to mention funds, can undertake speculative trading activities aimed to implement specific views, such as those described in Section 11.5.5 below, in which case hedging doesn't necessarily need to be implemented. However, even in this case, hedging might be required to isolate the particular views a trading desk of the bank is taking. For example, Section 11.5.5 will explain that to express the view that equity volatility will raise, say, we cannot simply go long call options, because call prices are increasing both in volatility and the price underlying the option. A better solution is to go long an option, delta-hedged through Black-Scholes, as we shall explain.

11.4.6 Endogenous volatility

[In progress]

Hedges and crashes. Delta-hedging can lead to financial turmoil. The 1987 crash, and the conclusion of the Brady commission. The theory in the Brady's report is that the market was initially hit by bad news, and fell, triggering further sell-offs originating from program trading. Uninformed investors, who did not fully understand the nature of these developments, exited the market, and the fall they created would trigger additional program trades until the market could not reach the usual equilibrium where it used to stay, switching to another equilibrium through a discrete change—an equity market crash of more than -22% in just one day, the largest of financial history.

The flash crash of May 6th, 2010 might be considered a modern prototype version of the 1987 crash. According to the explanation put forward in a joint report by the SEC and the CFTC (SEC-CFTC, 2010), a mutual fund sold a quite large number of E-mini futures on the S&P 500 for mere hedging motives, but high frequency trading firms would then initiate arbitrage adjustments, by going long them and simultaneously shorting the equities representing the SPY—the ETF that tracks the S&P 500 index. To unwind these positions would normally require traders with fundamental views, but the majority of the players over those few minutes of trading were high frequency trading firms, who would then transact with each other over a self-exciting process, where trading aggressiveness intensifies as volume increases, due to the nature of algorithmic trading. In the end, large volume and sell-offs reinforced each other, whereby increasing volumes triggered algorithms to sell into a falling market, leading to a crash of nearly -10% in the equity market over a few minutes.

[Give a short outline of this section]

11.4.6.1 Hedges

A well-known definition is that of the *Gamma* of a derivative, which is the second order partial of the derivative price with respect to the underlying. The Gamma is always positive for long calls and puts, as these derivatives have positive convexity, as illustrated by Figure 11.1. Naturally, short calls and puts have negative gamma. In order for the statement "when gamma is negative, delta hedging involves buying on the way up and selling on the way down" to be true, we also have to consider whether the delta is positive or not (that is, whether the derivative price is increasing or decreasing in the underlying asset price). So we have four instances of hedging portfolios:

(i) Positive gamma: Buying on the way up and selling on the way down.

  (i.1) Hedging portfolios with positive delta, as required, for example, to hedge against the *sale* of a call. Positive delta means that the hedging portfolio relies on *buying* the assets underlying the call. When the price of these assets are up, the delta is also up, which implies we need to keep on buying even more of the assets underlying the hedging portfolio. On the other hand, when prices are down, the delta is also down, which implies holding less of the assets underlying the hedging portfolio, thereby leading to sell some these assets precisely when the market is down.

  (i.2) Hedging portfolios with negative delta, as required, for example, to hedge against the *sale* of a put. Negative delta means that the hedging portfolio relies on *selling* the assets underlying the put. In this case, delta is up when when prices are up.

However, this now simply means that we need to sell less! For example delta might have been $-\frac{1}{2}$ before the market was up and now delta is $-\frac{1}{4}$: that is, we need to buy back some of the assets underlying the hedging portfolio. When, instead, prices are down, delta is also down, which means we need to sell even more into a depressed market.

(ii) Negative gamma: Buying on the way down and selling on the way up.

(ii.1) Hedging portfolios with positive delta, as required, for example, to hedge against having gone *long* a put. Positive delta means that the hedging portfolio relies on *buying* assets underlying the put. Negative gamma now means that as soon as the price of these asset goes up (resp. down), we need to buy less (resp. buy more), so we sell when prices go up and buy when prices go down.

(ii.2) Hedging portfolios with positive delta, as required, for example, to hedge against having gone *long* a call. We are now *selling* the assets underlying the call. Negative gamma, here, means that as the price of these assets goes up (resp. down), we need to sell more (resp. sell less), so once again, we sell when prices go up and buy when prices go down.

How to implement these hedging portfolios, in practice, is still an open question, as this issue is necessarily model-based. Section 11.5.4, for example, shows that delta hedging under the Black-Scholes assumptions would lead the bank to eliminate the risk of fluctuations in the underlying stock price. At the same time, however, hedging through Black-Scholes leads the derivatives book quite messy once the fundamental assumption underlying the Black-Scholes world does not hold, namely that volatility changes randomly. In this case, "hedging" would rather look like a volatility view. To appropriately hedge, one has to rely on more complicated hedging strategies. For example, to hedge against an option in a world of stochastic volatility, we would need to use a stock, a bond, and, another ... option!

11.4.6.2 Crashes

[In progress]

Use a simple model, by Grossman, to illustrate how volatility is pumped-up by automatic mechanisms. Then, discuss a streamlined version of Gennotte and Leland, with asymmetric information, to illustrate the 1987 crash.

11.4.7 Properties of options in diffusive models

We consider a simple model in which the stock price is solution to,

$$\frac{dS_t}{S_t} = \mu\left(S_t\right)dt + \sqrt{2v\left(S_t\right)}dW_t. \tag{11.22}$$

Let $C\left(S_t, t, T\right)$ be the price of a European-style option at time t, which pays off $\psi\left(S_T\right) > 0$ at time T. We assume that $\psi\left(S\right)$ is differentiable, with $\psi'\left(S\right) > 0$. In the absence of arbitrage, C satisfies the following partial differential equation,

$$0 = C_t + C_S rS + C_{SS}S^2 v\left(S\right) - rC, \tag{11.23}$$

subject to the boundary condition, $C\left(S, T, T\right) = \psi\left(S\right)$.

[Plan of the section, in progress]

11.4.7.1 Passage of time

Call options are sometimes referred to as "wasting" assets because their value "tends" to decreases over time, due to a decrease in the value of the optionality, in a sense to be explained next. Define $\epsilon \equiv \frac{S}{C} C_S$ as the elasticity of the option price with respect to the asset price. For a call option, the elasticity $\epsilon > 1$ as noted in Section 11.3, such that Eq. (11.23) and convexity of C leaves:

$$C_t = -rC(\epsilon - 1) - C_{SS} v(S) < 0. \tag{11.24}$$

Furthermore, note that convexity increases as time to maturity decreases, with the limit case arising at maturity when C_{SS} shrinks to Dirac's delta. Therefore, the drop in value is the most severe in correspondence of shorter maturities. What happens, then, if we sell call options to buy them back later? Do options provide us with arbitrage opportunities? Obviously not. Selling a call and bying it later leads to profits only if market volatility is stable and the underlying asset price does not move too much as a result. This remark is indeed the motivation of some basic trading strategies known as "calendar spreads," and further explained in Section 11.6.1. A trading strategy is not an aribtrage opportunity though, only a way to implement a particular view.

Note that for a Euroepan put option, the elasticity of the put price with respect to the asset is *negative*, and can actually lead the RHS of Eq. (11.24) to change sign, especially for far out-of-the-money options, as Figure 11.5 reveals.

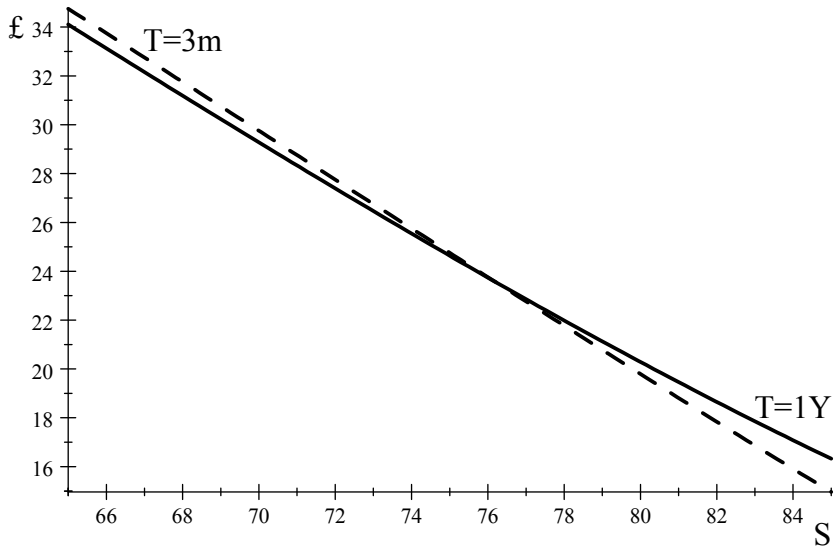


FIGURE 11.5. The value of a put option with strike $K = 100$, as time to maturity shrinks, as predicted by the Black & Scholes model, with volatility parameter equal to 20% and short-term rate $r = 1\%$. The solid line is the price corresponding to time to maturity $T =$ one year, and the dashed line is the price corresponding to time to maturity $T =$ three months.

11.4.7.2 Comparative statics of dynamic models

We derive properties of option prices in the context of diffusion processes, relying on methods suggested by Bergman, Grundy and Wiener (1996), and in Chapter 8 (see Proposition 8.1). We establish that if the stock price is solution to Eq. (11.22), the price of an European-style option inherits the properties of the final payoff: it is increasing and convex in the underlying price if

the payoff ψ is. These properties imply that the option price is increasing in the volatility of the underlying price.

Differentiate Eq. (11.23) with respect to S. The result is that $H \equiv C_S$ satisfies,

$$0 = H_t + H_S\left(rS + \tfrac{\partial}{\partial S}\left(S^2 v\left(S\right)\right)\right) + H_{SS}S^2 v\left(S\right), \tag{11.25}$$

subject to the boundary condition $H\left(S,\tau,T\right) = \psi'\left(S\right) > 0$. Therefore, we have that $H\left(S,\tau,T\right) = C_S\left(S,\tau,T\right) > 0$ for all S, due to results reviewed in Chapter 8 (see Proposition 8.1 and Appendix 1 in Chapter 8). That is, in a scalar diffusion setting, a European-style option price is increasing in the underlying whenever the $\psi\left(S\right)$ is increasing.

Next, we find conditions under which $C_{SS} > 0$. We differentiate Eq. (11.25) with respect to S, and $Z \equiv H_S = C_{SS}$ satisfies,

$$0 = Z_t + \left(rS + 2\tfrac{\partial}{\partial S}\left(S^2 v\left(S\right)\right)\right) Z_S + Z_{SS}S^2 v\left(S\right) - \left(-r - \tfrac{\partial^2}{\partial S^2}\left(S^2 v\left(S\right)\right)\right) Z, \tag{11.26}$$

subject to the boundary condition $Z\left(S,\tau,T\right) = \psi''\left(S\right)$. We now have that $H\left(S,\tau,T\right) = C_{SS}\left(S,\tau,T\right) > 0$ whenever the payoff is convex, $\psi''\left(S\right) > 0$ for all S. That is, in a scalar diffusion setting, a European-style option price is convex in the underlying, provided the final payoff is convex.

Finally, we implement the thought experiment of tilting the volatility of the underlying. Consider two markets A and B with prices $\left(C^i, S^i\right)_{i=A,B}$, in which the volatility of the asset price in market A is higher than that in market B, viz

$$\frac{dS^i_\tau}{S^i_\tau} = rd\tau + \sqrt{2v^i\left(S^i_\tau\right)}d\hat{W}_\tau, \quad i = A, B,$$

where $\hat{W}$ is Brownian motion under the risk-neutral probability, and $v^A\left(s\right) > v^B\left(s\right)$ for all s. The price difference in the two markets, $\Delta C \equiv C^A - C^B$, satisfies,

$$0 = \Delta C_t + \Delta C_S rS + \Delta C_{SS}S^2 v^A - r\Delta C + \left(v^A - v^B\right) C^B_{SS}S^2. \tag{11.27}$$

By the same results used to analyze Eq. (11.25) and (11.26), we now have that $\Delta C > 0$ whenever $C_{SS} > 0$. That is, if option prices are convex in the underlying price, they are increasing in the volatility of the asset price.

This result is reminiscent of the theory of mean-preserving spreads (see, e.g., Chapter 8). By increasing the volatility of the underlying, the holder of an European call (say) would benefit from the upside while not suffering losses on the way down—risk-neutral evaluation of traded assets benefits from an increasing volatility. We will see that that this conclusion might be reversed when it comes to assessing how the price of fixed income instruments reacts to changes in the volatility of the underlying fundamentals.

11.4.7.3 Counterexamples

[In progress]

11.4.7.4 Recovering risk-neutral probabilities

Consider the price of a European call,

$$C\left(S_t, t, T; K\right) = P\left(t,T\right)\int_K^\infty \left(x - K\right) q\left(x\middle|\, S_t\right) dx,$$

where Q is the risk-neutral probability and $q(x^+|\,x)dx \equiv dQ(x^+|\,x)$. Assuming that $\lim_{x\to\infty} xq\,(x|\,S) = 0$, and differentiating with respect to K leaves:

$$e^{r(T-t)}\frac{\partial C\,(S_t,t,T;K)}{\partial K} = -\int_K^\infty q\,(x|\,S_t)\,dx.$$

We can check this relation holds true in the Black-Scholes model, in Eq. (11.20). Let us differentiate again,

$$e^{r(T-t)}\frac{\partial^2 C\,(S_t,t,T;K)}{\partial K^2} = q\,(K|\,S_t)\,. \tag{11.28}$$

Eq. (11.28) allows us to recover the risk-neutral density using option prices. The Arrow-Debreu state density, $\mathcal{D}_{\text{AD}}\,(S_T = u|\,S_t)$, is given by,

$$\mathcal{D}_{\text{AD}}\,(S_T = u|\,S_t) = e^{r(T-t)}\,q\,(S_T|\,S_t)|_{S_T=u} = e^{2r(T-t)}\,\frac{\partial^2 C\,(S_t,t,T;K)}{\partial K^2}\bigg|_{K=u}.$$

These results are quite useful in applied work. They also help deal with the pricing of volatility contracts reviewed in Section 11.6, as explained in Appendix 4.

11.5 Stochastic volatility

11.5.1 Statistical models of changing volatility

11.5.1.1 ARCH and random variance models

Asset returns have time-varying volatility and their distributions are both heavy-peaked and tailed, as reviewed in Chapter 8. These empirical regularities are very well-known at least since the seminal work of Mandelbrot (1963) and Fama (1965). Engle (1982) and Bollerslev (1986) introduce the first parametric models aiming to capture these stylized facts through the celebrated Auto Regressive Conditionally Heteroskedastic (ARCH) models. ARCH models have played a prominent role in the analysis of many aspects of financial econometrics, such as the term structure of interest rates, the pricing of options, or the presence of time varying risk premiums in the foreign exchange market, as summarized by the classic survey of Bollerslev, Engle and Nelson (1994).

An ARCH model works as follows. Let $\{y_t\}_{t=1}^N$ be a record of observations on some asset returns, $y_t = \ln S_t/S_{t-1}$, where S_t is the asset price. The variance of y_t is, then, modeled as an autoregressive process, as follows:

$$y_t = a + \epsilon_t, \quad \epsilon_t|\,F_{t-1} \sim N(0,\sigma_t^2), \quad \sigma_t^2 = w + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{11.29}$$

where a, w, α and β are parameters and F_t denotes the information set as of time t. This model is known as the GARCH(1,1) model (Generalized ARCH). It was introduced by Bollerslev (1986), and collapses to the ARCH(1) model introduced by Engle (1982) once we set $\beta = 0$. In other words, the variance of the distribution of asset returns tomorrow, is linear in the expectation error, $(E_{t-1}\,(\epsilon_t^2) - \epsilon_t^2)$, and rises linearly with the current "realized variance," ϵ_t^2, viz

$$E_t\left(\epsilon_{t+1}^2\right) = w + (\alpha+\beta)\,\epsilon_t^2 + \beta\left(E_{t-1}\left(\epsilon_t^2\right) - \epsilon_t^2\right).$$

The quintessence of ARCH models is to make volatility dependent on the variability of past observations. An alternative formulation, initiated by Taylor (1986), makes volatility driven

by some unobserved components. This formulation gives rise to the *stochastic volatility* model. Consider, for example, the following stochastic volatility model,

$$\begin{aligned} y_t &= a + \epsilon_t, \quad \epsilon_t | \, F_{t-1} \sim N(0, \sigma_t^2); \\ \ln \sigma_t^2 &= w + \alpha \ln \epsilon_{t-1}^2 + \beta \ln \sigma_{t-1}^2 + \eta_t; \qquad \eta_t | \, F_{t-1} \sim N(0, \sigma_\eta^2) \end{aligned}$$

where a, w, α, β and σ_η^2 are parameters. The main difference between this model and the GARCH(1,1) model in Eq. (11.29) is that the volatility as of time t, σ_t^2, is not predetermined by the past forecast error, ϵ_{t-1}. Rather, this volatility depends on the realization of the stochastic volatility shock η_t at time t. This makes the stochastic volatility model considerably richer than a simple ARCH model. As for the ARCH models, SV models have also been intensively used, especially following the progress accomplished in the corresponding estimation techniques. The seminal contributions related to the estimation of this kind of models are mentioned in Mele and Fornari (2000). Early contributions that relate changes in volatility of asset returns to economic intuition include Clark (1973) and Tauchen and Pitts (1983), who assume that a stochastic process of information arrival generates a random number of intraday changes of the asset price.

11.5.1.2 ARCH and diffusive models

Under regularity conditions, ARCH models and stochastic volatility models behave essentially the same as the sampling frequency gets sufficiently high. Precisely, Nelson (1990) shows that ARCH models converge in distribution to the solution of the stochastic differential equations, in the sense that the finite-dimensional distributions of the volatility process generated by ARCH models converge towards the finite-dimensional distributions of some diffusion process, as the sampling frequence goes to infinity. Mele and Fornari (2000) (Chapter 2) contain a review of results relating to this type of convergence, and Corradi (2000) develops a critique related to the conditions underlying these convergence results. To illustrate, heuristically, consider the following model

$$\left\{ \begin{array}{lcl} d \ln S_\tau &=& (\cdots) \, d\tau + \sigma_\tau dW_\tau \\ d\sigma_\tau^2 &=& (\omega - \varphi \sigma_\tau^2) \, d\tau + \psi \sigma_\tau^2 dW_\tau^\sigma \end{array} \right. \tag{11.30}$$

where dW_τ and dW_τ^σ are correlated, with correlation ρ, and ω, φ, and ψ are some constants. Consider, further, the ARCH model

$$\left\{ \begin{array}{lcl} y_{\Delta,n+1} &=& \epsilon_{\Delta,n+1}, \quad \epsilon_{\Delta,n} = \sigma_{\Delta,n} \cdot u_{\Delta,n}, \quad u_{\Delta,n} \sim \text{NID}\,(0,1) \\ \sigma_{\Delta,n+1}^2 &=& w_\Delta + \alpha_\Delta \left(|\epsilon_{\Delta,n}| - \gamma \epsilon_{\Delta,n} \right)^2 + \beta_\Delta \sigma_{\Delta,n}^2 \end{array} \right. \tag{11.31}$$

where $y_{\Delta,n+1} = \ln (S_{\Delta,n+1}/S_{\Delta,n}) - E \left(\ln (S_{\Delta,n+1}/S_{\Delta,n}) \right)$, n and Δ refer to the indexing of observed data and the sampling frequency (weekly, say), and w_Δ, α_Δ, β_Δ are positive parameters, possibly depending on the sampling frequency, and $\gamma \in (-1, 1)$. The parameter γ allows to capture the Black-Christie-Nelson leverage effect (Black, 1976b; Christie, 1982; Nelson, 1991) discussed in Chapter 9. Note that the second of Eqs. (11.31) can be written as:

$$\begin{aligned} \sigma_{\Delta,n+1}^2 - \sigma_{\Delta,n}^2 = & \left[\Delta t^{-1} w_\Delta - \Delta t^{-1} \left(1 - \alpha_\Delta E \left(|u_{\Delta,n}| - \gamma u_{\Delta,n} \right)^2 - \beta_\Delta \right) \sigma_{\Delta,n}^2 \right] \Delta t \\ & + \Delta t^{-\frac{1}{2}} \alpha_\Delta \cdot \sigma_{\Delta,n}^2 \sqrt{\Delta t} \cdot \Delta W_{\sigma,n}, \end{aligned} \tag{11.32}$$

and $\Delta W_{\sigma,n} \equiv \left(|u_{\Delta,n}| - \gamma u_{\Delta,n} \right)^2 - E \left(|u_{\Delta,n}| - \gamma u_{\Delta,n} \right)^2$. The first two terms define the drift term for the variance process, and the last term is the diffusive component. Then, under regularity

conditions, the sample paths of S and σ^2 in Eqs. (11.31) converge to those of S and σ^2 in Eqs. (11.30), with a well-defined correlation coefficient ρ (see Fornari and Mele, 2006).[6]

11.5.2 Implied volatility, smiles and skews

In parallel to the previous time-series research, research on option prices in the 1980s was challenging the Black & Scholes assumption of a constant volatility. As we know, the Black & Scholes model relies on the assumption that the underlying asset price of is a geometric Brownian motion with constant volatility:

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t,$$

where W is a Brownian motion, and μ, σ are constants. As we also know, σ is the only parameter to enter the evaluation formula for a call option price.

This dependence leads to a crucial point. Not only is the assumption of a constant σ inconsistent with the time-series evidence reviewed in the previous section. It is also inconsistent with empirical evidence on the *cross-section* of option prices. Let $C_t^{\$}(K,T)$ denote the time t market price of a call expiring at T with strike K. Consider the price predicted by the Black-Scholes model, $C_{\text{BS}}(S_t, T-t; K, \sigma)$ in Eq. (11.15). Define the Black-Scholes *implied volatility*, IV say, as the value of σ that equates the Black-Scholes formula to the option market price

$$\text{IV} : C_t^{\$}(K,T) = C_{\text{BS}}(S_t, T-t; K, \text{IV}). \tag{11.33}$$

We know from Section 11.4.7 that the Black-Scholes option price is strictly increasing in σ. Therefore, there exists a unique value for IV such that Eq. (11.33) holds true. In fact, it is market practice to quote options in terms of implied volatilities, not prices. Moreover, implied volatility is the same for both the call and the put. Indeed, by the put-call parity in Eq. (11.7),

$$P_t^{\$}(K,T) = C_t^{\$}(K,T) - S_t + Ke^{-r(T-t)}. \tag{11.34}$$

This equation also holds for the Black-Scholes model for each σ, i.e. $P_{\text{BS}}(S_t, T-t; K, \sigma) = C_{\text{BS}}(S_t, T-t; K, \sigma) - S_t + Ke^{-r(T-t)}$. Subtracting this equation from Eq. (11.34) shows that the implied volatilities for a call and for a put options are the same.

If the Black & Scholes model holds, implied volatilities would be the same for each K. Yet, empirically, and at least since 1987, the cross section of implied volatilities exhibit striking characteristics. Define, in particular, the "moneyness of the option" as

$$mo \equiv \frac{S_t}{Ke^{-r(T-t)}}. \tag{11.35}$$

Before 1987, the pattern of implied volatilities was unclear, or U-shaped in $\frac{1}{mo}$ at best—a "smile." After the 1987 crash, this pattern would change to a "smirk," also referred to as "volatility skew." One possible explanation for these changes is that call and put options that

[6] Notably, assume that the following limits exist, $\lim_{\Delta t \downarrow 0} \Delta t^{-1} w_\Delta = \omega$, $\lim_{\Delta t \downarrow 0} \Delta t^{-1} \left(1 - \alpha_\Delta E\left(|u_{\Delta,n}| - \gamma u_{\Delta,n}\right)^2 - \beta_\Delta\right) = \varphi$, and $\lim_{\Delta t \downarrow 0} \Delta t^{-1/2} \sqrt{\text{var}\left(\Delta W_{2,n}\right)} \alpha_\Delta = \psi$. For example, suppose that $\gamma = 0$, in which case the random component of the diffusive term in Eq. (11.32) collapses to $\Delta W_{\sigma,n} = u_{\Delta,n}^2 - E(u_{\Delta,n}^2)$, and the moment condition for the diffusive component is $\psi = \lim_{\Delta t \downarrow 0} \Delta t^{-1/2} \sqrt{2} \alpha_\Delta$. Intuitively, in this case ΔW_σ is an IID sequence of centered chi-square variates with one degree of freedom (and variance $\varpi = 2$), and stands for the discrete version of the Brownian motion increments dW_σ in the second of Eqs. (11.30).

are deep-in-the-money and call or put options that are deep-out-of the money are relatively less liquid than at-the-money options. These characteristics lead to a liquidity risk-premium. Since the Black-Scholes option price is increasing in volatility, the implied volatility is U-shaped in $\frac{1}{mo}$.

Figures 11.6 and 11.7 illustrate how smiles and smirks arise as a result of a different assumption, namely that asset returns have random volatility. We rely on the celebrated Heston's (1993) model in which volatility is random, and refer the reader to Section 11.5.4 (see Eq. (11.55)) for additional details regarding this model.[7]

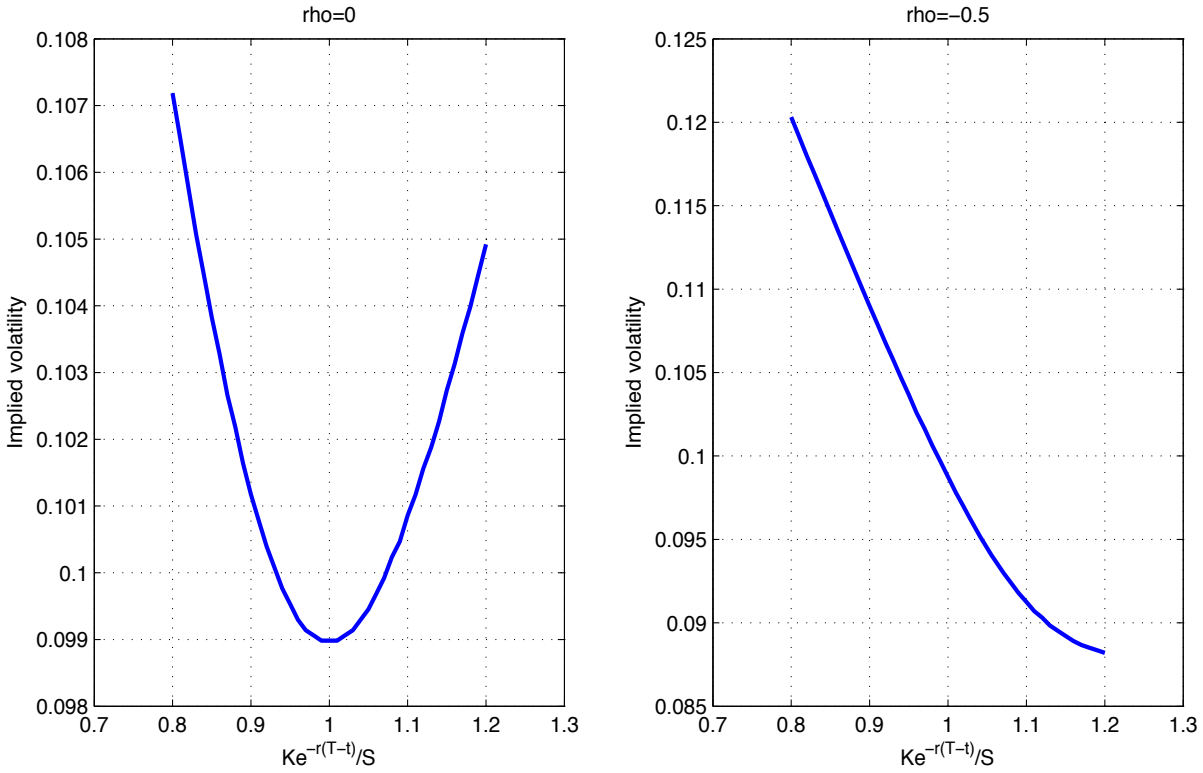


FIGURE 11.6. Smile and smirk predicted by the Heston model in Eq. (11.55), with parameters fixed at $\kappa = 2$, $\omega = 0.01$, $\xi = 0.1$ and, for the left-hand panel (the smile) $\rho = 0$, and the right-hand panel (the skew) $\rho = -0.5$. The initial values of the asset price and volatility are $S = 100$, and $\sqrt{v_t} = \sqrt{\omega}$, and the short-term rate $r = 0$, and the maturity of the option is six months.

The intuition underlying the patterns in Figure 11.6 is the following. The Black & Scholes model relies on the assumption asset returns are log-normally distributed. However, this assumption is not correct: in a stochastic volatility model, market pricing operates through an alternative distribution, one for which tail events are given a larger weight, due to market fears of extreme outcomes. To illustrate, the market may fear the stock price will fall below a given level, say $\underline{K}$, more than the Black & Scholes model predicts. As a result, the market density should have a left tail ticker than the log-normal, for values of $S < \underline{K}$.

[7]The densities in Figure 11.7 are those of $1 - \Phi\left(d - \sigma\sqrt{T-t}\right)$ in Eq. (11.15) for Black & Scholes, and of $1 - P_2\left(\ln S, \sigma^2, t\right)$ in Eq. (11.60) for Heston—both densities are with respect to K.

This possibility is illustrated by the left panel of Figure 11.7, which depicts the risk-neutral distributions of the Black & Scholes model as well as our Heston's model (with random volatility), taken to be the "truth." Heston's model does generate thick tails, as discussed below. Then, a market density with a left tail thicker than that of Black & Scholes implies that the probability deep-out-of-the-money *puts* (i.e., those with low strikes) will be exercized is higher under the market density than under the log-normal. For this reason, the implied volatility needed to price deep-out-of-the-money *puts* is higher than that required to price at-the-money calls and puts.[8]

For example, suppose that given the current stock price, $S_t = 100$, the implied Black-Scholes volatility is 10% for a given maturity τ and a strike $K = 90$, i.e.,

$$P_{\text{BS}}\left(100, T-t; K, 10\%\right) = P_t^{\$}\left(K, T\right), \quad K = 90.$$

Next, consider an even lower strike, say $K = 80$. In this case, the Black-Scholes put price would most likely not equal the market price: while the Black-Scholes and the market prices both lower with a decreased strike, the Black-Scholes price decreases more because the probability of exercise is lower under this theoretical market than under the actual market,

$$P_{\text{BS}}\left(100, T-t; K, 10\%\right) < P_t^{\$}\left(K, T\right), \quad K = 80.$$

That is, the implied Black-Scholes volatility would now need to raise from 10% to 11%, say: when $K = 80$, $P_{\text{BS}}\left(100, T-t; K, 11\%\right) = P_t^{\$}\left(K, T\right)$.

At the other extreme, the market may attach a higher likelihood that the stock price will be above some $\bar{K}$ than predicted by Black & Scholes: the market density might have a right tail ticker than the log-normal for values of $S > \bar{K}$, implying a higher probability that deep-out-of-the-money *calls* (i.e., those with high strike prices) will be exercized, compared to the log-normal. The left panel of Figure 11.7 illustrates that in this case, the implied volatility needed to price deep-out-of-the-money *calls* exceeds the implied vol needed to price at-the-money calls and puts. The second effect has disappeared since the 1987 crash, as mentioned, and for reasons similar to those underlying the right panel of Figure 11.7, leaving the "smirk" of the right panel of Figure 11.6.

This section suggests that out-of-the money options contain important information regarding market expectations about future volatility. Indeed, the CBOE-VIX index does aggregate the prices of out-of-the-money options to convey an estimate of the volatility expected to arise corrected by risk. Section 11.8 develops details on this index and explains that it links to the fair value of a variance swap, i.e. a contract in which a counterparty is insured against fluctuations in future volatility.

[8]Note that if a call (put) option is out-of-the-money for a given strike, a put (call) is in-the-money option for the same strike.

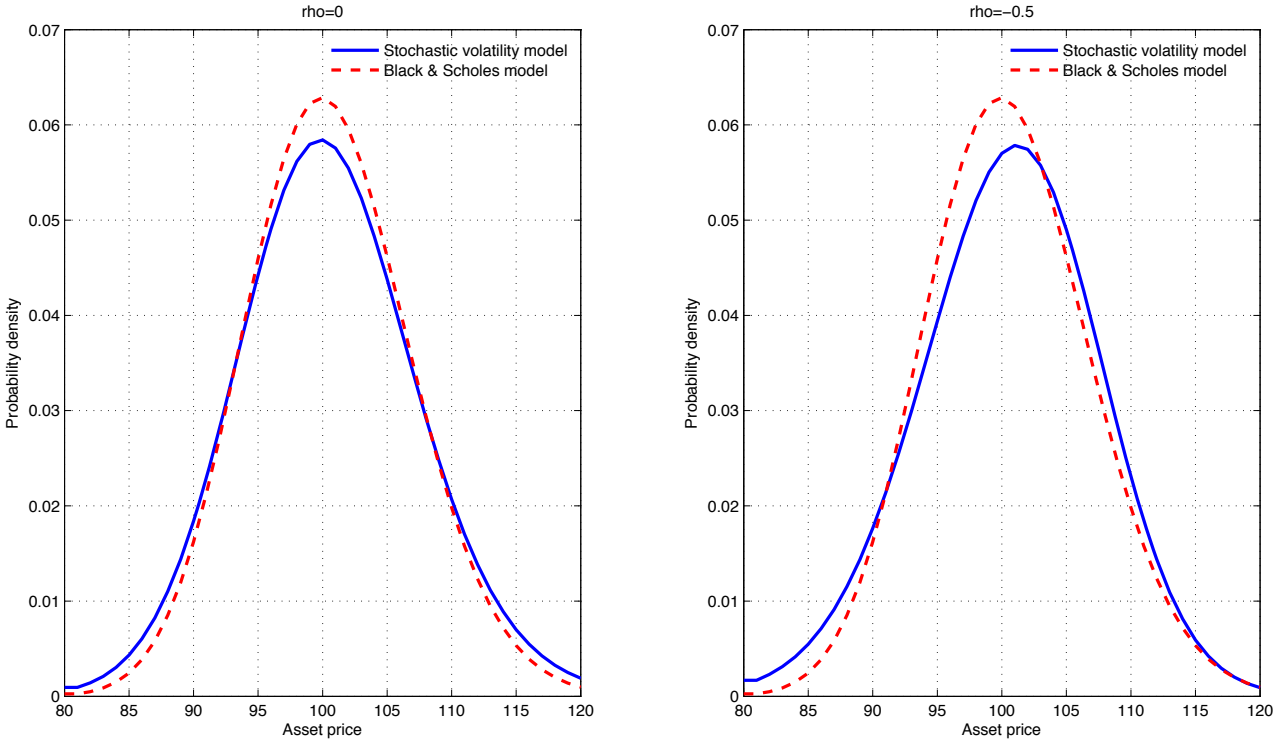


FIGURE 11.7. Risk-neutral densities predicted by the Black & Scholes model (dashed line) and the Heston model in Eq. (11.55). The Black & Scholes volatility parameter is $\sigma = 9\%$, and Heston's parameters are fixed at $\kappa = 2$, $\omega = 0.01$, $\xi = 0.1$ and, for the left-hand panel $\rho = 0$, and the right-hand panel $\rho = -0.5$. The initial values of the asset price and volatility are $S = 100$, and $\sqrt{v_t} = \sqrt{\omega}$, the short-term rate $r = 0$, and maturity is six months.

While the previous conclusions rely on numerical results, an explanation of smiles is available since the early 1990s (see, e.g., Ball and Roma, 1994; Renault and Touzi, 1996). To illustrate, consider the continuous time model,

$$\left\{ \begin{array}{rcl} \dfrac{dS_t}{S_t} & = & \mu dt + \sigma_t dW_t \\ d\sigma_t^2 & = & b(S_t, \sigma_t)dt + a(S_t, \sigma_t)dW_t^{\sigma} \end{array} \right. \tag{11.36}$$

where W_σ is Brownian motion correlated with W, with instantaneous correlation equal to ρ, and b and a are some functions satisfying regularity conditions. The drift function, b, is needed to generate mean-reverting behavior in stochastic volatility, σ_t^2, a characteristic we require in exactly the same spirit of assumptions required regarding the dynamics of interest rates (see Chapter 13).

The option price is $C\left(S_t, \sigma_t^2, T-t\right) = e^{-r(T-t)}\mathbb{E}_Q\left[\left(S_T - K\right)^+ \middle| S_t, \sigma_t^2\right]$, where $\mathbb{E}_Q\left[\cdot\right]$ is the expectation under some risk-neutral probability Q. Assume that the correlation $\rho = 0$. Then, the implied volatility predicted by this model satisfies,

$$\text{IV} : C\left(S_t, \sigma_t^2, T-t\right) = C_{\text{BS}}\left(S_t, T-t; K, \text{IV}\right),$$

and is U-shaped with respect to $\frac{1}{mo}$. Indeed, let $X = \ln(mo)$, where mo denotes the option moneyness, as defined in Eq. (11.35). In Appendix 3, we show that it holds, approximately, that in a model with stochastic volatility and zero correlation, IV is, approximately, a quadratic function of X, $\text{IV}(X, T-t)$ say, with a minimum occurring at $X = 0$, just as in the left panel of Figure 11.6, viz

$$\text{IV}(X, T-t) \approx \mu_V(t) + \frac{1}{2}\left(\frac{X^2 - \left(\frac{1}{2}\mu_V^2(t)(T-t)\right)^2}{\mu_V^3(t)(T-t)}\right) \cdot var_t(\sqrt{\tilde{V}_{t,T}}), \tag{11.37}$$

where $\mu_V(t) = \mathbb{E}_t(\sqrt{\tilde{V}_{t,T}})$ and $\tilde{V}_{t,T}$ is the "average variance," or future realized variance:

$$\tilde{V}_{t,T} \equiv \frac{1}{T-t}\int_t^T \sigma_\tau^2 d\tau. \tag{11.38}$$

These properties interestingly link to a compelling lesson we learnt over the early statistical literature on ARCH and random variance models: random changes in volatility lead to a return distribution with tails thicker than the normal—one with kurtosis larger than three (Mandelbrot, 1963; Fama, 1965; Nelson, 1990; Mele and Fornari, 2000)—a feature that the Heston's model clearly illustrates in Figure 11.7. For example, we know from Nelson (1990), that even if unexpected returns are conditionally normally distributed, they are approximately, and unconditionally, Student's t, once we assume their variance follows a GARCH(1,1) process.

Mathematically, denote the unexpected returns as of time t with ϵ_t, and suppose that $\epsilon_t = z_t\sigma_t$, where $z_t \sim \text{NID}(0,1)$ and σ_t, the conditional volatility of ϵ_t, is some random process. Then we have, by Jensen's inequality, that $E(\epsilon_t^4) = E(z_t^4)E(\sigma_t^4) \geq E(z_t^4)\left[E(\sigma_t^2)\right]^2 = E(z_t^4)\left[E(\epsilon_t^2)\right]^2$, which is an equality when σ_t is not random. It follows that the kurtosis, $\text{Kurt} \equiv \frac{E(\epsilon_t^4)}{[E(\epsilon_t^2)]^2} \geq E(z_t^4) = 3$. That is, random volatility makes the unconditional return density leptokurtotic even when the conditional is normal. Although these calculations relate to unconditional densities, similar conclusions would apply to conditional: random volatility makes a T-day conditional density leptokurtotic even when the one-day conditional is normal.

As a result of this leptokurtoticity in asset returns, the probability out-of-the money options is exercized is larger than that implied by the log-normal distribution—the smile effect. As for the smirk effect, we need $\rho < 0$, as shown by Figure 11.7. Intuitively, when $\rho < 0$, the left tail of the return distribution is thicker than the right, thereby making out-of-the money puts most valuable.

The model in Eqs. (11.36) has been extended to one with jumps, where the variance process follows a mean-reverting process such as:

$$d\sigma_t^2 = \kappa\left(\omega - \sigma_t^2\right)dt + \psi\sigma_t dW_t^\sigma + \mathcal{S} \cdot dJ_t,$$

where J_t is a Poisson process with intensity v (see Chapter 5, Section 5.7), $\mathcal{S} > 0$ is the size of the jump, which we suppose to be constant for illustration purposes only, and, finally, κ, ω and ψ are constants. In this model, the presence of positive jumps, $v > 0$, makes the left tail of the return distribution thicker, when $\rho < 0$. Therefore, we need a high κ to avoid a too thicker distribution. With $v = 0$, instead, a thicker distribution can only be obtained through lower values of κ.

Naturally, the effects illustrated in this section mostly refer to explanations related to stochastic volatility although in general, they might arise for other reasons leading to leptokurticity, such feedback effects. [In progress, mention the literature on feedback effects of the 1990s.]

An even simple channel is one in which stock prices are simply driven by jumps, which make the left tail of the distribution thicker than the right one, as in the following model,

$$\frac{dS_t}{S_t} = (\cdots)\, dt + \sigma dW_t - \mathcal{S} \cdot dJ_t,$$

where σ is not constant. Naturally, the point of assuming σ constant is to isolate "crashophobia" as an explanation for the skew. [In progress]

11.5.3 Option pricing with stochastic volatility

Stochastic volatility might lead to *market incompleteness*,[9] the situation in which we cannot hedge against future contingencies based on trading available securities (see Chapter 5). In the context we are dealing with in this section, we cannot replicate the payoff of a European stock option, because the number of assets available for trading (one stock) is less than the sources of risk—the two Brownian motions in Eq. (11.36), W_τ^σ and W_τ. Intuitively, the option price is $C_\tau \equiv C\left(S_\tau, \sigma_\tau^2, T-\tau\right)$, thus being driven by both the asset price, S_τ, and stochastic variance, σ_τ^2. The point is that σ_τ^2 and, then, C_τ, are driven by W_τ^σ, while the value of a "would be replicating" portfolio (only comprising the underlying asset) would only be driven by W_τ. Therefore, this portfolio cannot replicate the option, and the option price cannot be preference-free. In a nutshell, and as explained in Part I, preference-free pricing relies on the possibility to perfectly replicate a derivative and the unique price of this derivative is then that of the replicating portfolio, independently of risk-appetite.[10] Naturally, markets can be "completed" by the option, although in this case, option pricing is not preference-free as we shall show in the next sections.

To summarize, stochastic volatility entails two inextricable consequences: (i) There is an infinity of option prices consistent with absence of arbitrage, which correspond to the many risk-neutral probabilities consistent with the model: there are many risk-adjustments that we can make to the drift term of the variance process in Eqs. (11.36); (ii) there cannot be perfect hedging strategies only relying on the underlying asset. As regards point (ii), we might, alternatively, either (a) use a strategy, which albeit not self-financed, would still allow for a perfect replication of the claim, or (b) a self-financed strategy that would apply to some misspecified model. In case (a), the strategy leads to a hedging cost process. In case (b), the strategy leads to a tracking error process, although there might be situations where the claim can be "super-replicated," as explained below.

We start with a short detour about how to understand replicability in a general context, and proceed to option evaluation in subsequent subsections.

11.5.3.1 Spanning and cloning

A set of securities spans a set of payoffs, if any point in that set can be generated by a linear combination of the security prices. As explained in Chapter 5, the set of payoffs may include those promised by a contingent claim, for example, that promised by a European call, or final

[9] Stochastic volatility might not necessarily lead to market incompleteness. Mele (1998) (p. 88) considers a "circular" market with m asset prices, where (i) asset price no. i exhibits stochastic volatility, and (ii) this stochastic volatility is driven by the Brownian motion driving the $(i-1)$-th asset price. Therefore, in this market, each asset price is solution to Eqs. (11.36) and yet markets are complete.

[10] Note that the payoff of the option at the expiration T is $(S_T - K)^+$ and does not depend on σ_τ^2. However, the value of the option at any $\tau < T$ does depend on σ_τ^2 because, intuitively, the risk-neutral expectation of $(S_T - K)^+$ conditional on the information set at τ is calculated through the risk-neutral transition density π say, which obviously depend on σ_τ^2: $\pi\left(S_T | S_\tau, \sigma_\tau^2\right)$.

consumption, as in Harrison and Kreps (1979) and Duffie and Huang (1985). Chapter 5 relies on this "spanning" property and solves for consumption-portfolio choices through martingale methods. In this section, we show how spanning helps define replicating strategies with the purpose of pricing "redundant" assets. Consider the following model (introduced in Chapter 5), where asset prices are assumed to be driven by a d-dimensional Brownian motion W_t,

$$dy_t = \varphi(y_t)\,dt + \Sigma(y_t)\,dW_t, \tag{11.39}$$

where φ and Σ are vector and matrix valued functions. The value of a portfolio is $V_t = \theta_t \cdot S_{+,t}$, where S_+ denotes the vector of the security prices and the money market account, and θ_t is a portfolio process with the same notation as in Chapter 5. As explained in Chapter 5 (Section 5.3.1) the value of a self-financed portfolio satisfies $dV_t = \theta_t \cdot dS_{+,t}$, or,

$$dV_t = \left(\pi_t^\top(\mu_t - \mathbf{1}_m r_t) + r_t V_t - C_t\right)dt + \pi_t^\top \sigma_t dW_t, \tag{11.40}$$

where $\pi \equiv (\pi_1, \cdots, \pi_m)^\top$, $\pi_i \equiv \theta_i S_i$, $\mu \equiv (\mu_1, \cdots, \mu_m)^\top$, S_i is the price of the i-th asset, μ_i is its drift and σ is the volatility matrix of the price process. Chapter 5 utilizes the risk-neutral probability, $\mathcal{Q}$, to help characterize the securities span. Let us, now, do the same under the physical probability, P. In our context, asset prices are semimartingales under P,

$$dA_t = dF_t + \tilde{\gamma}_t dW_t, \tag{11.41}$$

where F is a process with finite variation, and $\tilde{\gamma}$ satisfies regularity conditions. Let us conjecture that $V_t = A_t$ for all t. By the *unique decomposition property* mentioned in Chapter 5 (Section 5.2), drifts and diffusion components of V and A must be the same. Regarding the diffusion terms, the portfolio π_t in Eq. (11.40) needs to satisfy:

$$\tilde{\gamma}_t = \pi_t^\top \sigma_t. \tag{11.42}$$

Regarding the drifts, we equate those of V and F, obtaining,

$$\frac{dF_t}{dt} = \pi_t^\top(\mu_t - \mathbf{1}_m r_t) + r_t V_t = \pi_t^\top(\mu_t - \mathbf{1}_m r_t) + r_t A_t, \tag{11.43}$$

where the second equality holds by conjecture that $A_t = V_t$ for all t.

The conjecture that $A_t = V_t$ is indeed correct as A_t is the *price of a traded asset* (by a slight generalization of the argument in Section 5.2 of Chapter 5), and the *assumption of complete markets*. However, Eq. (11.42) has no solutions for π when $m < d$, such that there are no self-financing strategies that replicate A_t: the number of available assets is too small to span all possible events—markets are incomplete according to the definitions in Chapter 5. We shall return to this topic a few times in this chapter.

Option evaluation fits perfectly within this framework. For example, a call price could be undertood as being a differentiable function $C(t, \boldsymbol{y}_t)$, and by Itô's lemma, a special case of Eq. (11.41),

$$dC = (LC)\,dt + (C_y \cdot \Sigma)\,dW,$$

where LC is the usual infinitesimal generator, C_y is the vector containing the partials of C with respect to each component of $\boldsymbol{y}$. It can be replicated if, at least, $m = d$, in which case $C_y\Sigma = \pi^\top\sigma$. For example, in the context of Black & Scholes and Merton of Section 11.4.1, we have that $m = d = 1$, and the asset price S is the only state variable in Eq. (11.39)—a geometric Brownian motion with parameters (μ, σ), i.e. $\varphi(S) = \mu S$ and $\Sigma(S) = \sigma S$. Then, $C_y \cdot \Sigma = C_S \sigma S$, $\pi = C_S S$, leading to the celebrated Black-Scholes hedge ration in Eq. (11.19).

How does replicability work in the context of stochastic volatility? It may actually work, although then we might not have preference-free formulae, as we now explain.

11.5.3.2 Replication

Let us suppose that the price of the underlying asset is solution to Eqs. (11.36). The rational pricing function of a European-style option is $C_\tau^T = C\left(S_\tau, \sigma_\tau^2, T-\tau\right)$. Suppose two such options are traded, with prices $C_\tau^{T_1}$ and $C_\tau^{T_2}$, where we take $T_1 < T_2$. We cannot replicate $C_\tau^{T_1}$ by trading the underlying asset and the money market account. Indeed, let V be the value of a self-financed strategy including the asset price and the money market account, which obviously satisfy:

$$dV = \left(\pi(\mu - r) + rV\right) dt + \pi\sigma dW, \tag{11.44}$$

where π is the value of the invested underlying asset. Instead, the price of the first option satisfies:

$$dC^{T_1} = \mathcal{L}\left(C^{T_1}\right) dt + C_S^{T_1}\sigma S dW + C_{\sigma^2}^{T_1} a dW_\sigma, \tag{11.45}$$

where $\mathcal{L}$ is the infinitesimal generator, $\mathcal{L}C^{T_1} \equiv \frac{\partial C^{T_1}}{\partial t} + \mu S C_S^{T_1} + bC_{\sigma^2}^{T_1} + \frac{1}{2}\sigma^2 S^2 C_{SS}^{T_1} + \frac{1}{2}a^2 C_{\sigma^2\sigma^2}^{T_1} + \rho a\sigma S C_{S\sigma^2}^{T_1}$. We see that we cannot match the diffusion coefficients of the option price in Eq. (11.45) through that in Eq. (11.44).

Instead, we might replicate the price of the first option, C_{T_1}, through a self-financed portfolio strategy including (i) the underlying asset, (ii) the option expiring at T_2, and (iii) the money market account. The value V of this strategy satisfies:

$$dV = \left(\pi_1\left(\mu - r\right) + \pi_2\left(\frac{\mathcal{L}(C^{T_2})}{C_{T_2}} - r\right) + rV\right) dt + \pi_1\sigma dW + \pi_2\left(\frac{C_S^{T_2}}{C^{T_2}}\sigma S dW + \frac{C_{\sigma^2}^{T_2}}{C^{T_2}} a dW_\sigma\right), \tag{11.46}$$

where π_1 is the invested asset value, and π_2 is the value of the investment in the second option. We match the diffusion coefficients of Eq. (11.45) and Eq. (11.46), and obtain:

$$\pi_1 = \left(\frac{C_S^{T_1}}{C_S^{T_2}} - \frac{C_{\sigma^2}^{T_1}}{C_{\sigma^2}^{T_2}}\right) C_S^{T_2} S, \quad \pi_2 = \frac{C_{\sigma^2}^{T_1}}{C_{\sigma^2}^{T_2}} C^{T_2}. \tag{11.47}$$

Replacing these expressions into Eq. (11.46), and equating the drift of Eq. (11.46) to that of Eq. (11.45), leaves:

$$\frac{\mathcal{L}\left(C^{T_1}\right) - rC^{T_1} - C_S^{T_1}\left(\mu - r\right) S}{C_{\sigma^2}^{T_1}} = \frac{\mathcal{L}\left(C^{T_2}\right) - rC^{T_2} - C_S^{T_2}\left(\mu - r\right) S}{C_{\sigma^2}^{T_2}}. \tag{11.48}$$

These two ratios agree. They must then be equal to some process $a \cdot \Lambda^\sigma$, say, which is independent of the maturity of the option. Therefore, we obtain that,

$$\frac{\partial C}{\partial t} + rSC_S + \left(b - a\Lambda^\sigma\right) C_{\sigma^2} + \frac{1}{2}\sigma^2 S^2 C_{SS} + \frac{1}{2}a^2 C_{\sigma^2\sigma^2} + \rho a\sigma S C_{S\sigma^2} = rC. \tag{11.49}$$

The interpretation of Λ^σ is that of the unit risk-premium required to face the risk of stochastic fluctuations in volatility. The problem, absence of arbitrage does not suffice to recover a unique Λ^σ. By the Feynman-Kac stochastic representation of the solution to a partial differential equation, there are many solutions to Eq. (11.49),

$$C\left(S_t, \sigma_t^2, T-t\right) = e^{-r(T-t)}\mathbb{E}_{Q^\Lambda}\left(\left(S_T - K\right)^+ \middle| S_t, \sigma_t^2\right), \tag{11.50}$$

where Q^Λ is *a* risk-neutral probability, induced by the many Λ that are consistent with absence of arbitrage.

The previous derivation suggests two possible uses of the portfolio strategies of Eqs. (11.47). The first, obvious, is hedging. We can always hedge the first option with π_1 and π_2 in Eqs. (11.47). The second is more subtle. If we really think our evaluation model for the first option is better than the market, we can always synthesize the first option with the portfolio in Eqs. (11.47), and replicate the payoff of the first option at expiration, $\psi(S_{T_1})$.

Eqs. (11.48) and (11.49) can be interpreted as APT relations. Indeed, let us define the unit risk-premium related to the fluctuations of the asset price, $\lambda = (\mu - r)/\sigma$. Then, Eq. (11.48) or Eq. (11.49) imply that,

$$\frac{\mathcal{L}(C)}{C} = E\left(\frac{dC}{C}\right) = r + \underbrace{\frac{C_S}{C}\sigma S}_{\equiv\beta_S} \cdot \lambda + \underbrace{\frac{C_{\sigma^2}}{C}a}_{\equiv\beta_{\sigma^2}} \cdot \Lambda^\sigma,$$

where β_S is the beta related to the volatility of the option price induced by fluctuations in the stock price, S, and β_{σ^2} is the beta related to the volatility of the option price induced by fluctuations in the return volatility. It is a generalization of the APT relation in Eq. (11.16) that holds for the Black & Scholes model.

11.5.3.3 Market completeness

Derivatives can complete markets. Remind the reader this is the case when $C_\sigma > 0$ (Section 5.5.3). Develop conditions under which this is true. It's a generalization to the one-dimensional case analyzed in Section 11.4.7. [In progress]

11.5.3.4 Pricing formulae

Hull and White (1987), Scott (1987) and Wiggins (1987) develop the first option pricing models with stochastic volatility. Heston (1999b) provides an analytical solution assuming an affine model for the variance process, for otherwise, we need to solve through numerical methods such as Montecarlo simulation or the numerical solution to partial differential equations.

Hull & White

Hull and White (1987) derive a first pricing formula based on a continuous-time model where asset returns and volatility are uncorrelated,

$$\left\{\begin{array}{lcl} \dfrac{dS_t}{S_t} & = & rdt + \sigma_t dW_t \\ \dfrac{d\sigma_t^2}{\sigma_t^2} & = & \mu_v d\tau + \xi dW_t^\sigma \end{array}\right. \tag{11.51}$$

where W and W_σ are uncorrelated Brownian motions, defined under the risk-neutral probability. They show that the option price takes the following form:

$$C\left(S_t, \sigma_t^2, T-t\right) = \mathbb{E}_t^{\tilde{V}}[C_{\text{BS}}(S_t, T-t; K, \sqrt{\tilde{V}_{t,T}})], \tag{11.52}$$

where $C_{\text{BS}}(S_t, T-t; K, \sqrt{\tilde{V}_{t,T}})$ denotes the usual Black-Scholes formula in Eq. (11.15), evaluated at the average variance defined as in Eq. (11.38), and $\mathbb{E}_t^{\tilde{V}}$ is the conditional risk-neutral expectation taken with respect to the laws generating $\tilde{V}_{t,T}$. According to Eq. (11.52), the option price is simply the Black & Scholes formula averaged over all possible "values" taken by

future average variance, $\tilde{V}_{t,T}$. Accordingly, the authors provide a Taylor's expansion around the conditional expectation of $\tilde{V}_{t,T}$, $m_V(t) = \mathbb{E}_t(\tilde{V}_{t,T})$,

$$\begin{aligned} C\left(S_t, \sigma_t^2, T-t\right) = & C_{\text{BS}}(S_t, T-t; K, \sqrt{m_V}) \\ & + \frac{1}{2} \left. \frac{\partial^2 C_{\text{BS}}(S_t, T-t; K, \sqrt{V})}{\partial V^2} \right|_{V=m_V(t)} \cdot var_t(\tilde{V}_{t,T}) \\ & + \frac{1}{6} \left. \frac{\partial^3 C_{\text{BS}}(S_t, T-t; K, \sqrt{V})}{\partial V^3} \right|_{V=m_V(t)} \cdot skew_t(\tilde{V}_{t,T}) + \cdots \end{aligned} \tag{11.53}$$

In fact, Eq. (11.53) is a general formula applying to models more general than that in Eq. (11.51), say one of the models encompassed by Eqs. (11.36), provided of course W and W_σ are uncorrelated, as explained more formally in Appendix 3. In Appendix 3, we also explain that Hull & White Equation (11.52) can be generalized to the case in which $W_\tau = \rho_\tau W_\tau^\sigma + \sqrt{1-\rho_\tau^2} dZ_\tau$, where Z is a standard Brownian motion and ρ_τ is a stochastic correlation. Romano and Touzi (1997) show that in this case, and provided $\rho(\cdot)$, and $b(\cdot), a(\cdot)$ in Eqs. (11.36) are independent of $S(\cdot)$,

$$\begin{aligned} C\left(S_t, \sigma_t^2, \rho_\tau, T-t\right) &= \mathbb{E}_{\tilde{V}}[C_{\text{BS}}(S_t L_{t,T}, T-t; K, \sqrt{\tilde{V}_{t,T}^\rho})] \\ L_{t,T} \equiv e^{-\frac{1}{2}\int_t^T \rho_\tau^2 \sigma_\tau^2 d\tau + \int_t^T \sigma_\tau \rho_\tau dW_\tau^\sigma}, & \quad \tilde{V}_{t,T}^\rho = \frac{1}{T-t}\int_t^T \sigma_\tau^2 \left(1-\rho_\tau^2\right) d\tau \end{aligned} \tag{11.54}$$

Heston

Heston (1993b) develops an analytical solution for European options in a model with stochastic volatility, in which the asset price is solution to

$$\left\{ \begin{array}{lcl} d\ln S_t &=& \left(r - \frac{1}{2}\sigma_t^2\right) dt + \sigma_t dW_{1t} \\ d\sigma_t^2 &=& \kappa\left(\omega - \sigma_t^2\right) dt + \xi\sigma_t \left(\rho dW_{1t} + \sqrt{1-\rho^2} dW_{2t}\right) \end{array} \right. \tag{11.55}$$

where $\kappa, \omega, \xi, \rho$ are parameters. Thus, the instantaneous variance is a square-root process. We can evaluate the payoff of call relying on the general formula in Eq. (11.10) of Section 11.2, by which

$$e^{-r(T-t)}\mathbb{E}_t\left(S_T - K\right)^+ = S_t \cdot \hat{Q}_t\left(S_T \geq K\right) - e^{-r(T-t)} K \cdot Q_t\left(S_T \geq K\right), \tag{11.56}$$

where $\hat{Q}_t(S_T \geq K)$ and $Q_t(S_T \geq K)$ are two probabilities with (time t-conditional) densities $\hat{q}_t^{\text{m}}(S_T)$ and $q_t^{\text{m}}(S_T)$; $q_t^{\text{m}}(S_T)$ is the risk-neutral marginal density of S_T, $\hat{q}_t^{\text{m}}(S_T)$ is another marginal density of S_T, with Radon-Nikodym derivative with respect to $q_t^{\text{m}}(S_T)$ given in Eq. (11.11),

$$\eta_t(T) = \frac{\hat{q}_t^{\text{m}}(S_T)}{q_t^{\text{m}}(S_T)} = \frac{e^{-r(T-t)} S_T}{S_t}. \tag{11.57}$$

By Girsanov theorem, the density process associated to $\eta_t(T)$ in Eq. (11.57) satisfies

$$\frac{d\eta_t(\tau)}{\eta_t(\tau)} = -\left(-\sigma_\tau\right) dW_{1\tau},$$

such that the stock price is solution to:

$$\left\{\begin{array}{lcl} d\ln S_t &=& \left(r+\frac{1}{2}\sigma_t^2\right)dt+\sigma_t d\hat{W}_{1t} \\ d\sigma_t^2 &=& \left(\kappa\omega-\left(\kappa-\rho\xi\right)\sigma_t^2\right)dt+\xi\sigma_t\left(\rho d\hat{W}_{1t}+\sqrt{1-\rho^2}dW_{2t}\right)\end{array}\right. \tag{11.58}$$

and $\hat{W}_{1t}$ is a Brownian motion under the new probability with density $\hat{q}_t^{\mathrm{m}}$, with $dW_{1t}=d\hat{W}_{1t}+\sigma_t dt$.

Let $x\equiv\ln S$. In the Black-Scholes case, σ_t^2 is a constant, and the two probabilities, $\hat{Q}_t\left(x_T\geq\ln K\right)$ and $Q_t\left(x_T\geq\ln K\right)$, can be expressed in closed-form, using Eq. (11.58) and Eq. (11.55), respectively, leading to the celebrated formula in Eq. (11.15), as explained in Section 11.4.2.

In the Heston's model, the two probabilities, $P_1\left(x_t,\sigma_t^2,t\right)\equiv\hat{Q}_t\left(x_T\geq\ln K\right)$ and $P_2\left(x_t,\sigma_t^2,t\right)\equiv Q_t\left(x_T\geq\ln K\right)$, are solutions to:

$$\hat{L}P_1\left(x,\sigma^2,t\right)=0,\quad LP_2\left(x,\sigma^2,t\right)=0, \tag{11.59}$$

with the same boundary condition $P_j\left(x,\sigma^2,T\right)=\mathbb{I}_{x\geq\ln K}$, $j=1,2$, and where $\hat{L}$ and L are the infinitesimal generators associated to Eq. (11.58) and Eq. (11.55). Indeed, we have that: (i) probabilities are martingales, due to $P_2\left(x_t,\sigma_t^2,T-t\right)=\mathbb{E}_t\left(\mathbb{I}_{x_T\geq\ln K}\right)=\Pr\left(x\geq\ln K\right)$, and similarly for $P_1\left(\cdot\right)$; and (ii) probabilities are diffusion processes because the stock price is. Eq. (11.59) then follows by the Feynman-Kac representation theorem.

The solution to the two partial differential equations is unknown in closed-form, though. However, their characteristic functions can be shown to be exponential affine in x and σ^2. Precisely, define the two characteristic functions:

$$f_1\left(x,\sigma^2,t;\phi\right)=\hat{\mathbb{E}}_t\left(e^{-\mathrm{i}\phi x_T}\right),\quad f_2\left(x,\sigma^2,t;\phi\right)=\mathbb{E}_t\left(e^{-\mathrm{i}\phi x_T}\right),\quad \mathrm{i}=\sqrt{-1},$$

where $\hat{\mathbb{E}}_t$ denotes the expectation taken with respect to $\hat{q}_t^{\mathrm{m}}$, and $\mathbb{E}_t$ denotes the conditional expectation taken against q_t^{m}. The two functions f_j satisfy the same partial differential equations (11.59), but they can be solved in closed-form, because their boundary conditions are simply $f_j\left(x,\sigma^2,T\right)=e^{-\mathrm{i}\phi x}$. Indeed, a fundamental definition is that a model is affine if its characteristic function is exponential-affine in its state variables. Affine models were already used to analyze the term structure of interest rates, since at least Vasicek (1977) and Cox, Ingersoll and Ross (1985), as we shall discuss in Chapter 13. Heston's model is the option pricing counterpart to these models of the yield curve.

The solution to the two characteristic functions is given by:

$$f_j\left(x,\sigma^2,t;\phi\right)=e^{C_j(T-t;\phi)+D_j(T-t;\phi)\sigma^2+\mathrm{i}\phi x},$$

where

$$\begin{array}{rcl} C_j\left(T-t;\phi\right) &=& r\phi\mathrm{i}\left(T-t\right)+\frac{\kappa\omega}{\xi^2}\left[\left(b_j-\rho\xi\phi\mathrm{i}+d_j\right)\left(T-t\right)-2\ln\left(\frac{1-g_je^{d_j(T-t)}}{1-g_j}\right)\right] \\ D_j\left(T-t;\phi\right) &=& \frac{b_j-\rho\xi\phi\mathrm{i}+d_j}{\xi^2}\left(\frac{1-e^{d_j(T-t)}}{1-g_je^{d_j(T-t)}}\right) \\ g_j &=& \frac{b_j-\rho\xi\phi\mathrm{i}+d_j}{b_j-\rho\xi\phi\mathrm{i}-d_j},\quad d_j=\sqrt{\left(b_j-\rho\xi\phi\mathrm{i}\right)^2-\xi^2\left(2u_j\phi\mathrm{i}-\phi^2\right)} \\ b_1 &=& \kappa-\xi\rho,\quad b_2=\kappa,\quad u_1=\frac{1}{2},\quad u_2=-\frac{1}{2}\end{array}$$

such that,

$$P_j\left(x_t, \sigma_t^2, t\right) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \text{Re}\left[\frac{e^{-\text{i}\phi \ln K} f_j\left(x_t, \sigma_t^2, t; \phi\right)}{\text{i}\phi}\right] d\phi. \tag{11.60}$$

[Write a small technical Appendix on inversions of characteristic functions] Replacing these two probabilities into Eq. (11.56), yields the celebrated Heston's formula.

11.5.3.5 Case study: assessing accumulators in markets with stochastic volatility

Is stochastic volatility important, in practice? Consider the "accumulators" described in Section 11.3.2, which are portfolios short n puts having the same strike K_p, and long one call with strike K_C, such that the initial cost of the portfolio is equal to zero. Let $K_C = 100$, and assume all options expire in six months. We consider two markets. The first market is one where asset prices are generated by the Black & Scholes model. In the second market, asset prices follow the Heston's model in Eq. (11.55). We assume that the parameter values are the same as those in Figure 11.7. We also assume that random fluctuations in volatility are not priced, such that the parameters κ and ω under the physical probability are the same as those under the risk-neutral. We address a number of issues aiming at exploring how stochastic volatility affects risk and return of these accumulators.

First, assume the current level of the index is $S = 98$. How many puts with strike $K_p = 90$ should we sell, to finance the long position in the call option, in both the Black & Scholes and Heston's markets? Assuming that the current volatility in Heston's model equals that related to the steady state variance, $\sigma_\tau = \sqrt{\omega} = \sqrt{0.01}$, we find that in the Heston's market, $n = 3.9399$, whereas in Black & Scholes, $n = 6.5362$.

Next, we assume the current level of the index is $S = 102$ and re-do the previous calculations assuming, now, that the strike of the put is $K_p = 95$. We find that in Heston's market, $n = 5.9338$, whereas in Black & Scholes, $n = 8.8767$. The interpretation is that by increasing the strike from $K_p = 90$ to $K_p = 95$, we move towards more expensive puts. At the same time, by assuming that $S = 102$ (instead of $S = 98$), we are also considering a better market, which has an opposite effect on the put price than that arising following an increase in K_p. Moreover, a better market obviously makes call options more expensive. These effects lead to a higher number of puts to sell, in order to finance a long position in the call.

We can assess how the portfolio changes, once we consider a value of the index still equal to $S = 98$, but two levels of volatility: one higher than that corresponding to the steady state, $\sigma_t = \sqrt{0.02}$, and another, lower, $\sigma_t = \sqrt{0.008}$. In the high volatility case, $\sigma_t = \sqrt{0.02}$, we find that $n = 2.9252$, and in the low volatility case, we find that $\sigma_t = \sqrt{0.008}$, $n = 4.3249$. An increase in volatility makes both puts and calls increase in value, although the price of the two puts increases more than that of the call.

Finally, assume that the rate of appreciation μ of the asset price is constant under the physical probability, and equal to $\mu = 3\%$. We can use Montecarlo simulations to calculate the average and standard deviation of the Profits & Losses generated by the accumulator. We summarize our findings in the table below.

P&Ls for a six month accumulator

	average	std dev
B&S	1.3716	6.5114
Heston	1.2925	7.0352
Heston (high vol)	1.3571	8.9136
Heston (low vol)	1.2700	6.5906

Figure 11.8 depicts the relative frequency of the P&Ls corresponding to all cases we consider.

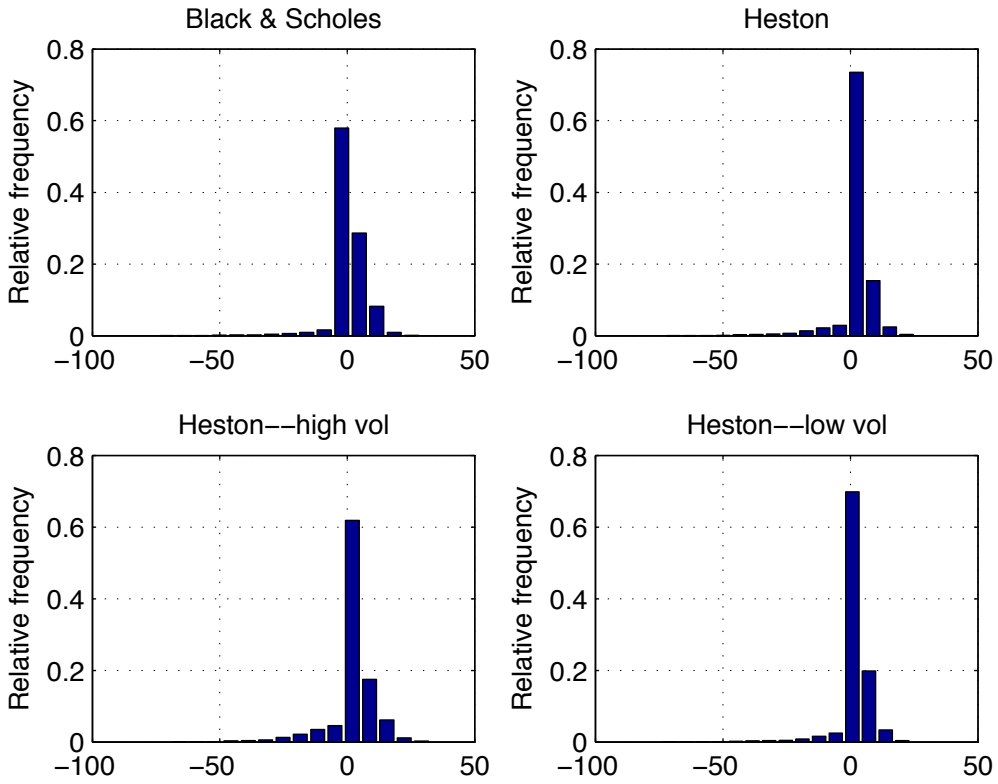


FIGURE 11.8. Frequency of Profit and Losses regarding a six month "accumulator" under different market assumptions: Black & Scholes (NW quadrant), Heston's with current volatility fixed at its long-term value (NE), Heston's with current volatility lower than its long-term value (SE), and Heston's with current volatility higher than its long-term value (SW).

Note that although the average P&Ls are positive in all cases, the frequency distributions of the P&Ls exhibit quite high standard deviations, with long left-tails—downside risk is quite substantial, consistently with the payoff structure that we depicted in Figure 11.4. In the Heston's market, average profits are higher than in Black & Scholes because accumulators necessitate less puts. Within the Heston's market, profits lower when we move to both a low volatility and high volatility scenarios. In the low volatility scenario, average profits lower because the number of puts in the accumulator needs to increase. In the high volatility scenario, profits lower because whilst the number of puts in the accumulator decreases, a more volatile market also makes the accumulator more likely to generate adverse outcomes, thereby leveling down the expected profits.

11.6 Trading volatility with options

11.6.1 Payoffs

Positioning in volatility is a trading strategy relying on expectations about future volatility movements. It is a non-directional strategy, as it does not rely on the direction of the markets, only on their changes. It is market practice to distinguish between two types of volatility trading:

(i) *Vega trading*, or *volatility surface* trading. It refers to a trade aiming to profit from a view that the term-structure of implied volatilities will change—for example, from the expectation of a flattening or a steepening term-structure of implied volatilities. It requires positioning into multiple types of options according to the nature of the expectation. For example, a "bull flattener" relies on the expectation that long-term implied volatilities will decrease faster than short-term implied volatilities, leading the term-structure of implied volatilities to flatten, which could be implemented through a portfolio which is: (i.1) long short-term options, and (i.2) short the long-term ones. (This portfolio would need to be delta-hedged, for reasons explained below).

(ii) *Gamma trading*. It is a trade that aims to generate profits from a realized volatility exceeding the current implied volatility. It relies on directional views regarding ongoing volatility developments. For this reason, gamma trading has an horizon that is typically shorter than that of vega trading.

Option-based strategies might allow us to have views about these volatility developments, and include trading "straddles," "strangles," "butterflies," "calendars," or delta-hedged option positions, as we shall explain below. They consist of portfolios comprising options and assets underlying these options, and aim to make P&Ls consistent with views about volatility developments.

A natural question arises. We know option prices are, generally, increasing in volatility. So why do we need to create portfolios of options and underlyings, in order to trade volatility? The reason is that option prices are increasing in both volatility and the asset price. For example, in a stochastic volatility setting, the option price is $C\left(S_t, \sigma_t^2, t\right)$, and if the volatility σ_t increases, the option price $C\left(S_t, \sigma_t^2, t\right)$ increases as well, in general. However, it might be possible that the increase in volatility occurs exactly when the asset price decreases. Incidentally, this circumstance is quite likely to occur, given the empirical evidence about the negative correlation between σ_t and S_t reviewed in Section 11.5. The implication would be that the increase in C determined by an increase in σ_t might be offset by the fall in C following the drop in S_t. To isolate movements in the asset price volatility, we need to consider portfolios reverse-engineered so as to be insensitive to changes in the underlying asset price. [Mention here and in the next section Goldman Sachs approach to VIX]

To mitigate the effects of the movements in the underlying price, we may consider Black-Scholes hedges, such that the long position in the call option is offset by the short-position in the "Black-Scholes replicating" portfolio—which, by construction, only neutralizes movements in S_t, not σ_t. An alternative is a portfolio comprising options with final payoffs driven by the stock price, and negatively correlated, such as a European put and call options. For example, a *straddle* is a portfolio of one call option and one put option that have the same strike price and the same maturity. (A *strangle* is the same as a straddle, with the difference that the strike of the call differs from that of the put.)

Figure 11.9 depicts an example of payoffs arising from going long a straddle. The left panel shows the final payoff, equal to $(K - S_T)^+ + (S_T - K)^+$, for $K = 100$, as well as the value of this payoff at t, assuming a Black and Scholes market, with a risk-free rate $r = 1\%$, instantaneous volatility $\sigma = 20\%$, and under two assumptions about the maturity of the straddle, three months and one month. The right panel shows, instead, the P&L of this straddle, defined as $(K - S_T)^+ + (S_T - K)^+ - e^{r(T_i - t)}\text{Cost}_{i,t}$, where $\text{Cost}_{i,t} = p_{\text{BS}}(S_t, T_i - t; K, \sigma) + C_{\text{BS}}(S_t, T_i - t; K, \sigma)$, $C_{\text{BS}}(\cdot)$ is the Black-Scholes formula in Eq. (11.15), $p_{\text{BS}}(\cdot)$ is the corresponding put price, and $T_i - t$ is the maturity of the straddle, with $T_1 - t = \frac{1}{12}$ and $T_2 - t = \frac{3}{12}$. We assume that the index level at t is $S_t = 100$, such that the straddles are approximately at the money—the strike leading to at-the-money straddles is: $K_i = S_t e^{r(T_i - t)}$, but for comparison reasons, we keep on setting $K = 100$ for both straddles.

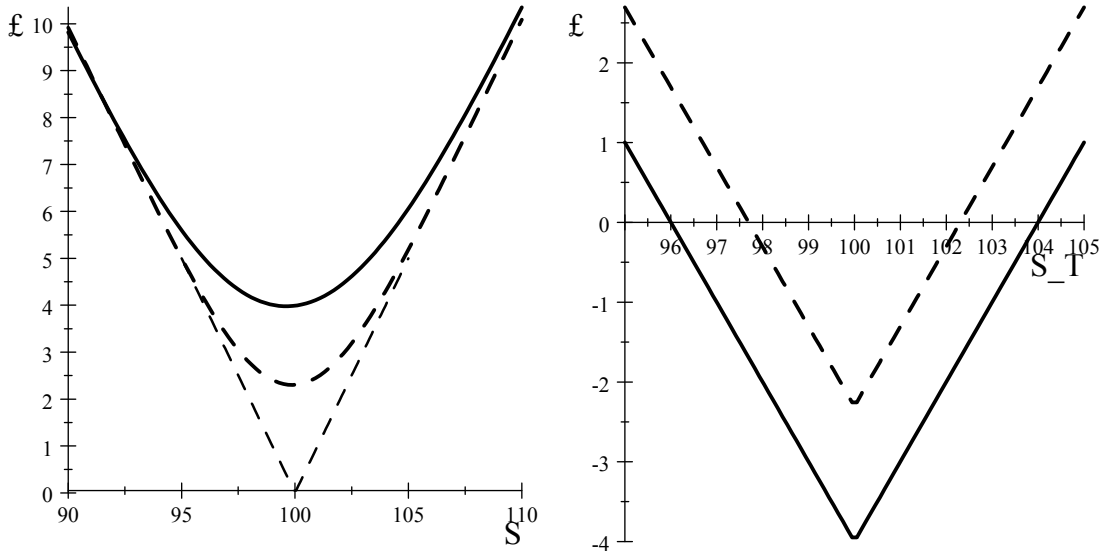


FIGURE 11.9. The left panel depicts the payoff of a "straddle" with strike price $K = 100$ (thin dashed line), as well as the value of a straddle with maturity $T_2 - t = \frac{3}{12}$ (solid line) and $T_1 - t = \frac{1}{12}$ (dashed line). The right panel depicts the P&Ls of roughly at-the-money straddles bought at time t, defined as $(K - S_T)^+ + (S_T - K)^+ - e^{r(T-t)}\text{Cost}_{i,t}$, where $\text{Cost}_{i,t} = p_{\text{BS}}(S_t, T_i - t; K, \sigma) + C_{\text{BS}}(S_t, T_i - t; K, \sigma)$, with $T_2 - t = \frac{3}{12}$ months (solid line), and $T_1 - t = \frac{1}{12}$ (dashed line), $r = 1\%$, $\sigma = 10\%$, and the index level $S_t = 100$.

The logic behind a straddle is that a call and a put have deltas that roughly compensate with each other, thereby allowing this portfolio to change primarily because of volatility movements. Figure 11.9 illustrates that a straddle helps express views about volatility, in that it pays off whenever the stock price moves sufficiently away from the initial level of the index, $S_t = 100$.

Note a technical complication, arising because the delta of the straddle is not precisely always zero, especially when the index level drifts away from moneyness. By Eq. (11.21), and the put-call parity in Eq. (11.7), it is: $2\Phi(d) - 1$. For example, we have that in the Black & Scholes market,

$$\text{Straddle} = C + P = S(2\Phi(d) - 1) - Ke^{-r\tau}\left(2\Phi\left(d - \sigma\sqrt{\tau}\right) - 1\right),$$

such that once we choose $K : S = Ke^{-r\tau}$ at inception, we have, *at inception*,

$$\begin{aligned}
\text{Straddle}_o^{atm} &\equiv 2S\left[\Phi\left(\frac{1}{2}\sigma\sqrt{\tau}\right) - \Phi\left(-\frac{1}{2}\sigma\sqrt{\tau}\right)\right] \\
&= 2S\left[2\Phi\left(\frac{1}{2}\sigma\sqrt{\tau}\right) - 1\right] \\
&\approx 2S\left[2\left(\Phi(0) + \phi(0)\frac{1}{2}\sigma\sqrt{\tau}\right) - 1\right] \\
&= 2S\frac{1}{\sqrt{2\pi}}\sigma\sqrt{\tau},
\end{aligned}$$

where the third line follows by a Taylor's linear approximation for $\sigma\sqrt{\tau}$ small, and the fourth by a simple calculation. Figure 11.10 depicts the delta of the straddle as a function of the index, *after inception*.[11]

The reason it deviates from zero while the index is significantly away from its initial level is that the straddle becomes a short or a long position as soon as the index moves. When the index is up, the delta of the call is higher than the delta of the put, in absolute value, and when the index is down, the opposite happens.

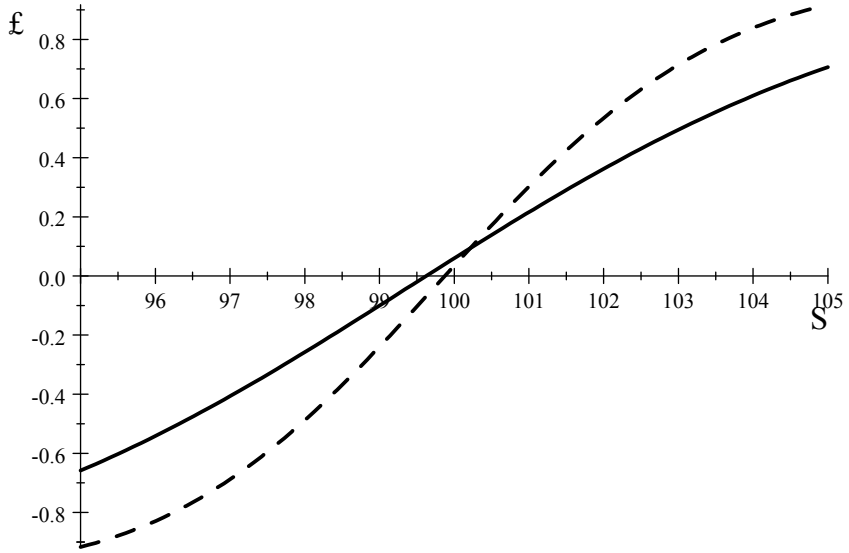


FIGURE 11.10. The delta of a straddle, $2\Phi(d) - 1$, where $\Phi(d)$ is the Black and Scholes delta in Eq. (11.21), the strike $K = 100$, for maturity $T_2 - t = \frac{3}{12}$ (solid line) and $T_1 - t = \frac{1}{12}$ (dashed line), and $r = 1\%$ and $\sigma = 10\%$.

Naturally, shorting a straddle leads to payoffs with opposite sign than those in Figure 11.9—shorting a straddle relies on the expectation that markets are going to be stable. Straddles bear some inglorious history. In 1995, the 233-year old Barings Bank collapsed, because of the famous short-straddle one of its traders, Nick Leeson, was implementing on the Nikkei Index. A short-straddle is, of course, a view volatility will not raise. However, in January 1995, a violent

[11] Note that by standard homogeneity properties, $\frac{\partial}{\partial S}\text{Straddle} = 2\Phi(d) - 1$, and that at inception, and for an (approximately) at-the-money straddle, $\frac{\partial}{\partial S}\text{ Straddle}|_{Ke^{-r\tau}=S} \approx \frac{1}{\sqrt{2\pi}}\sigma\sqrt{\tau}$.

earthquake made the Nikkei index crash by almost 7% in a week. The straddle was "naked," i.e. delta-hedged, at most, and led to losses Leeson was not only unable to absorb, but also to amplify, given he was insisting on having views the Index would stabilize. The Index did not.

Potential losses arising from a short position in straddles can be reduced, by going long one additional portfolio comprising: (i) an out-of-the money put, which pays exactly when the underlying goes down, and (ii) an out-of-the money call, which pays when the underlying goes up. Combining this portfolio with a short-straddle leads to what is known as *butterfly spread*. Figure 11.11 depicts payoffs and P&L relating to a butterfly, where the straddle has strike $K = 100$ and maturity one month (as one of the straddles in Figure 11.9), and the strikes of the out-of-the-money call and put are $K_C = 102$ and $K_p = 98$. The right panel shows the P&L of the butterfly, defined as $-\left[(K - S_T)^+ + (S_T - K)^+\right] + (K_p - S_T)^+ + (S_T - K_C)^+ - e^{r(T-t)}\text{Cost}_t$, where $T - t = \frac{1}{12}$, and Cost_t is the value of the butterfly at time t in the same Black & Scholes market considered in Figures 11.9-11.10.

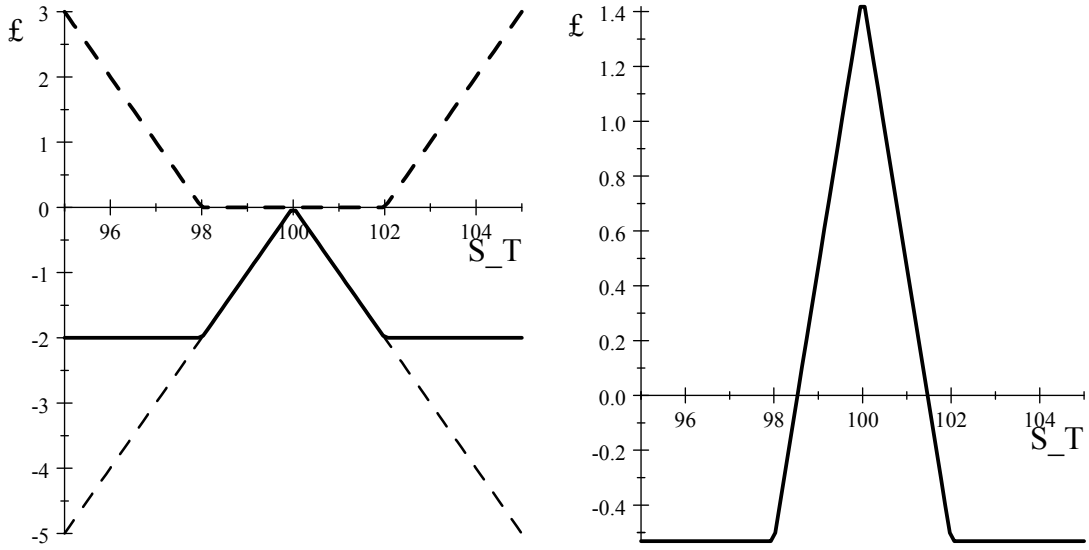


FIGURE 11.11. The left panel depicts the payoff of a "butterfly" with maturity equal to one month (solid line), which is (i) short one straddle (the thin dashed line) with strike $K = 100$, and (ii) long one out-of-the-money put, with strike $K_p = 98$, and one out-of-the-money put, with strike $K_C = 102$ (the dashed line). The right panel depicts the P&L of the butterfly, $-\left[(K - S_T)^+ + (S_T - K)^+\right] + (K_p - S_T)^+ + (S_T - K_C)^+ - e^{r(T-t)}\text{Cost}_t$, where $T - t = \frac{1}{12}$, and Cost_t is the value of the butterfly at time t, obtained in a Black & Scholes market with $r = 1\%$ and $\sigma = 10\%$ and the index level $S_t = 100$.

Calendar spreads are alternative strategies to straddles with which to express views about volatility. They are portfolios long one call with maturity $T_1 - t$ and short one call with maturity $T_2 - t$, where $T_1 < T_2$, and where the two calls have the same strike price. If the underlying asset price does not move too much, the calendar spread value drops, because the price decay due to the passage of time (see Section 11.4.6.2) is more severe for the call with lower time to maturity.

However, at any given point in time, the calendar spread increases in value as soon as the underlying price moves, regardless of whether this movement is positive or negative. This property is due to convexity. Let us explain. Note that as the option maturity decreases, the option

value becomes more convex with respect to the underlying price. Therefore, at any given point in time, when the underlying increases, the short-dated option value increases more than the long-dated; and when the underlying decreases, the short-dated option value decreases less than the long-dated.[12]

Therefore, going long a calendar is consistent with the view that the asset prices will quite fluctuate away from their current levels (regardless of whether on their way up or down), that is, that market (realized) volatility is about to increase. Figure 11.12 depicts the payoff and the P&Ls of a calendar, with strike $K = 100$, an index level $S_t = 100$, maturities $T_1 - t = \frac{1}{12}$ and $T_2 - t = \frac{3}{12}$, and assuming the same Black & Scholes market as in Figures 11.9-11.11. The payoff one month after the initial positioning is $(S_T - K)^+ - C_{\text{BS}}\left(S_T, \frac{2}{12}; K, \sigma\right)$, whereas the P&L is given by $(S_T - K)^+ - C_{\text{BS}}\left(S_T, \frac{2}{12}; K, \sigma\right) - e^{r\frac{1}{12}}\left(C_{\text{BS}}\left(S_t, T_1 - t; K, \sigma\right) - C_{\text{BS}}\left(S_t, T_2 - t; K, \sigma\right)\right)$.

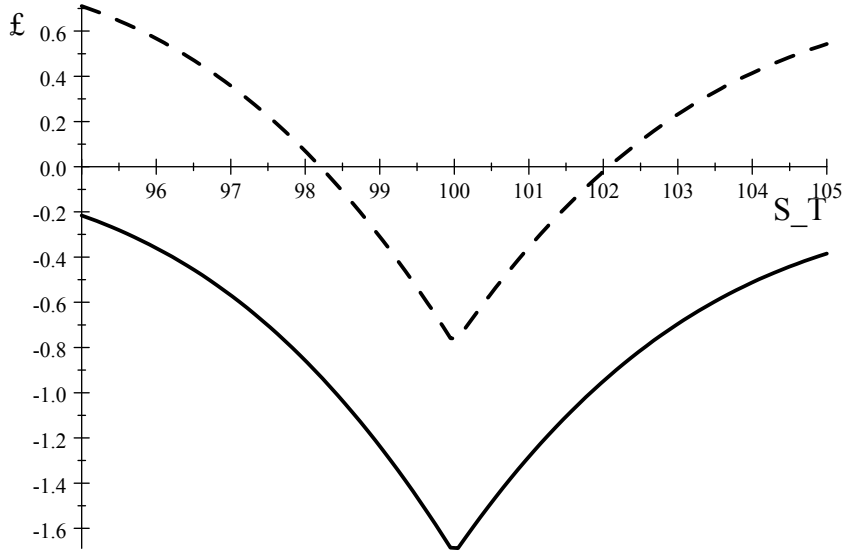


FIGURE 11.12. Payoff and P&L of a "calendar," a portfolio which is (i) long a call option with strike price $K = 100$ and time to maturity $T_1 - t = \frac{1}{12}$, and (ii) short a call option with strike price $K = 100$ and time to maturity $T_2 - t = \frac{3}{12}$. The solid line plots the payoff after one month from inception, defined as $\pi\left(S_T\right) \equiv (S_T - K)^+ - C_{\text{BS}}\left(S_T, \frac{2}{12}; K, \sigma\right)$, whereas the dashed line is the payoff inclusive of the initial value of the position, $\pi\left(S_T\right) - e^{r\frac{1}{12}}\left(C_{\text{BS}}\left(S_t, T_1 - t; K, \sigma\right) - C_{\text{BS}}\left(S_t, T_2 - t; K, \sigma\right)\right)$, with the index level $S_t = 100$.

11.6.2 P&Ls of Δ-hedged strategies

Straddles, calendars, or Black-Scholes hedged strategies, are not necessarily the best way to formulate views regarding ongoing volatility developments. To understand volatility trading through option-based strategies, consider the simplest strategy, where one buys an option and hedges it through the Black and Scholes formula.[13] Suppose to live in a world with stochastic volatility, where the asset price moves as in Eqs. (11.36). Assume that at time t, we go long a call option with market price equal to $C\left(S_t, \sigma_t^2, t\right)$. Let us build up a self-financed portfolio

[12] This argument is best understood while comparing the intrinsic value of an option to the option value before maturity. The former is flat for relatively small values of the underlying price. Otherwise, it increases one-to-one with underlying.

[13] The following arguments also apply to the hypothetical situation where an investment bank, say, purchases an option for a mere market making scope, and then tries to hedge against it through Black-Scholes. It is, however, an unrealistic situation, as investment banks hedge through books, not through the single units adding up to the books.

with value V_t,

$$V_t = a_t S_t + b_t B_t, \tag{11.61}$$

where $B_t = e^{rt}$ is the money market account,

$$V_0 = C\left(S_0, \sigma_0^2, 0\right), \quad a_t = \Delta_{\text{BS}}\left(S_t, t; \text{IV}_0\right), \tag{11.62}$$

and IV_0 is the Black-Scholes implied volatility as of time $t = 0$, i.e. the time at which we are to take a view on future volatility.

Consider, first, the following heuristic arguments. Assume a Black & Scholes market, where the short-term rate, r, is zero and μ is also zero. While volatility is constant in this market, there might be periods where the realized instantaneous variance, $\left(\frac{\Delta S_t}{S_t}\right)^2$, is higher than IV_0^2. What is the P&L of a call option delta-hedged through Black & Scholes? Note that a call option delta-hedged with Black & Scholes is simply a portfolio with value equal to $\Pi_t = C_t - \Delta_{\text{BS}} S_t - b_t B_t$, such that, approximately,

$$\begin{aligned}\Delta\Pi_t = & \underbrace{\Theta_t \Delta t + \frac{1}{2}\Gamma_t \left(\Delta S_t\right)^2 + \Delta_{\text{BS}} \Delta S_t}_{=\Delta C_t} - \Delta_{\text{BS}} \Delta S_t \\ = & \left(-\frac{1}{2}\Gamma_t S_t^2 \text{IV}_0^2\right)\Delta t + \frac{1}{2}\Gamma_t \left(\Delta S_t\right)^2 = \frac{1}{2}\Gamma_t S_t^2 \left[\left(\frac{\Delta S_t}{S_t}\right)^2 - \text{IV}_0^2 \Delta t\right],\end{aligned} \tag{11.63}$$

where $\Theta = \frac{\partial C}{\partial t}$, $\Delta_{\text{BS}} = \frac{\partial C}{\partial S}$, the Delta, $\Gamma = \frac{\partial^2 C}{\partial S^2}$, the Gamma, and the second equality follows by the Black-Scholes pricing equation (11.14). Aggregating up to the maturity of the option, delivers the P&L at T:

$$\text{P\&L}_T \equiv \sum_{t=1}^{T} \Delta\Pi_t = \frac{1}{2}\sum_{t=1}^{T} \Gamma_t S_t^2 \left[\left(\frac{\Delta S_t}{S_t}\right)^2 - \text{IV}_0^2 \Delta t\right]. \tag{11.64}$$

Note that the Black & Scholes delta is needed to compensate for those portions of the call price movements arising due to the asset price movements—the term $\Delta_{\text{BS}} \Delta S_t$ in the brackets of Eq. (11.63) contributes positively to the P&L, when $\Delta S_t > 0$, and negatively otherwise. So the hedging is natural: the call price can go up because of realized volatility (i.e., ΔS_t^2) or because the underlying price goes up. To isolate the views about volatility, we need to hedge against movements in the underlying price. As is clear, hedging through the Black-Scholes delta helps neutralize this effect.[14]

Hedging can only be effective in the very short-term, as the period-by-period profits $\Delta\Pi_t$ in Eq. (11.63), only depend on how far realized volatility is from the initial Black-Scholes implied volatility.[15] In general, hedging might lead to a P&L inconsistent with views, because the difference between realized volatility and Black-Scholes implied volatility at time t, $\left(\frac{\Delta S_t}{S_t}\right)^2 - \text{IV}_0^2 \Delta t$, is weighted with the Dollar Gamma, $\Gamma_t S_t^2$, which is positive as the Black-Scholes price is convex in S. In other words, we may well end up with a negative $\text{P\&L}_T$, even if the terms

[14] Naturally, in the continuous time limit, and assuming Black & Scholes is true, $\left(\frac{\Delta S_t}{S_t}\right)^2 \longrightarrow \text{IV}_0^2 dt$, such that Eq. (11.64) shrinks to zero. Therefore, in the Black-Scholes market, views on realized volatility can be channelled only due to discrete-rebalancing. Below, we shall explain that a non-zero P&L obtains even in the continuous time limit, once the stock price displays stochastic volatility.

[15] A similar P&L obtains regarding straddles and calendars.

$\left(\frac{\Delta S_t}{S_t}\right)^2 - \mathrm{IV}_0^2 \Delta t$ are positive for most of the time, a feature known as "price-dependency." We now illustrate these facts through a continuous-time model with random volatility.

Consider a general situation where volatility is not constant, such that the model is misspecified. El Karoui, Jeanblanc-Picqué and Shreve (1998) make the following observations. Consider the value of the self-financed portfolio in Eq (11.61). Because this portfolio is self-financed,

$$\begin{aligned} dV_t &= a_t dS_t + r b_t B_t dt \\ &= r V_t dt + a_t \left(dS_t - r S_t dt\right) \\ &= \left[r V_t + a_t \left(\mu - r\right) S_t\right] dt + a_t \sigma_t S_t dW_t. \end{aligned}$$

Moreover, by Itô's lemma,

$$dC_{\mathrm{BS}}\left(S_t, T-t; K, \mathrm{IV}_0\right) = \left(\frac{\partial C_{\mathrm{BS}}}{\partial t} + \mu S_t \frac{\partial C_{\mathrm{BS}}}{\partial S} + \frac{1}{2}\sigma_t^2 S_t^2 \frac{\partial^2 C_{\mathrm{BS}}}{\partial S^2}\right) dt + \sigma_t S_t \frac{\partial C_{\mathrm{BS}}}{\partial S} dW_t,$$

where

$$\begin{aligned} &\frac{\partial C_{\mathrm{BS}}}{\partial t} + \mu S_t \frac{\partial C_{\mathrm{BS}}}{\partial S} + \frac{1}{2}\sigma_t^2 S_t^2 \frac{\partial^2 C_{\mathrm{BS}}}{\partial S^2} \\ &= \underbrace{\frac{\partial C_{\mathrm{BS}}}{\partial t} + r S_t \frac{\partial C_{\mathrm{BS}}}{\partial S} + \mathrm{IV}_0^2 S_t^2 \frac{\partial^2 C_{\mathrm{BS}}}{\partial S^2}}_{\equiv\, r C_{\mathrm{BS}}} + \left(\mu - r\right) S_t \frac{\partial C_{\mathrm{BS}}}{\partial S} + \frac{1}{2}\left(\sigma_t^2 - \mathrm{IV}_0^2\right) S_t^2 \frac{\partial^2 C_{\mathrm{BS}}}{\partial S^2} \\ &= r C_{\mathrm{BS}} + \left(\mu - r\right) S_t \frac{\partial C_{\mathrm{BS}}}{\partial S} + \frac{1}{2}\left(\sigma_t^2 - \mathrm{IV}_0^2\right) S_t^2 \frac{\partial^2 C_{\mathrm{BS}}}{\partial S^2}. \end{aligned}$$

Therefore, the tracking error, or P&L$_t$, defined as the difference between the Black-Scholes price and the portfolio value,

$$\mathrm{P\&L}_t \equiv C_{\mathrm{BS}}\left(S_t, T-t; K, \mathrm{IV}_0\right) - V_t,$$

satisfies

$$d\mathrm{P\&L}_t = \left(r \mathrm{P\&L}_t + \frac{1}{2}\left(\sigma_t^2 - \mathrm{IV}_0^2\right) S_t^2 \frac{\partial^2 C_{\mathrm{BS}}}{\partial S^2}\right) dt.$$

At maturity T,

$$\begin{aligned} \mathrm{P\&L}_T &\equiv C_{\mathrm{BS}}\left(S_T, T-T; K, \mathrm{IV}_0\right) - V_T \\ &= \max\left\{S_T - K, 0\right\} - V_T \\ &= \frac{1}{2} e^{rT} \int_0^T e^{-rt} \left(\sigma_t^2 - \mathrm{IV}_0^2\right) \frac{\partial^2 C_{\mathrm{BS}}}{\partial S^2} S_t^2 dt. \end{aligned} \tag{11.65}$$

This expression is the continuous-time counterpart to Eq. (11.64) in a market with stochastic volatility. Moreover, it can be shown that a delta-hedged straddle strategy leads to twice the expression in Eq. (11.65), with the second partial of the straddle replacing the Black-Scholes Gamma. Because the Black-Scholes price is convex, Eq. (11.65) tells us that even if we do not exactly know the law of movement of volatility, but still hold the view it will be persistently higher than the initial Black-Scholes implied volatility, we can obtain positive profits through (i) a long position in a call, and (ii) a short position in the Black-Scholes replicating portfolio. Naturally, it isn't an arbitrage opportunity. The critical assumption is that volatility will increase.

Eq. (11.65) is problematic. Even if the volatility σ_t is higher than IV_0 for most of the time, the final P&L may not necessarily lead to a profit. The reason is that each volatility view, $\sigma_t^2 - \text{IV}_0^2$, is weighted by the Dollar Gamma, $S_t^2 \frac{\partial^2 C_{\text{BS}}}{\partial S^2}$. It may be that "bad" realization of the volatility views, i.e. $\sigma_t^2 < \text{IV}_0^2$, occur precisely when the Dollar Gamma is large—the "price-dependency" issue raised whilst discussing Eq. (11.64). Moreover, the strategy is costly, as it relies on Δ-hedging. The volatility contracts of Section 11.6 overcome these difficulties.

11.7 Local volatility

This section generalizes the previous models with stochastic volatility, which allow to price the entire cross-section of options (the "skew") without errors. Section 11.8 explains that there exists a beautiful connection between the model in this section and the *price* of volatility, i.e., the price to be paid in dedicated "variance swaps" aiming to protect investors from changes in volatility over a fixed horizon.

11.7.1 Issues

Stochastic volatility models might provide interesting explanations, such as the smile effect, as discussed in Section 11.5.2. However, these models cannot allow for a *perfect* fit of the smile. Towards the end of 1980s and the beginning of the 1990s, a modeling approach emerged to cope with issues relating to a perfect fit of the yield curve. As reviewed in the next two chapters, this modeling approach was a response to the need of pricing interest rate derivatives while relying on models in which the underlying assets in the banks books (bonds say) were priced without errors. In the early and mid 1990s, methods were developed to deal with equity options (Derman and Kani, 1994; Dupire, 1994; Derman, 1998), which we succinctly review in this section.

Why is it important to exactly fit all of the already existing options? Trading deals with both plain vanilla and less liquid, or "exotic" derivatives. Suppose we wish to price exotic derivatives, say a new product. We want to make sure the model we use to price the new product predicts that the plain vanilla option prices are identical to those we are trading. How can we trust a model that is not even able to pin down all outstanding contracts? Such a model could open to arbitrage opportunities to unscrupulous users. To achieve these tasks, we need to feed the model not only with information regarding the current price, but also with information linking to the entire collection of available option prices.

[Plan of this Section]

11.7.2 Implied binomial trees

We begin with the usual binomial tree of Cox, Ross and Rubinstein (1979), then deal with implied binomial trees. The idea underlying implied binomial trees is to enrich the basic Black & Scholes model with information regarding the very same asset we are pricing (i.e., the options). While Black-Scholes and Cox-Ross-Rubinstein feed the evaluation model with information only relating to the initial stock price, implied binomial trees also use information available through derivatives data, aiming to model developments in the stock price. The resulting model is one with time-varying volatility, in which each derivative price used to feed the model is fit without errors. The model can be used to price exotic derivatives, while by design able to price existing derivatives without errors.

11.7.2.1 Binomial trees

[Use material from D:\antonio\lectures\fixed_income_LSE_FM413\classes\simple_trees, note it's almost copy-paste]

11.7.2.2 Implied binomial trees

It is the discrete-time version of the local volatility model in continuous time. In discrete time, Derman and Kani (1994) and Rubinstein (1994), and in continuous time, Dupire (1994).

We assume that the short-term rate r is constant, or exogenously given. We shall rely on Arrow-Debreu securities, i.e., securities that pay off one unit of numéraire in a given state of the world, and zero otherwise (see Chapter 2 in Part I). Let $p_s(\tau+1)$ be the price of an Arrow-Debreu security that pays off £1 at time $\tau+1$ and in state s, such that at time τ, we have the following tree

$$\begin{array}{lll} & & \pounds 0 \\ (s,\tau): \quad e^{-r}(1-q_s(\tau)) & \nearrow & \\ & \searrow & (s,\tau+1):\pounds 1 \\ & \nearrow & \\ (s-1,\tau): \quad e^{-r}q_{s-1}(\tau) & \searrow & \\ & & \pounds 0 \\ \cdots & \cdots & \cdots \\ \text{time } \tau & & \text{time } \tau+1 \end{array}$$

In this tree, $q_s(\tau)$ is the risk-neutral probability of an upward movement in the stock price at time τ in state s. Displayed on the second row of the tree are the values of the Arrow-Debreu security that pays at $(s,\tau+1)$, both at (s,τ) and $(s-1,\tau)$. By explanations given in Chapter 2, we have that

$$p_s(\tau+1) = p_s(\tau)\,e^{-r}(1-q_s(\tau)) + p_{s-1}(\tau)\,e^{-r}q_{s-1}(\tau). \tag{11.66}$$

Eq. (11.66) is known as the forward equation for the set of all the Arrow-Debreu security prices. It can be solved recursively once we know $q_s(\tau)$—r is given, as mentioned. To illustrate, suppose that risk-neutral probabilities are constant and equal to $\bar{q}$, such that the solution to Eq. (11.66) is, simply, the discounted value of the risk-neutral probability that s upward movements occur over $\tau+1$ trials:

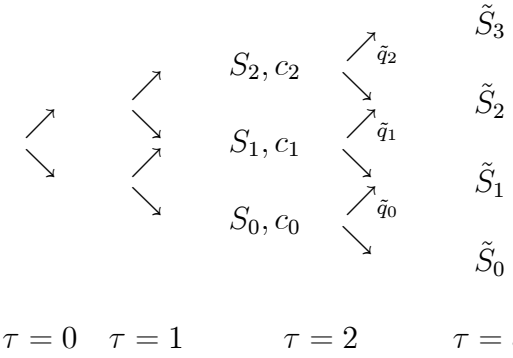
$$p_s(\tau+1) = e^{-r(\tau+1)}\binom{\tau+1}{s}\bar{q}^s(1-\bar{q})^{\tau+1-s}.$$

This example can be used to price Arrow-Debreu securities in the context of the Cox-Ross-Rubinstein model of the previous section.

We now turn to a more general problem. Suppose we are given a set of European option prices. What are the values of $q_s(\tau)$ to be assigned in each node of the tree such that these options are priced without error? More generally, what are the values of $q_s(\tau)$ and those of the stock price in each node of the tree such that all the European options are priced without error?

Our general approach in Chapter 12 is to freeze the risk-neutral probabilities while allowing the short-term rate r to be stochastic. That is, things are "coupled" in that the need arises to determine an implied binomial tree for the short-term rate that allows for a perfect fit of the

entire yield curve. In this chapter, we take r as given as explained, and determine an implied binomial tree for the risk-neutral probabilities. Suppose that we have solved everything up to time τ, for example, $\tau = 2$ as in the example depicted below.



We denote with $S_i \equiv S_i(2)$ the stock prices at time $\tau = 2$ and in node $i \in \{0, 1, 2\}$, and assume, as mentioned, that we have already determined them. We want to determine the values of the stock price at $\tau = 3$, i.e. $\tilde{S}_i \equiv S_i(3)$ in node $i \in \{0, 1, 2, 3\}$, *and* the risk-neutral probabilities at $\tau = 2$, i.e. $\tilde{q}_i \equiv q_i(2)$ in node $i \in \{0, 1, 2\}$. In total, we want to determine 7 parameters, $\tilde{S}_i$ and $\tilde{q}_i$.

In general, and for any τ, given a certain level of the tree at $\tau \equiv n - 1$, we want to determine $n + 1$ values of the stock price *and* n risk-neutral probabilities for a total of $2n + 1$ parameters. Denote with $C_{\$}(K, \tau + 1)$ the market price of a European call option expiring at time $\tau + 1$, and struck at K. We consider n options struck at the values of the stock at time τ. The price of each of these options is the average of its values one period before expiration, weighted by the Arrow-Debreu security prices for τ, viz

$$C_{\$}(S_i, \tau + 1) = \sum_{j=0}^{\tau} p_j(\tau)\, c_j(S_i), \quad i = 0, \cdots, \tau, \tag{11.67}$$

where $c_j(S_i)$ denotes the price at τ in node j, of the same option struck at S_i, and expiring at $\tau + 1$, which equals

$$c_j(S_i) = \begin{cases} e^{-r}\left(\tilde{q}_j(\tilde{S}_{j+1} - S_i) + (1 - \tilde{q}_j)(\tilde{S}_j - S_i)\right) & \text{if } j > i \\ e^{-r}\tilde{q}_j(\tilde{S}_{j+1} - S_i), & \text{if } j = i \\ 0 & \text{if } j < i \end{cases} \tag{11.68}$$

The strikes in Eq. (11.67) are chosen such that the options are (roughly) at-the-money one period before their expiration and, accordingly, the expressions in Eqs. (11.68) for the prices of the call options one period before the expiration follow, because by construction, $\tilde{S}_i \leq S_i \leq \tilde{S}_{i+1}$. While market prices for these options do not necessarily exists, we can interpolate the skew, and predict the missing points needed to implement Eq. (11.67).

To solve for the $2n + 1$ parameters, we use (i) the n pricing equations in (11.67), (ii) the n martingale conditions satisfied by the stock price at time τ,

$$S_i = e^{-r}\left(\tilde{q}_i \tilde{S}_{i+1} + (1 - \tilde{q}_i)\,\tilde{S}_i\right), \quad i = 0, \cdots, \tau, \tag{11.69}$$

and, finally, (iii) a $(2n + 1)$-th condition, a renormalization condition, which we shall discuss below.

Replacing Eq. (11.69) into the the first of Eqs. (11.68) leaves,

$$e^r c_j\left(S_i\right)=e^r S_j-S_i, \quad \text{if } j>i. \tag{11.70}$$

Therefore, by Eqs. (11.67), (11.68) and (11.70),

$$e^r C_{\$}\left(S_i,\tau+1\right)=p_i\left(\tau\right)e^r c_i\left(S_i\right)+\sum_{j=i+1}^{\tau}p_j\left(\tau\right)e^r c_j\left(S_i\right)=p_i\left(\tau\right)\tilde{q}_j(\tilde{S}_{i+1}-S_i)+\sum_{j=i+1}^{\tau}p_j\left(\tau\right)\left(e^r S_j-S_i\right). \tag{11.71}$$

Let us solve Eq. (11.69) for $\tilde{q}_i$,

$$\tilde{q}_i=\frac{e^r S_i-\tilde{S}_i}{\tilde{S}_{i+1}-\tilde{S}_i}. \tag{11.72}$$

Replacing Eq. (11.72) into Eq. (11.71) and rearraning terms, leaves a recursion for the stock price over the nodes at time $\tau+1$,

$$\tilde{S}_{i+1}=\frac{\tilde{S}_i\left(e^r C_{\$}\left(S_i,\tau+1\right)-\sum_{j=i+1}^{\tau}p_j\left(\tau\right)\left(e^r S_j-S_i\right)\right)-p_i\left(\tau\right)S_i\left(e^r S_i-S_i\right)}{\left(e^r C_{\$}\left(S_i,\tau+1\right)-\sum_{j=i+1}^{\tau}p_j\left(\tau\right)\left(e^r S_j-S_i\right)\right)-p_i\left(\tau\right)\left(e^r S_i-S_i\right)}. \tag{11.73}$$

Eq. (11.73) needs a re-normalization, as mentioned. [In progress] Once we solve for $\tilde{S}_{i+1}$, we can solve for $\tilde{q}_i$, using Eq. (11.72), then update the price of the Arrow-Debreu securities in Eq. (11.66), and solve everything recursively.

This algorithm might lead to negative risk-neutral probabilities, which implies arbitrage in which case we change the options to use. [In progress]

We now turn to the continuous time approach.

11.7.3 The perfect fit, in continuous time

We know that the only input to the option pricing problem is the instantaneous volatility of the underlying asset price. Which volatility should we use? At least, we know that option prices are a function of this volatility. The idea is to find a volatility function such that this very same volatility delivers back the prices of all the already traded options. It's an "inverse" problem. Let us outline the steps we need to price new derivatives, while avoiding any pricing errors regarding the existing ones:

(i) We take as given the prices of a set of actively traded European options. Let K and T be strikes and time-to-maturity of these liquid options. We aim to match model predictions to data, viz

$$C_{\$}\left(K,T\right)=C\left(K,T\right), \quad K,T \text{ varying}, \tag{11.74}$$

where $C_{\$}\left(K,T\right)$ and $C\left(K,T\right)$ denote, respectively, market option prices and model predictions. Note that we are assuming that a continuum of options exists, and that the prices of these options are differentiable as much as we need to make our problem well-defined; technically, we need that $\sigma_{\text{loc}}\left(K,T\right)$ in Eq. (11.76) below be well-defined. We ask: does it mathematically make sense to consider a diffusive model for the asset price, such that the initial option prices collection $C_{\$}\left(K,T\right)$ is predicted without errors, just as required by Eq. (11.74)?

(ii) The answer is in the affirmative. Consider a diffusion process for the stock price:

$$\frac{dS_t}{S_t} = rdt + \sigma\left(S_t, t\right) d\hat{W}_t, \tag{11.75}$$

where $\hat{W}$ is a Brownian motion under the risk-neutral probability. Then, Eq. (11.74) holds, after "calibratation" of the volatility function $\sigma\left(S_t, t\right)$, as explained next.

(iii) The Appendix shows that Eq. (11.74) holds if and only if $\sigma\left(S_t, t\right) = \sigma_{\text{loc}}\left(S_t, t\right)$, where

$$\sigma_{\text{loc}}\left(K, T\right) = \sqrt{2\frac{\dfrac{\partial C_{\$}\left(K,T\right)}{\partial T} + rK\dfrac{\partial C_{\$}\left(K,T\right)}{\partial K}}{K^2\dfrac{\partial^2 C_{\$}\left(K,T\right)}{\partial K^2}}}. \tag{11.76}$$

The function $\sigma_{\text{loc}}\left(S, t\right)$ is known as *local volatility*. Its square is the local variance, defined as the conditional expectation under Q of the instantaneous variance given the market level S at some future date τ,

$$\sigma^2_{\text{loc}}\left(S, \tau\right) = \mathbb{E}_t\left[\sigma^2\left(S_\tau, \tau\right)\middle|\, S_\tau = S\right] \tag{11.77}$$

where $\mathbb{E}_t\left[.\middle|.\right]$ is a time-t conditional expectation under the risk-neutral probability. All in all, local volatility is the function $\sigma_{\text{loc}}\left(S, t\right)$ in Eq. (11.76) such that the theoretical price generated by the model in Eq. (11.75) equals the market price of the initial options collection.

(iv) Finally, we can price illiquid options through numerical methods, for example through simulations. In the simulations, we use

$$\frac{dS_t}{S_t} = rdt + \sigma_{\text{loc}}\left(S_t, t\right) d\hat{W}_t.$$

Empirical evidence suggests that the local volatility surface $\sigma_{\text{loc}}\left(S, t\right)$ is typically decreasing in S for fixed t, a phenomenon known as the Black-Christie-Nelson leverage effect (see Section 9.10 in Chapter 9 and the previous Section 11.5.2). This empirical regularity may be relied upon while assuming from the outset that $\sigma(x, t) = x^\alpha f(t)$, for some function f and some constant $\alpha < 0$. The resulting momdel is known as CEV (Constant Elasticity of Variance) model.

One can also introduce models combining "local vols" with "stoch vol," as follows:

$$\begin{aligned} \frac{dS_t}{S_t} &= rdt + \sigma(S_t, t) \cdot v_t \cdot d\hat{W}_t \\ dv_t &= \phi(v_t)dt + \psi(v_t)d\hat{W}^v_t \end{aligned} \tag{11.78}$$

where $\hat{W}^v$ is another Brownian motion, and ϕ, ψ are some functions. The appendix shows that in this specific case, the initial set of all European options prices is pinned down by:

$$\tilde{\sigma}_{\text{loc}}\left(K, T\right) = \frac{\sigma_{\text{loc}}\left(K, T\right)}{\sqrt{\mathbb{E}\left(v^2_T\middle|\, S_T = K\right)}}, \tag{11.79}$$

where $\sigma_{\text{loc}}(K,T)$ is the same as in Eq. (11.76). For this model, we simulate

$$\begin{cases} \dfrac{dS_t}{S_t} = rdt + \tilde{\sigma}_{\text{loc}}(S_t,t)\cdot v_t \cdot d\hat{W}_t \\ dv_t = \phi(v_t)dt + \psi(v_t)d\hat{W}_t^v \end{cases}$$

Practitioners are also heavily rely on the so-called SABR model, which is parametric in nature. This model is dealt with in Chapter 13 (see Section 13.10).

A note on recalibration. [In progress] Local surfaces are obviously functions of the initial state where the calibration starts off. The calibration has to be re-performed all the time to reflect new information.

11.7.4 Relations with implied volatility

11.7.4.1 Implied volatility as expected local volatility

Section 11.5.5 provides the expression of the P&L relating to a long position in a call option, delta-hedged with Black and Scholes using an implied volatility fixed at an initial level IV_0,

$$\text{P\&L}_T \equiv C_{\text{BS}}(S_T,T;K,T,\text{IV}_0) - V_T = \frac{1}{2}e^{rT}\int_0^T e^{-rt}\left(\sigma_t^2 - \text{IV}_0^2\right)S_t^2\frac{\partial^2 C_{\text{BS}}}{\partial S^2}dt. \tag{11.80}$$

Naturally, we have that $e^{-rT}\mathbb{E}(\text{P\&L}_T) = 0 \Leftrightarrow C_{\text{BS}}(S_0,0;K,T,\text{IV}_0) = V_0 = C(S_0,\sigma_0^2,0)$, the true market price, consistently with Eq. (11.62), such that, setting the expression of the last equality of Eq. (11.80), and solving for IV_0, delivers:

$$\text{IV}_0^2 = \frac{\mathbb{E}\left[\int_0^T e^{-rt}S_t^2\frac{\partial^2 C_{\text{BS}}}{\partial S^2}\cdot\sigma_t^2 dt\right]}{\mathbb{E}\left[\int_0^T e^{-rt}S_t^2\frac{\partial^2 C_{\text{BS}}}{\partial S^2}dt\right]}.$$

Alternatively, we may consider another hedging positioning, suggested by Gatheral (2006, Chapter 3), where the delta-hedging is made through some fictitious time-varying instantaneous, but deterministic, volatility, equal to $\bar{\sigma}_t$, say, where

$$\bar{\sigma}_t^2 = \frac{1}{T-t}\int_t^T \nu_u du, \quad \bar{\sigma}_0^2 = \text{IV}_0^2, \tag{11.81}$$

for some deterministic v_t. In this case, the P&L would be similar to that in Eq. (11.80), with

$$\text{P\&L}_T = \frac{1}{2}e^{rT}\int_0^T e^{-rt}\left(\sigma_t^2 - \nu_t\right)S_t^2\frac{\partial^2 C_{\text{BS}}}{\partial S^2}dt.$$

Imposing the zero profit condition under Q, leaves:

$$\nu_t = \frac{\mathbb{E}\left(S_t^2\frac{\partial^2 C_{\text{BS}}}{\partial S^2}\sigma_t^2\right)}{\mathbb{E}\left(S_t^2\frac{\partial^2 C_{\text{BS}}}{\partial S^2}\right)} = \mathbb{E}_{Q^\Gamma}\left(\sigma_t^2\right), \tag{11.82}$$

where $\mathbb{E}_{Q^\Gamma}$ is the expectation taken under the probability Q^Γ, defined as,

$$\frac{dQ^\Gamma}{dQ} = \frac{S_t^2\frac{\partial^2 C_{\text{BS}}}{\partial S^2}}{\mathbb{E}\left(S_t^2\frac{\partial^2 C_{\text{BS}}}{\partial S^2}\right)}.$$

We term Q^Γ "Dollar Gamma" probability. By Eqs. (11.81) and (11.82),

$$\mathrm{IV}_0^2 = \frac{1}{T}\int_0^T \mathbb{E}_{Q^\Gamma}\left(\sigma_t^2\right) dt. \tag{11.83}$$

So implied vols *are* expectations of future realized vols, but only under the Dollar Gamma probability. Clearly, then, they cannot be used as the fair value of a variance swaps, unless we tilt the variance contract by a "random multiplier" coinciding with the Dollar Gamma.

We can elaborate on Eq. (11.83). We have:

$$\begin{aligned}
\mathbb{E}_{Q^\Gamma}\left(\sigma_t^2\right) &= \mathbb{E}\left[\mathbb{E}\left(\sigma_t^2 \frac{dQ^\Gamma}{dQ}\middle| S_t\right)\right] \\
&= \mathbb{E}\left[\frac{dQ^\Gamma}{dQ}\mathbb{E}\left(\sigma_t^2\middle| S_t\right)\right] \\
&= \mathbb{E}\left[\frac{dQ^\Gamma}{dQ}\sigma_{\mathrm{loc}}^2\left(S_t,t\right)\right] \\
&= \mathbb{E}_{Q^\Gamma}\left[\sigma_{\mathrm{loc}}^2\left(S_t,t\right)\right] \\
&= \int \sigma_{\mathrm{loc}}^2\left(S_t,t\right)\frac{S_t^2\frac{\partial^2 C_{\mathrm{BS}}}{\partial S^2}q\left(S_t\middle| S_0\right)}{\mathbb{E}\left(S_t^2\frac{\partial^2 C_{\mathrm{BS}}}{\partial S^2}\right)}dS_t \\
&\approx \sigma_{\mathrm{loc}}^2(\tilde{S}_t,t),
\end{aligned} \tag{11.84}$$

where the first equality follows by the law of iterated expectations, $q\left(S_t\middle| S_0\right)$ denotes the conditional density of S_t given S_0, and $\sigma_{\mathrm{loc}}^2\left(S_t,t\right)$ is the local variance, as defined in Section 11.6.2. Finally, $\tilde{S}_t$ is a deterministic, "most likely path" of S_t, after Gatheral (2006, Chapter 6), a sort of certainty equivalent for the local variance, for a fixed t. We also know that at $t=T$, $\partial^2 C_{\mathrm{BS}}/\partial S^2$ is Dirac's delta centered at K, such that we may safely condition $\tilde{S}_T = K$ and, then, view $\tilde{S}_t$ as a bridge starting from S_0 and ending at K. As a simple example, $\tilde{S}_t \approx S_0\left(K/S_0\right)^{t/T}$. As a second example, $\mathbb{E}\left(S_t\middle| S_T = K\right)$, which we may approximate assuming S_t is a Geometric Brownian motion with parameters r and σ, in which case $\tilde{S}_t \approx S_0 e^{rt}\left(\frac{K}{S_0 e^{rT}}\right)^{t/T} e^{\frac{1}{2}\sigma^2}$. Gatheral argues, with a numerical example, that these approximations are quite reasonable, at least for options with time to maturity less than a year.

Using the approximation in Eq. (11.84) delivers:

$$\mathrm{IV}_0^2 = \frac{1}{T}\int_0^T \sigma_{\mathrm{loc}}^2(\tilde{S}_t,t)dt. \tag{11.85}$$

Surfaces depend on the initial state, as mentioned in Section 11.6.2. "Sticky smiles" can roughly be defined as those where the skew does not depend on the initial state. Suppose a very simple example, where $\mathrm{IV}\left(t,T,K;S_t\right) = a - b\left(K+S_t\right) = a - b\left(\frac{K}{S_t}+1\right)S_t$. As S_t falls, the skews goes up, consistent with the leverage effect. This skew does not depend on the initial price S_0. We can generate this skew, by assuming the local variance does not depend on time, $\sigma_{\mathrm{loc}}^2\left(K,t\right) = a - b\left(K+S_0\right)$. Indeed, in light of Eq. (11.85), we would then have that $\mathrm{IV}^2\left(0,T,K;S_0\right) = \mathrm{IV}_0^2 = a - b\left(K+S_0\right)$ and then, $\mathrm{IV}\left(t,T,K;S_t\right) = a - b\left(K+S_t\right)$ for each t.

11.7.4.2 Local volatility as a function of implied volatility

[In progress]

11.8 The price of (equity) volatility

How much volatility do we expect to prevail in the future, after controlling for risk? The answer to this question has long been the volatility implied by at-the-money options. Eq. (11.85) seems to suggest this answer is correct. In fact, it is not. Expected volatility, adjusted for risk, is a certain weighted average of implied volatilities of a continuum of options, as explained below. It is not mere academic purism. Knowing expected volatility under the risk-neutral probability allows to trade assets with payoffs linked to future realized volatility, known as variance swaps. In fact, in September 2003, the Chicago Board Options Exchange (CBOE) changed its volatility index VIX to approximate the variance swap rate of the S&P 500 index return (for 30 days), as explained below. In March 2004, the CBOE launched the CBOE Future Exchange for trading futures on the new VIX; and options referenced to VIX futures are also available for trading.

There are compelling reasons explaining the interest investors may have in these contracts. Undeniably, one is the possibility to formulate views about developments in stock market volatility, without incurring into the price-dependency issues pointed out in Section 11.5.4. Passive funds managers might also find these contracts useful, as in times of high volatility, tracking errors widen and, then, index tracking performance deteriorates. Hedge funds might find this type of contracts attractive as well, as they invest in "relative value" strategies, attempting to profit from temporary price discrepancies. In times of high volatility, price discrepancies typically widen, and variance swaps help these institutions hedge against these events.

11.8.1 One introductory example: range-based volatility

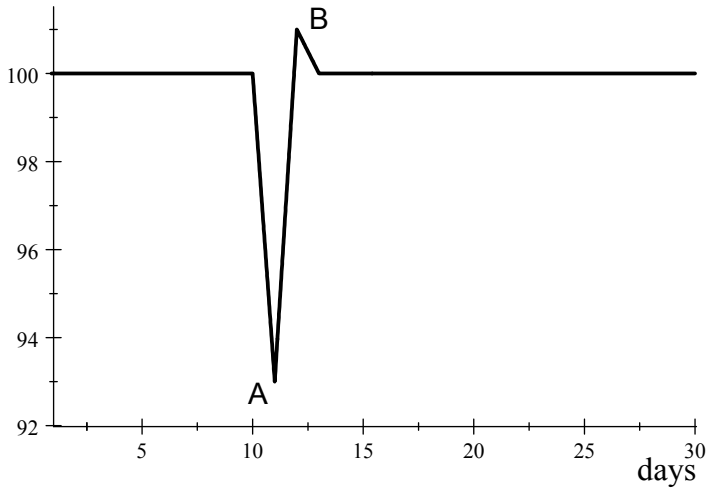
Straddles give rise to price-dependency, as explained in Section 11.5.4: even in the presence of a highly volatile market within a reference period, the straddle might not deliver a positive payoff, unless the underlying price moves beyond the straddles thresholds (see Figure 11.9). This introductory section develops an example of a simple payoff that aims to track changes in volatility.

Consider a derivative with payoff equal to the following estimate of stock volatility over a given time period $[t, T]$,

$$\pi(t,T) \equiv \max_{\tau\in[t,T]} (\ln S_\tau) - \min_{\tau\in[t,T]} (\ln S_\tau),$$

where S_τ denotes the stock price at τ.

This estimator of volatility relates to the Parkinson's (1980) "range estimator" of volatility, that is, one based on price ranges, defined as the difference between the highest and the lowest log price over a fixed sampling interval. This measure of realized volatility has drawbacks as a basis for volatility trading. Consider the following example. Suppose that for one month, the stock price never moves, except for three days (only then), when it goes down by 7% (10th day) (discretely compounded), then overshoots by 1% (11th day) and, finally, gets back to its initial value (12th day) (a "flash-crash"?). The payoff $\pi(t,T)$ is equal to $\ln \frac{B}{A} \approx 8.2\%$.



While this contract is clearly tracking volatility, it does so imperfectly. The stock price has not been volatile except for a few isolated days. We would like to make reference to contracts that pay off when volatility has been "sustained" over the whole month. The next section describes such contracts.

11.8.2 "Fear gauge" contracts

11.8.2.1 Variance swaps and the VIX index

Let us consider the following price process S_t under the risk-neutral probability:

$$\frac{dS_t}{S_t} = rdt + \sigma_t d\hat{W}_t, \tag{11.86}$$

where σ_t is $\mathcal{F}_t$-adapted, i.e. $\mathcal{F}_t$ can be larger than that generated by the stock price, $\mathcal{F}_t^S \equiv \sigma\left(S_\tau : \tau \leq t\right)$. Next, define the realized "integrated" variance within the time interval $[T_1, T_2]$, with $T_1 > t$:

$$\text{var}\left(T_1, T_2\right) \equiv \int_{T_1}^{T_2} \sigma_u^2 du.$$

We define a *variance swap* as a contracts with zero value at the inception date, t, and payoff at maturity T given by:

$$\pi_T^{\text{var}} = \left(\text{var}\left(t,T\right) - \mathbb{P}_{\text{var}}\left(t,T\right)\right) \times \text{N}, \tag{11.87}$$

where N is the notional value of the contract, and $\mathbb{P}_{\text{var}}\left(t,T\right)$ is the variance swap rate agreed at t, and paid off at time T.[16] Therefore, this contract is a forward, not a swap really. If r is deterministic, then the swap rate must satisfy:

$$\mathbb{P}_{\text{var}}\left(t,T\right) = \mathbb{E}_t\left(\text{var}\left(t,T\right)\right).$$

[16]Note that this contract relies on some notion of *realized* variance, as a continuous record of returns is obviously unavailable. Moreover, it has long been market practice to define the variance notional in such a way that $\text{N} = \frac{\text{VN}}{2\sqrt{\mathbb{P}_{\text{var}}}}$, where VN is the "vega notional," that is, the notional expressed in volatility percentage points. Suppose, for example, that realized volatility is 1 "vega" (i.e., one volatility point) above the square root of the variance swap rate, $\text{var}\left(t,T\right) = (\sqrt{\mathbb{P}_{\text{var}}} + 1)^2$, such that $\pi_T^{\text{var}} = (1 + \frac{1}{2\sqrt{\mathbb{P}_{\text{var}}}}) \times \text{VN} \approx \text{VN}$. That is, vega notional is approximately the notional for each vega realized volatility that exceeds the square root of the variance swap rate.

It remains to determine $\mathbb{E}_t\left(\text{var}\left(t,T\right)\right)$; below, we show that this is the same as a portfolio of out-of-the money options. First, let us consider a simple case, in which $r=0$. In this case, we can solve for $\mathbb{E}_t\left(\text{var}\left(t,T\right)\right)$ while relying on previous results in the previous section regarding local volatility. Indeed, note that by Eqs. (11.76) and (11.77), and the connection between risk-neutral densities and convexity of the option price, Eq. (11.28), we have that

$$\mathbb{E}_t\left(\sigma_T^2\right)=2\int_0^{\infty}\frac{\frac{\partial C_t(K,T)}{\partial T}+rK\frac{\partial C_t(K,T)}{\partial K}}{K^2}dK, \tag{11.88}$$

where $C_t\left(K,T\right)$ is the price as of time t of a call option expiring at T and struck at K. We can compute the risk-neutral expectation of this realized variance, under the assumption that $r=0$. By Eq. (11.88),

$$\mathbb{E}_t\left(\text{var}\left(T_1,T_2\right)\right)=2\int_0^{\infty}\frac{C_t\left(K,T_2\right)-C_t\left(K,T_1\right)}{K^2}dK. \tag{11.89}$$

Eq. (11.89) can be generalized to the case where $r>0$. In the Appendix, we show that for $T_1=t$, $T_2\equiv T$,

$$\mathbb{E}_t\left(\text{var}\left(t,T\right)\right)=2e^{r(T-t)}\left(\int_0^{F_t}\frac{P_t\left(K,T\right)}{K^2}dK+\int_{F_t}^{\infty}\frac{C_t\left(K,T\right)}{K^2}dK\right), \tag{11.90}$$

where F_t is the forward price: $F_t=e^{r(T-t)}S_t$, and $P_t\left(K,T\right)$ is the price as of time t of a put option expiring at T and struck at K. A proof of Eq. (11.90) is in the Appendix. It is a weighted average of out-of-the-money options, and it is a "fear gauge" as a result—the market assessment of extreme movements is aptly captured by an average of out-of-the-money options.

The new VIX index maintained by the CBOE is an estimate of the square root of $\mathbb{E}_t\left(\text{var}\left(t,T\right)\right)$ in Eq. (11.90), annualized:

$$\text{VIX}\left(t,T\right)\equiv\sqrt{\frac{1}{T-t}\mathbb{E}_t\left(\text{var}\left(t,T\right)\right)}, \tag{11.91}$$

where $T-t$ is expressed as a fraction of a year. The approximation relies on a finite number of out-of-the-money options.

Note an interesting point in this section. Up to the previous sections, we were used to think that volatility determines option prices (see, e.g., Section 11.5). We now have a theoretic construct that makes option prices sum up to a sufficient statistics for (market-adjusted, i.e., risk-neutral) *expected* volatility.

The VIX index is typically referred to as a "fear-gauge." Eqs. (11.90)-(11.91) illustrate that this definition is quite appropriate. The VIX index depends on those of the out-of-the money options, that is, the options that are exercized in case of tail events. Moreover, tail events are those arising when worst-case scenarios occur. This connection suggests an interpretation of VIX behavior and in terms of aversion to Knightian uncertainty as surveyed in Chapter 9, an issue not explored yet in theoretical research. It is well-known that the VIX index spikes exactly when equity markets drop. However, Eqs. (11.90)-(11.91) indicate the index potentially reflects market fears regarding both tail-events.

11.8.2.2 A crash derivation

The previous results are extraordinary. We know since Section 11.2 that for any generic risk $\tilde{X}_T$, its forward price is $\mathbb{E}(\tilde{X}_T)$, provided interest rates are constant, such that $\mathbb{E}(\tilde{X}_T) = e^{r(T-t)}X_t$ as soon as X is traded, for otherwise we would need to rely on a model to determine the expectation $\mathbb{E}(\tilde{X}_T)$. Note that here, $\text{var}\,(t,T)$ is not traded, and yet we can still express its price, $\mathbb{E}_t\left[\text{var}\,(t,T)\right]$, without relying on any model. In other words, we can price volatility in a model-free fashion—the annualized price of a variance swap is simply the square of the VIX index.

It is useful to review the steps that lead to Eq. (11.90). The starting point is the so-called log-contract, a concept first introduced by Neuberger (1994). This idealized contract is designed as to ensure a a payoff equal to $\ln \frac{F_T}{F_t}$. Its fair value is obviously $\mathbb{E}_t\left(\ln \frac{F_T}{F_t}\right)$, and is negative, as we shall shaw in a moment.

The intuitive reason the price of the log-contract is negative is that the payoff $\ln \frac{F_T}{F_t}$ is skewed to the left, due to the concavity of the log function, such that the expected losses on the downside are larger than the expected gains on the upside. These facts are confirmed by an application to Itô's lemma to Eq. (11.86), which yields:

$$\mathbb{E}_t\left(\ln \frac{F_T}{F_t}\right) = -\frac{1}{2}\mathbb{E}_t\left[\text{var}\,(t,T)\right]. \tag{11.92}$$

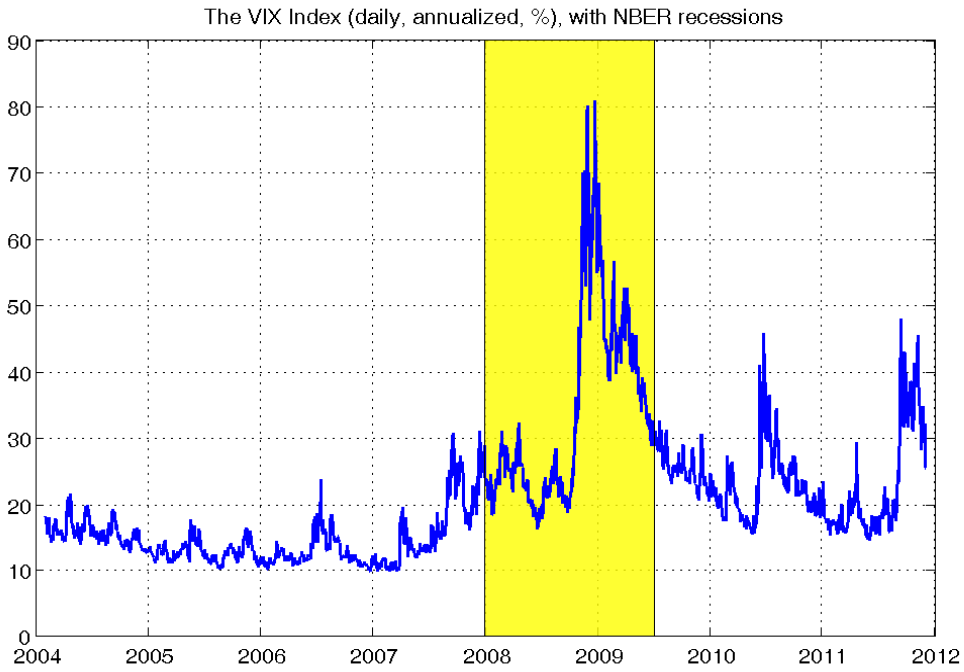
It is remarkable. The price of volatility is the same as the value of going short a log-contract. Naturally, we are not done, because we still don't know how to price the log-contract in the first place! The pricing of this, and related contracts, relies on so-called "spanning arguments," as explained for example by Bakshi and Madan (2000) and Carr and Madan (2001). In the Appendix, it is shown that the payoff of the log-contract can be written as

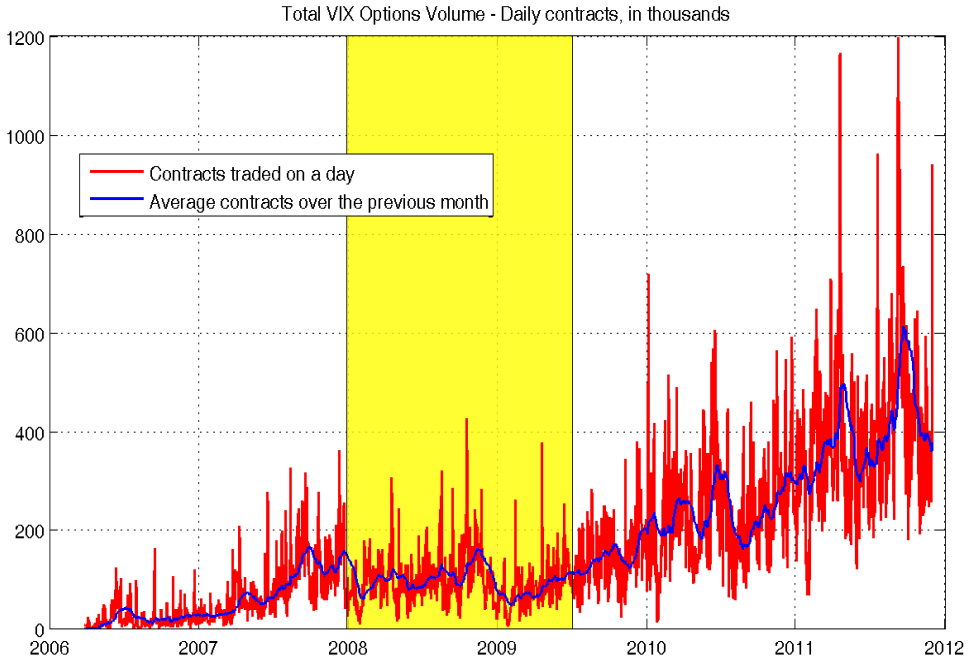
$$\ln \frac{F_T}{F_t} = \frac{1}{F_t}\left(F_T - F_t\right) - \left(\int_0^{F_t} \left(K - S_T\right)^+ \frac{1}{K^2} dK + \int_{F_t}^{\infty} \left(S_T - K\right)^+ \frac{1}{K^2} dK\right). \tag{11.93}$$

The expectation on the R.H.S. of Eq. (11.93) is (minus) half that on the R.H.S. of Eq. (11.90), assuming interest rates are constant. Eq. (11.90) then follows by Eq. (11.92).

11.8.2.3 The market for volatility and further developments

The behavior of the VIX index has been a topic intensively studied in empirical research. [Cite references.] The following pictures depict the time series behavior of the new VIX index since its inception as well as dynamics of volume on VIX options.


The VIX Index (daily, annualized, %), with NBER recessions
90
80
70
60
50
40
30
20
10
0
2004
2005
2006
2007
2008
2009
2010
2011
2012


Total VIX Options Volume - Daily contracts, in thousands
Contracts traded on a day
Average contracts over the previous month
1200
1000
800
600
400
200
0
2006
2007
2008
2009
2010
2011
2012

Derivatives referenced to the VIX index have been traded quite widely. Their role is to replace expensives straddles and makes books less messy with outcomes consistent with views—eliminates pric dependency. The following table depicts transaction data for options on the VIX index as compared to other options cleared by CBOE. (Note that the notionals relating to VIX options and futures are not the same, being $1000 for VIX futures and $100 for VIX options, as of August 2011.)

CBOE trading volume (contracts)—Average per day, August 2011

Total trading volume	$\sim$ 12,000,000		
CBOE Index Options	$\sim$ 2,000,000	(i)	
S&P 500 Options	$\sim$ 1,300,000	(ii)	
CBOE VIX Options	$\sim$ 582,000	$\sim \frac{1}{6}$ of (i)+(ii)	$\sim \frac{1}{2}$ of (ii)
CBOE VIX Futures	$\sim$ 79,000		

Section 11.6.3 explains how the skew relates to local volatility, but how is the expected variance in Eq. (11.90) related to the skew? Demeterfi, Derman, Kamal and Zou (1999) show that if the implied volatility varies linearly with the strike,

$$\text{IV} = \text{IV}_{\text{atm}} - b\frac{K-S}{S},$$

for some constant b, then,

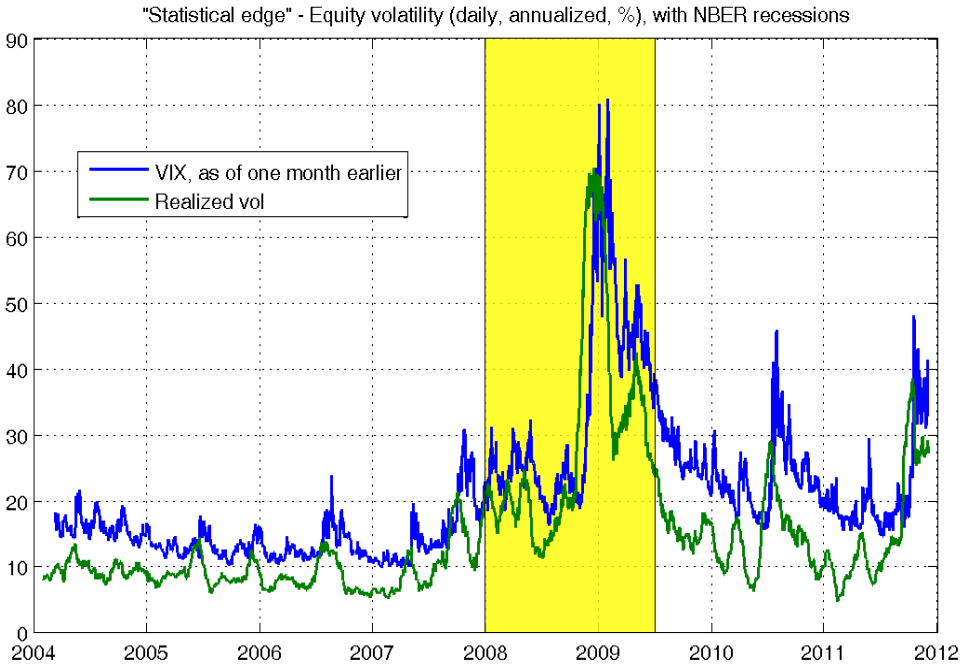
$$\frac{1}{T-t}\mathbb{E}\left[\text{var}\left(t,T\right)\right] \approx \text{IV}_{\text{atm}}^2 \cdot \left(1+3\left(T-t\right)b^2\right).$$

That is, the existence of a skew, $b \neq 0$, increases the value of the fair variance above the at-the-money implied volatility.

Sometimes it is said that variance swaps are profitable to protection sellers, because "the derivative house has the statistical edge," meaning that the realized variance from t to T, say, is general lower than future expected variance under the risk-neutral probability, reflecting variance risk-premiums, as shown in the following picture. The difference between expected variance under the two probabilities is referred to as a risk premium for standard reasons: note, heuristically, that for any variable measurable at T, X_T, we have that

$$-cov\left(\frac{dQ}{dP}, X_T\right) = E_t\left(X_T\right) - \mathbb{E}_t\left(X_T\right).$$

When interest rates are constants, the covariance between a pricing kernel and X_T is indeed a risk-premium. Sometimes, we say that the variance risk-premium is negative based on the previous relation and the empirical evidence in the next figure.



11.8.3 Forward volatility trading

Let us consider the following example of structured volatility trading. Suppose we hold the view that expected market volatility will rise in one year time, to an extent that is inefficiently priced in by the term structure of the currently traded variance swaps. Precisely, our view is that the spot price of the variance swap in one year will exceed the "implied forward variance swap price," i.e.

$$\mathbb{P}_{\text{var}}(1,2) > \mathbb{P}_{\text{var}}(0,2) - \mathbb{P}_{\text{var}}(0,1). \tag{11.94}$$

To implement a trade consistent with this view, we may proceed as follows:

$$\begin{array}{lll} \text{(i)} & \text{long a two year variance swap, struck at } \mathbb{P}_{\text{var}}(0,2), & \text{with notional one} \\ \text{(ii)} & \text{short a one year variance swap, struck at } \mathbb{P}_{\text{var}}(0,1), & \text{with notional } e^{-r} \end{array} \tag{11.95}$$

Obviously, this strategy does not cost, at time zero.

The strategy in (11.95) generates profits whenever the inequality in (11.94) holds true. Indeed, suppose Eq. (11.94) holds true at time 1. Then, come time 1, we can short another one year variance swap, struck at $\mathbb{P}_{\text{var}}(1,2)$. Intuitively, we do so because "we bought it cheap," according to (11.94). Shorting this variance swap at time 1 generates the following payoff at time 2:

$$\pi_1(2) \equiv \mathbb{P}_{\text{var}}(1,2) - \text{var}(1,2). \tag{11.96}$$

Moreover, the two year variance swap we went long at time zero (component (i) of 11.95) gives rise to the following payoff at time 2:

$$\pi_2(2) \equiv \text{var}(0,2) - \mathbb{P}_{\text{var}}(0,2). \tag{11.97}$$

Adding Eq. (11.96) and Eq. (11.97), and using the relation, $\text{var}(0,2) = \text{var}(0,1) + \text{var}(1,2)$, leads to:

$$\pi(2) \equiv \pi_1(2) + \pi_2(2) = \mathbb{P}_{\text{var}}(1,2) + \text{var}(0,1) - \mathbb{P}_{\text{var}}(0,2).$$

Finally, the one year variance swap with notional e^{-r} we shorted at time zero (component (ii) of 11.95) leads to the following payoff at time 1:

$$\pi(1) \equiv (\mathbb{P}_{\text{var}}(0,1) - \text{var}(0,1))\, e^{-r}. \tag{11.98}$$

Investing $\pi(1)$ for a further year at the safe interest rate delivers $\pi(1)\, e^r$ at time 2, such that the total profits at time 2 are:

$$\pi_{\text{tot}} \equiv \pi(2) + \pi(1)\, e^r = \mathbb{P}_{\text{var}}(1,2) - \mathbb{P}_{\text{var}}(0,2) + \mathbb{P}_{\text{var}}(0,1) > 0, \tag{11.99}$$

where the inequality follows by Eq. (11.94).

11.8.4 Marking to market

Suppose a variance contract expiring at time T is issued at time t, when it is costless. How is this contract worth at time $\tau \in (t,T)$? Let us take the time τ risk-neutral discounted expectation of π_T^{var} in Eq. (11.87),

$$\begin{aligned} \frac{\mathbb{E}_\tau(\pi_T^{\text{var}})}{\text{Notional}} &= e^{-r(T-\tau)}\mathbb{E}_\tau(\text{var}(t,\tau) + \text{var}(\tau,T) - \mathbb{P}_{\text{var}}(t,T)) \\ &= e^{-r(T-\tau)}(\text{var}(t,\tau) + \mathbb{P}_{\text{var}}(\tau,T) - \mathbb{P}_{\text{var}}(t,T)). \end{aligned} \tag{11.100}$$

where $\mathbb{E}_\tau$ denotes the risk-neutral expectation conditional upon the information available at time τ.

Marking to market suggests an alternative way to implement the forward volatility trading of the previous section. Suppose, then, again, to have the view that markets for volatility will be such that (11.94) holds true at time 1, and, accordingly, consider the strategy in (11.95). If (11.94) holds true at time 1, we may close the position (i) in (11.95) at time 1. By Eq. (11.100), the market value of the two year variance swap we went long at time 0 is,

$$\hat{\pi}(1) \equiv (\text{var}(0,1) + \mathbb{P}_{\text{var}}(1,2) - \mathbb{P}_{\text{var}}(0,2))\, e^{-r}. \tag{11.101}$$

At time 1, we obtain $\hat{\pi}(1) + \pi(1)$, which we can invest at the safe interest rate for one additional period, delivering the profit π_{tot} in Eq. (11.99), for time 2.

11.8.5 Stochastic interest rates

When interest rates are stochastic, but still independent of volatility, the expressions given for the contract and indexes do not hold anymore, and there are a number of qualifications, which we make in Remark A.1 of Appendix 5. Moreover, the forward volatility trading strategy in 11.95 should be modified. For example, we may use the following strategy:

(i) long a two year variance swap, struck at $\mathbb{P}_{\text{var}}(0,2)$, with notional one
(ii) short a one year variance swap, struck at $\mathbb{P}_{\text{var}}(0,1)$, with notional $\frac{P(0,2)}{P(0,1)}$

If come time 1, Eq. (11.94) holds true, we may liquidate (i), thereby accessing the payoff relating to (ii), for a total payoff equal to:

$$\begin{aligned}
&\left(\text{var}\left(0,1\right)+\mathbb{P}_{\text{var}}\left(1,2\right)-\mathbb{P}_{\text{var}}\left(0,2\right)\right)P\left(1,2\right)+\left(\mathbb{P}_{\text{var}}\left(0,1\right)-\text{var}\left(0,1\right)\right)\frac{P\left(0,2\right)}{P\left(0,1\right)}\\
=&\left(\mathbb{P}_{\text{var}}\left(1,2\right)-\mathbb{P}_{\text{var}}\left(0,2\right)+\mathbb{P}_{\text{var}}\left(0,1\right)\right)P\left(1,2\right)\\
&+\left(\text{var}\left(0,1\right)-\mathbb{P}_{\text{var}}\left(0,1\right)\right)\left(P\left(1,2\right)-\frac{P\left(0,2\right)}{P\left(0,1\right)}\right),
\end{aligned}$$

where the first term on the left hand side arises by the liquidation of (i) and by Eq. (11.101), and the second term on the left hand side arises by (ii). By Eq. (11.94), the first term on the right hand side is positive. If the short-term interest rate was deterministic, $P\left(1,2\right)=\frac{P(0,2)}{P(0,1)}$, and the second term on the right hand side would be zero. When interest rates are stochastic, the second term can take on any sign although then, its absolute value should be quite low, compared to the first term on the right hand side.

11.8.6 Hedging

A financial institution might be merely interested in intermediating the contract, which then needs to be hedged against. Suppose, for example, that the financial institution sells protection at time t, thereby promising to pay the realized integrated variance $\text{var}\left(t,T\right)$ at time T. We want to replicate this integrated variance. By Itô's lemma:

$$\text{var}\left(t,T\right)=2\int_{t}^{T}\frac{1}{S_{u}}dS_{u}-2\ln\left(\frac{S_{T}}{S_{t}}\right)=2\left(\int_{t}^{T}\frac{1}{S_{u}}dS_{u}-r\left(T-t\right)\right)-2\ln\frac{F_{T}}{F_{t}}. \tag{11.102}$$

The first term can be replicated by continuously rebalancing a stock position, which is always long $\theta_{t}=\frac{2}{S_{t}}$ shares of the stock, adjusted for the time value of money. Precisely, consider a self-financed portfolio $(\theta_{\tau},\psi_{\tau})$, such that its value satisfies:

$$V_{\tau}=\theta_{\tau}S_{\tau}+\psi_{\tau}M_{\tau},$$

where M_{τ} denotes the money market account. We choose:

$$\hat{\theta}_{\tau}=\frac{1}{S_{\tau}}\frac{M_{\tau}}{M_{T}},\quad\hat{\psi}_{\tau}=\left[\int_{t}^{\tau}\frac{1}{S_{u}}dS_{u}-1-r\left(\tau-t\right)\right]\frac{1}{M_{T}}. \tag{11.103}$$

It is easy to see that

$$\hat{V}_{\tau}=\left[\int_{t}^{\tau}\frac{1}{S_{u}}dS_{u}-r\left(\tau-t\right)\right]\frac{M_{\tau}}{M_{T}}, \tag{11.104}$$

such that: (i) $\hat{V}_{t}=0$, and (ii) $\hat{V}_{T}=\int_{t}^{T}\frac{1}{S_{u}}dS_{u}-r\left(T-t\right)$. In Appendix 5, we show that $(\hat{\theta}_{\tau},\hat{\psi}_{\tau})$ is self-financed. The bottom line is that we can hedge the first term in Eq. (11.102) through a self-financed portfolio that costs nothing at time t. This portfolio is simply $(2\hat{\theta}_{\tau},2\hat{\psi}_{\tau})$.

To replicate the second term in Eq. (11.102), the payoff of the so-called log-contract, note that we simply have to make reference to twice Eq. (11.93). Therefore, the log-contract can be replicated by shorting $2/F_{t}$ units of forwards, which are of course costless at time t, and going long a continuum of out-of-the-money options with weights $2/K^{2}$, which cost

$$2\int_{0}^{F_{t}}\frac{P_{t}\left(K,T\right)}{K^{2}}dK+2\int_{F_{t}}^{\infty}\frac{C_{t}\left(K,T\right)}{K^{2}}dK=e^{-r(T-t)}\mathbb{E}\left[\text{var}\left(t,T\right)\right]$$

where the equality follows by Eq. (11.90). We borrow $e^{-r(T-t)}\mathbb{E}\left[\text{var}\left(t,T\right)\right]$ to purchase these options, and once this is done, we are guaranteed $\text{var}\left(t,T\right)$ is replicated at time T, as we now have replicated both the first term and the second term in Eq. (11.102). Finally, come time T, we pay back the loan, worth $\mathbb{E}\left[\text{var}\left(t,T\right)\right]$, and receive a payoff equal to $\text{var}\left(t,T\right)-\mathbb{E}\left[\text{var}\left(t,T\right)\right]$, due to the sale of insurance. Since $\text{var}\left(t,T\right)$ is replicated, no additional funds are needed at time T.

11.9 A digression on skewness

What is the value to hedge against left tail risks? In part, equity volatility accounts for left tail risk, due to its negative correlation with the market. However, volatility does not uniquely track these risks. Skewness does. Model-free skewness contracts may be designed to deal with tail risk. Indexes based on skewness are maintained by CBOE since January 2011. These indexes are interesting because it seems that, at times, their value correlates with the variance risk-premium: the higher the variance risk-premium, the higher the expected skewness (under Q).

To evaluate a skewness contract, consider the following general perspective. Appendix 6 shows that for any twice differentiable function f,

$$e^{-r(T-t)}\mathbb{E}\left[f\left(F_T\right)-f\left(F_t\right)\right]=\int_0^{F_t}f''\left(K\right)\cdot\text{Put}_t\left(K,T\right)dK+\int_{F_t}^{\infty}f''\left(K\right)\cdot\text{Call}_t\left(K,T\right)dK, \tag{11.105}$$

where $F_t=e^{r(T-t)}S_t$, the forward rate, and the remaining is usual notation. Eq. (11.105) shows that any Markov payoff can be spanned through a set of European options. For example, if $f\left(x\right)=\ln x$, we can price a log-contract, which leads to price variance swaps (see Section 11.8). We now rely on Eq. (11.105) to evaluate skewness contracts.

Consider the following payoff, $f_{\text{v}}\left(F_T\right)\equiv\left(\ln\frac{F_T}{F_t}\right)^2$. By construction, $f_{\text{v}}\left(F_t\right)=0$, such that, by Eq. (11.105),

$$\mathbb{E}\left[f_{\text{v}}\left(F_T\right)\right]=e^{r(T-t)}\left(\int_0^{F_t}\omega_{\text{v},t}\left(K\right)\cdot\text{Put}_t\left(K,T\right)dK+\int_{F_t}^{\infty}\omega_{\text{v},t}\left(K\right)\cdot\text{Call}_t\left(K,T\right)dK\right),$$

$$\omega_{\text{v},t}\left(K\right)\equiv f_{\text{v}}''\left(F_T\right)=\frac{2}{K^2}\left(1-\ln\frac{K}{F_t}\right).$$

Which contract does the payoff $f_{\text{v}}\left(F_T\right)$ relate to? Clearly, $f_{\text{v}}\left(F_T\right)$ relates to the second moment of the cumulative return, $\ln\frac{F_T}{F_t}$. But this second moment is quite distinct from the realized variance, which we always defined as the sum of instantaneous return variances, $\int_t^T\left(\frac{dF_u}{F_u}\right)^2$. Precisely, note that by Itô's lemma,

$$\mathbb{E}\left[f_{\text{v}}\left(F_T\right)\right]=\mathbb{E}\left[2\int_t^T\ln\left(\frac{F_\tau}{F_t}\right)\frac{dF_\tau}{F_\tau}+\int_t^T\left(1-\ln\frac{F_\tau}{F_t}\right)\left(\frac{dF_\tau}{F_\tau}\right)^2\right]. \tag{11.106}$$

This variance contract is a bit unusual at the time of writing: the standard notion of variance we use to price is $\int_t^T\left(\frac{dF_u}{F_u}\right)^2$, not $\left(\ln\frac{F_T}{F_t}\right)^2$.

The current literature and practice on skewness contracts display a similar "cumulative return" flavor. Consider the following payoff, introduced by Bakshi, Kapadia and Madan (2003),

$$f_{\text{sk}}\left(F_T\right)\equiv\left[\ln\frac{F_T}{F_t}-\mathbb{E}\left(\ln\frac{F_T}{F_t}\right)\right]^3,\quad f_{\text{sk}}\left(F_t\right)=0.$$

The payoff $f_{\text{sk}}(F_T)$ relates to the third moment of the cumulative return over a certain investment horizon. Instead, the notion of a "realized skewness" relies on the third moments of the instantaneous returns, and averaged over a given horizon. Pricing realized skewness is unexplored territory at the time of writing. Let us consider, instead, the payoff $f_{\text{sk}}(F_T)$, and consider the definition of skewness, adjusted for risk,

$$\text{Skew} \equiv \frac{\mathbb{E}\left[f_{\text{sk}}(F_T)\right]}{\mathbb{E}\left[f_{\text{vv}}(F_T)\right]^{\frac{3}{2}}},$$

where

$$f_{\text{vv}}(F_T) \equiv \left[\ln\frac{F_T}{F_t} - \mathbb{E}\left(\ln\frac{F_T}{F_t}\right)\right]^2, \quad f_{\text{vv}}(F_t) = 0.$$

This concept of skewness is "adjusted for risk" because expectations are taken under the risk-neutral probability. Note that $f_{\text{vv}}(F_T)$ is the de-meaned version of $f_{\text{v}}(F_T)$, such that its expectation is

$$\mathbb{E}\left[f_{\text{vv}}(F_T)\right] = \mathbb{E}\left[f_{\text{v}}(F_T)\right] - \mathbb{E}\left[f_{\log}(F_T)\right]^2,$$

where $f_{\log}(F_T) = \ln\frac{F_T}{F_t}$, the log-contract,

$$-\mathbb{E}\left[f_{\log}(F_T)\right] = e^{r(T-t)}\left(\int_0^{F_t}\frac{1}{K^2}\cdot\text{Put}_t(K,T)\,dK + \int_{F_t}^{\infty}\frac{1}{K^2}\cdot\text{Call}_t(K,T)\,dK\right).$$

Likewise,

$$\mathbb{E}\left[f_{\text{sk}}(F_T)\right] = e^{r(T-t)}\left(\int_0^{F_t}\omega_{\text{sk},t}(K)\cdot\text{Put}_t(K,T)\,dK + \int_{F_t}^{\infty}\omega_{\text{sk},t}(K)\cdot\text{Call}_t(K,T)\,dK\right),$$
$$\omega_{\text{sk},t}(K) \equiv f''_{\text{sk}}(F_T) = \frac{3}{K^2}\left(\ln\frac{K}{F_t} - \mathbb{E}\left[f_{\log}(F_T)\right]\right).$$

11.10 Dealing with market imperfections

11.11 Appendix 1: The original arguments of Black & Scholes

The arguments in Black and Scholes (1973) and Merton (1973) rely on the assumption the option is traded. Accordingly, create a self-financed portfolio of $\bar{n}_S$ units of the underlying asset and n_C units of the European call option, where $\bar{n}_S$ is an arbitrary number. Such a portfolio is worth $V = \bar{n}_S S + n_C C$ and since it is self-financed it satisfies:

$$\begin{aligned} dV &= \bar{n}_S dS + n_C dC \\ &= \bar{n}_S dS + n_C \left[C_S dS + \left(C_\tau + \frac{1}{2}\sigma^2 S^2 C_{SS} \right) d\tau \right] \\ &= \left(\bar{n}_S + n_C C_S \right) dS + n_C \left(C_\tau + \frac{1}{2}\sigma^2 S^2 C_{SS} \right) d\tau \end{aligned}$$

where the second line follows by Itô's lemma. Therefore, the portfolio is locally riskless whenever

$$n_C = -\bar{n}_S \frac{1}{C_S},$$

in which case V would appreciate at the safe rate,

$$\frac{dV}{V} = \frac{n_C \left(C_\tau + \frac{1}{2}\sigma^2 S^2 C_{SS} \right) d\tau}{\bar{n}_S S + n_C C} = \frac{-\frac{1}{C_S} \left(C_\tau + \frac{1}{2}\sigma^2 S^2 C_{SS} \right)}{S - \frac{1}{C_S} C} d\tau = r d\tau.$$

The last equality, and the boundary condition, lead to the Black-Scholes partial differential equation (11.14).

11.12 Appendix 2: Black (1976)

We want to evaluate the following expectation:

$$\mathbb{E}\left(x_T - K\right)^+,$$

where

$$x_T = x_t e^{-\frac{1}{2}\int_t^T \gamma_\tau^2 d\tau + \int_t^T \gamma_\tau d\tilde{W}_\tau}.$$

Let $\mathbb{I}_{\text{exe}}$ be the indicator of all events s.t. $x_T \geq K$. We have

$$\begin{aligned}
\mathbb{E}_t\left(x_T - K\right)^+ &= \mathbb{E}_t\left(x_T \cdot \mathbb{I}_{\text{exe}}\right) - K \cdot \mathbb{E}_t\left(\mathbb{I}_{\text{exe}}\right) \\
&= x_t \cdot \mathbb{E}_t\left(\frac{x_T}{x_t} \cdot \mathbb{I}_{\text{exe}}\right) - K \cdot \mathbb{E}_t\left(\mathbb{I}_{\text{exe}}\right) \\
&= x_t \cdot \mathbb{E}_{Q^x}\left(\mathbb{I}_{\text{exe}}\right) - K \cdot \mathbb{E}_t\left(\mathbb{I}_{\text{exe}}\right) \\
&= x_t \cdot Q^x\left(x_T \geq K\right) - K \cdot Q\left(x_T \geq K\right).
\end{aligned}$$

where the probability Q^x is defined as:

$$\frac{dQ^x}{dQ} = \frac{x_T}{x_t} = e^{-\frac{1}{2}\int_t^T \gamma_\tau^2 d\tau + \int_t^T \gamma_\tau d\tilde{W}_\tau},$$

a Q-martingale starting at one. Under Q^x, $dW_\tau^x = d\tilde{W}_\tau - \gamma d\tau$ is a Brownian motion, such that

$$x_T = x_t e^{\frac{1}{2}\int_t^T \gamma_\tau^2 d\tau + \int_t^T \gamma_\tau dW_\tau^x}.$$

Note that $Q\left(x_T \geq K\right) = \Phi(d_2)$ and $Q^x\left(x_T \geq K\right) = \Phi(d_1)$, where

$$d_{2/1} = \frac{\ln\left(\frac{x_t}{K}\right) \mp \frac{1}{2}\int_t^T \gamma_\tau^2 d\tau}{\sqrt{\int_t^T \gamma_\tau^2 d\tau}}.$$

11.13 Appendix 3: Stochastic volatility

11.13.1 Hull & White equation

We assume that the asset price and its volatility solve Eqs. (11.36), but that the two Brownian motions W and W_σ are uncorrelated. Note, first that conditionally on the variance path $\left(\sigma_\tau^2\right)_{\tau\in[t,T]}$, $\ln\left(\frac{S_T}{S_t}\right)$ is normally distributed under the risk-neutral probability,

$$\ln\left(\frac{S_T}{S_t}\right) = r(T-t) - \frac{1}{2}\int_t^T \sigma_\tau^2 d\tau + \int_t^T \sigma_\tau dW_\tau,$$

with:

$$\begin{aligned}
\mathbb{E}_t\left[\ln\left(\frac{S_T}{S_t}\right)\middle|\left(\sigma_\tau^2\right)_{\tau\in[t,T]}\right] &= r(T-t) - \frac{1}{2}(T-t)\,\tilde{V}_{t,T} \\
var_t\left[\ln\left(\frac{S_T}{S_t}\right)\middle|\left(\sigma_\tau^2\right)_{\tau\in[t,T]}\right] &= \int_t^T \sigma_\tau^2 d\tau = (T-t)\,\tilde{V}_{t,T}
\end{aligned}$$

Then, we use the law of iterated expectations and elaborate on Eq. (11.50), and arrive to Eq. (11.52) as follows:

$$\begin{aligned}
C\left(S_t,\sigma_t^2,T-t\right) &= e^{-r(T-t)}\mathbb{E}_t\left[\left(S_T-K\right)^+\middle|\,S_t,\sigma_t^2\right] \\
&= \mathbb{E}_t\left[\mathbb{E}\left(e^{-r(T-t)}\left(S_T-K\right)^+\middle|\,S_t,\left(\sigma_\tau^2\right)_{\tau\in[t,T]}\right)\middle|\,S_t,\sigma_t^2\right] \\
&= \mathbb{E}_t[C_{\text{BS}}(S_t,T-t;K,\sqrt{\tilde{V}_{t,T}})\Big|\,S_t,\sigma_t^2] \\
&= \mathbb{E}_t[C_{\text{BS}}(S_t,T-t;K,\sqrt{\tilde{V}_{t,T}})\Big|\,\sigma_t^2] \\
&= \int C_{\text{BS}}(S_t,T-t;K,\sqrt{\tilde{V}_{t,T}})\phi(\tilde{V}_{t,T}\Big|\,\sigma_t^2)d\tilde{V}_{t,T} \\
&\equiv \mathbb{E}_t^{\tilde{V}}[C_{\text{BS}}(S_t,T-t;K,\sqrt{\tilde{V}_{t,T}})],
\end{aligned} \tag{11A.1}$$

where $\phi(\tilde{V}_{t,T}\Big|\,\sigma_t^2)$ denotes the density of $\tilde{V}_{t,T}$ conditional upon the current level of the variance, σ_t^2. The third and fourth equalities follow by the assumption W and W_σ are uncorrelated, such that

$$\phi(\tilde{V}_{t,T}\Big|\,S_t,\sigma_t^2) = \phi(\tilde{V}_{t,T}\Big|\,\sigma_t^2),$$

for otherwise the current level of the index, S_t, would help predict $\tilde{V}_{t,T}$. In other words, Eq. (11A.1) reveals that the price of an option in this market with stochastic volatility is Black & Scholes weighted with the probability density of the realized variance.

11.13.2 Extensions

Romano and Touzi (1997) extend the Hull & White equation to the case where asset returns and volatility are correlated. Consider the following model:

$$\left\{\begin{array}{rcl}
\dfrac{dS_t}{S_t} &=& rdt + \sigma_t\left(\rho_t dW_t^\sigma + \sqrt{1-\rho_t^2}dZ_t\right) \\
d\sigma_t^2 &=& b\left(\sigma_t\right)dt + a\left(\sigma_t\right)dW_t^\sigma
\end{array}\right.$$

where the correlation process, ρ_t, does not depend on S_t, and W_σ and Z are two independent Brownian motions, such that,

$$S_T = S_t L_{t,T}\cdot e^{r(T-t)-\frac{1}{2}\int_t^T \sigma_\tau^2\left(1-\rho_\tau^2\right)d\tau+\int_t^T \sigma_\tau\sqrt{1-\rho_\tau^2}dZ_\tau},$$

where $L_{t,T}$ is as in Eq. (11.54) of the main text. We have, using the Law of Iterated Expectations,

$$\begin{aligned}
&C\left(S_t,\sigma_t^2,\rho_t,T-t\right)\\
&=e^{-r(T-t)}\mathbb{E}_t\left[\left(S_T-K\right)^+\middle|\,S_t,\sigma_t^2,\rho_t\right]\\
&=\mathbb{E}_t\left[\mathbb{E}\left(e^{-r(T-t)}\left(S_T-K\right)^+\middle|\,S_t,\left(\sigma_\tau^2,\rho_\tau\right)_{\tau\in[t,T]}\right)\middle|\,S_t,\sigma_t^2,\rho_t\right]\\
&=\mathbb{E}_t[C_{\mathrm{BS}}(S_tL_{t,T},T-t;K,\sqrt{\tilde{V}_{t,T}^{\rho}})\Big|\,S_t,\sigma_t^2,\rho_t],
\end{aligned}$$

where $\tilde{V}_{t,T}^{\rho}$ is as in Eq. (11.54) of the main text. The third equality follows because by assumption, both the variance and correlation processes are independent of $\{Z_\tau\}_{\tau\in[t,T]}$, such that conditionally upon the variance and the correlation paths, $\left(\sigma_\tau^2\right)_{\tau\in[t,T]}$ and $(\rho_\tau)_{\tau\in[t,T]}$, $\ln\left(\frac{S_T}{S_tL_{t,T}}\right)$ is normally distributed under the risk-neutral probability,

$$\ln\left(\tfrac{S_T}{S_tL_{t,T}}\right)=r\left(T-t\right)-\frac{1}{2}\int_t^T\sigma_\tau^2\left(1-\rho_\tau^2\right)d\tau+\int_t^T\sigma_\tau\sqrt{1-\rho_\tau^2}dZ_\tau,$$

with

$$\begin{aligned}
\mathbb{E}_t\left[\ln\left(\tfrac{S_T}{S_tL_{t,T}}\right)\middle|\left(\sigma_\tau^2,\rho_\tau\right)_{\tau\in[t,T]}\right]&=r(T-t)-\frac{1}{2}\left(T-t\right)\tilde{V}_{t,T}^{\rho}\\
var_t\left[\ln\left(\tfrac{S_T}{S_tL_{t,T}}\right)\middle|\left(\sigma_\tau^2,\rho_\tau\right)_{\tau\in[t,T]}\right]&=\int_t^T\sigma_\tau^2d\tau=\left(T-t\right)\tilde{V}_{t,T}^{\rho}
\end{aligned}$$

The fourth line follows by the same arguments leading to Eq. (11A.1).

11.13.3 Smile analytics

We develop the approximation stated in Eq. (11.37). The assumption is that the asset price and its volatility solve Eqs. (11.36), with the Brownians W and W_σ being uncorrelated, such that the market price of the option is that generated by the Hull & White equation (11.52). By definition, the Black & Scholes implied volatility, IV, satisfies,

$$C_t^{\$}=\mathbb{E}_t[C_{\mathrm{BS}}(S,T-t,K;\sqrt{\tilde{V}_{t,T}})]=C_{\mathrm{BS}}(S,T-t;\mathrm{IV}\left(X,T-t\right)),\quad X\equiv\ln\frac{S_te^{r(T-t)}}{K}.\tag{11A.2}$$

Let $\mu_V\left(t\right)=\mathbb{E}_t(\sqrt{\tilde{V}_{t,T}})$ the expected average volatility, and consider a Taylor's second order expansion of the Black & Scholes function about $\mu_V\left(t\right)$,

$$\begin{aligned}
&\mathbb{E}_t[C_{\mathrm{BS}}(S,T-t,K;\sqrt{\tilde{V}_{t,T}})]\\
&\approx C_{\mathrm{BS}}(S,T-t,K;\sqrt{\mu_V\left(t\right)})+\frac{1}{2}\left.\frac{\partial^2C_{\mathrm{BS}}(S,T-t;K,\sqrt{V})}{\partial\sqrt{V}^2}\right|_{\sqrt{V}=\mu_V(t)}\cdot var_t(\sqrt{\tilde{V}_{t,T}}),
\end{aligned}$$

and,

$$\begin{aligned}
&C_{\mathrm{BS}}\left(S,T-t;\mathrm{IV}\left(X,T-t\right)\right)\\
&\approx C_{\mathrm{BS}}(S,T-t,K;\sqrt{\mu_V\left(t\right)})+\left.\frac{\partial C_{\mathrm{BS}}(S,T-t;K,\sqrt{V})}{\partial\sqrt{V}}\right|_{\sqrt{V}=\mu_V(t)}\cdot\left(\mathrm{IV}\left(X,T-t\right)-\mu_V\left(t\right)\right).
\end{aligned}$$

By plugging these two approximations into Eq. (11A.2) leaves:

$$\mathrm{IV}\left(X, T-t\right) \approx \mu_V\left(t\right) + \frac{1}{2}\left.\frac{\partial^2 C_{\mathrm{BS}}(S, T-t; K, \sigma)/\,\partial\sigma^2}{\partial C_{\mathrm{BS}}(S, T-t; K, \sigma)/\,\partial\sigma}\right|_{\sigma=\mu_V(t)} \cdot var_t(\sqrt{\tilde{V}_{t,T}}). \tag{11A.3}$$

The vega, $\partial C_{\mathrm{BS}}/\,\partial\sigma$, and the volga, $\partial^2 C_{\mathrm{BS}}/\,\partial\sigma^2$, for the Black-Scholes model are well-known. They are:

$$\begin{aligned}
\frac{\partial C_{\mathrm{BS}}\left(S, T-t; \sigma\right)}{\partial\sigma} &= S\sqrt{T-t}\phi\left(\tfrac{X+\frac{1}{2}\sigma^2(T-t)}{\sigma\sqrt{T-t}}\right) \\
\frac{\partial^2 C_{\mathrm{BS}}(S, T-t; K, \sigma)}{\partial\sigma^2} &= \left(\tfrac{X^2-\left(\frac{1}{2}\sigma^2(T-t)\right)^2}{\sigma^3(T-t)}\right)\frac{\partial C_{\mathrm{BS}}(S, T-t; K, \sigma)}{\partial\sigma}.
\end{aligned}$$

Replacing these expressions into Eq. (11A.3) yields the approximation in Eq. (11.37) of the main text.

11.14 Appendix 4: Local volatility

The arguments underlying the derivation of the local volatility function in this appendix rely on Derman and Kani (1994), Dumas (1995) and Britten-Jones and Neuberger (2000). In all the proofs to follow, all expectations are taken to be conditional on $\mathcal{F}_t$, but to simplify notation, we write $\mathbb{E}(\cdot|\cdot) \equiv \mathbb{E}(\cdot|\cdot, \mathcal{F}_t)$.

Proof of Eqs. (11.79) and (11.88). We first derive Eq. (11.79), a result encompassing Eq. (11.76). We assume that the under the risk-neutral probability, the stock price is solution to,

$$\frac{dS_t}{S_t} = rdt + \sigma_t d\hat{W}_t, \tag{11A.4}$$

where σ_t is some $\mathcal{F}_t$-adapted process. For example, $\sigma_t \equiv \sigma(S_t, t) \cdot v_t$, all t, where v_t is solution to the second of Eqs. (11.78). Next, we assume that we are given a continuum of option prices $C_\$(K,T)$ along the two dimensions of strikes K and time-to-maturities T. We want to match the prediction of the model with the market prices,

$$C_\$(K,T) = e^{-r(T-t)}\mathbb{E}(S_T - K)^+, \tag{11A.5}$$

where S_τ is solution to Eq. (11A.4).

Let us expand the right-hand side of Eq. (11A.5) with respect to time-to-maturity, for fixed K,

$$d_T(S_T - K)^+ = \left(\mathbb{I}_{S_T \geq K} r S_T + \frac{1}{2}\delta(S_T - K)\sigma_T^2 S_T^2\right) dT + \mathbb{I}_{S_T \geq K}\sigma_T S_T d\hat{W}_T,$$

where δ is the Dirac's delta. By using the identity, $(S_T - K)^+ + K\mathbb{I}_{S_T \geq K} = S_T \mathbb{I}_{S_T \geq K}$,

$$\frac{d\mathbb{E}(S_T - K)^+}{dT} = r\left(\mathbb{E}(S_T - K)^+ + K\mathbb{E}(\mathbb{I}_{S_T \geq K})\right) + \frac{1}{2}\mathbb{E}\left(\delta(S_T - K)\sigma_T^2 S_T^2\right),$$

where

$$\begin{aligned}
\mathbb{E}\left(\delta(S_T - K)\sigma_T^2 S_T^2\right) &= \iint \delta(S_T - K)\sigma_T^2 S_T^2 \underbrace{\phi_T(\sigma_T|S_T)\phi_T(S_T)}_{\equiv \text{ joint density of } (\sigma_T, S_T)} dS_T d\sigma_T \\
&= \int \sigma_T^2 \left(\int \delta(S_T - K) S_T^2 \phi_T(S_T)\phi_T(\sigma_T|S_T) dS_T\right) d\sigma_T \\
&= K^2 \phi_T(K) \int \sigma_T^2 \phi_T(\sigma_T|S_T = K) d\sigma_T \\
&\equiv K^2 \phi_T(K)\mathbb{E}\left(\sigma_T^2 \middle| S_T = K\right),
\end{aligned}$$

such that,

$$\begin{aligned}
&e^{-r(T-t)}\frac{d\mathbb{E}(S_T - K)^+}{dT} \\
&= r\left(e^{-r(T-t)}\mathbb{E}(S_T - K)^+ + Ke^{-r(T-t)}\mathbb{E}(\mathbb{I}_{S_T \geq K})\right) + \frac{1}{2}K^2 e^{-r(T-t)}\phi_T(K)\mathbb{E}\left(\sigma_T^2 \middle| S_T = K\right).
\end{aligned} \tag{11A.6}$$

Next, differentiate both sides of Eq. (11A.5 with respect to T,

$$\frac{\partial}{\partial T}C_\$(K,T) = -rC_\$(K,T) + e^{-r(T-t)}\frac{\partial \mathbb{E}(S_T - K)^+}{\partial T}. \tag{11A.7}$$

and twice with respect to K,

$$\frac{\partial}{\partial K}C_{\$}\left(K,T\right)=-e^{-r(T-t)}\mathbb{E}\left(\mathbb{I}_{S_T\geq K}\right),\quad \frac{\partial^2 C_{\$}\left(K,T\right)}{\partial K^2}=e^{-r(T-t)}\phi_T\left(K\right), \tag{11A.8}$$

where the second relation is simply the famous relation in Eq. (11.28) of the main text. By replacing Eq. (11A.5), (11A.7) and Eqs. (11A.8) into Eq. (11A.6) leaves,

$$\frac{\partial}{\partial T}C_{\$}\left(K,T\right)=-rK\frac{\partial C_{\$}\left(K,T\right)}{\partial K}+\frac{1}{2}K^2\frac{\partial^2 C_{\$}\left(K,T\right)}{\partial K^2}\mathbb{E}\left(\sigma_T^2\middle|\, S_T=K\right). \tag{11A.9}$$

This is,

$$\mathbb{E}\left[\sigma_T^2\middle|\, S_T=K\right]=2\frac{\dfrac{\partial C_{\$}\left(K,T\right)}{\partial T}+rK\dfrac{\partial C_{\$}\left(K,T\right)}{\partial K}}{K^2\dfrac{\partial^2 C_{\$}\left(K,T\right)}{\partial K^2}}\equiv\sigma_{\text{loc}}^2\left(K,T\right). \tag{11A.10}$$

That is, if Eq. (11A.5) holds true, volatility must be restricted to satisfy Eq. (11A.10). As an example, let $\sigma_t\equiv\sigma\left(S_t,t\right)\cdot v_t$, where v_t is solution to the second of Eqs. (11.78). Then,

$$\begin{aligned}\sigma_{\text{loc}}^2\left(K,T\right)&=\mathbb{E}\left(\sigma_T^2\middle|\, S_T=K\right)\\&=\mathbb{E}\left(\sigma^2\left(S_T,T\right)\cdot v_T^2\middle|\, S_T=K\right)\\&=\sigma^2\left(K,T\right)\mathbb{E}\left(v_T^2\middle|\, S_T=K\right)\\&\equiv\tilde{\sigma}_{\text{loc}}^2\left(K,T\right)\mathbb{E}\left(v_T^2\middle|\, S_T=K\right),\end{aligned}$$

which proves Eq. (11.79).

The converse does also hold true. Consider, for example, the case where $v_t\equiv 1$. That is, suppose that $u\left(x,t;K,T\right)$ is solution to the *terminal value problem*,

$$\begin{cases}0=u_t+\dfrac{1}{2}\sigma^2\left(x,t\right)x^2u_{xx}+rxu_x-ru, & t\in[0,T]\\ u\left(x,t;K,T\right)=\left(x-K\right)^+\end{cases}$$

where x denotes the stock price at expiration. Then, define $C_{\$}\left(K,T\right)\equiv u\left(x,t;K,T\right)$, such that, by Eq. (11A.10), $C_{\$}\left(K,T\right)$ solves the *initial value problem*,

$$\begin{cases}0=-\dfrac{\partial}{\partial T}C_{\$}\left(K,T\right)-rK\dfrac{\partial C_{\$}\left(K,T\right)}{\partial K}+\frac{1}{2}K^2\sigma_{\text{loc}}^2\left(K,T\right)\dfrac{\partial^2 C_{\$}\left(K,T\right)}{\partial K^2}, & T>0\\ C_{\$}\left(K,0\right)=\left(x-K\right)^+\end{cases}$$

where x denotes the initial price. The previous partial differential equation is known as the *Dupire's equation*.

11.15 Appendix 5: Variance contracts

We provide proofs of results relating to volatility contracts.

Proof of Eq. (11.88). We have,

$$\begin{aligned}\mathbb{E}\left(\sigma_T^2\right) &= \int_0^\infty \mathbb{E}\left(\sigma_T^2 \middle| S_T = K\right)\phi_T(K)\,dK \\ &= 2\int_0^\infty \frac{\frac{\partial C(K,T)}{\partial T} + rK\frac{\partial C(K,T)}{\partial K}}{K^2\frac{\partial^2 C(K,T)}{\partial K^2}}\phi_T(K)\,dK \\ &= 2e^{r(T-t)}\int_0^\infty \frac{\frac{\partial C(K,T)}{\partial T} + rK\frac{\partial C(K,T)}{\partial K}}{K^2}dK,\end{aligned}$$

where the second line follows by Eq. (11A.10), and the third line follows by Eq. (11.28). ■

Proof of Eq. (11.90). By a Taylor expansion with remainder, we have that for any function f smooth enough,

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \int_{x_0}^x (x-t) f''(t)\,dt. \tag{11A.11}$$

Let F_t be the forward price, $F_t = e^{r(T-t)}S_t$. By applying this formula to $\ln F_T$,

$$\begin{aligned}\ln F_T &= \ln F_t + \frac{1}{F_t}(F_T - F_t) - \int_{F_t}^{F_T}(F_T - t)\frac{1}{t^2}dt \\ &= \ln F_t + \frac{1}{F_t}(F_T - F_t) - \int_0^{F_t}(K - F_T)^+\frac{1}{K^2}dK - \int_{F_t}^\infty (F_T - K)^+\frac{1}{K^2}dK \\ &= \ln F_t + \frac{1}{F_t}(F_T - F_t) - \int_0^{F_t}(K - S_T)^+\frac{1}{K^2}dK - \int_{F_t}^\infty (S_T - K)^+\frac{1}{K^2}dK,\end{aligned} \tag{11A.12}$$

where the second equality follows because $\int_{x_0}^x (x-t)\frac{1}{t^2}dt = \int_0^{x_0}(t-x)^+\frac{1}{t^2}dt + \int_{x_0}^\infty (x-t)^+\frac{1}{t^2}dt$, and the third equality follows because the forward price at T satisfies $F_T = S_T$. Hence, by $\mathbb{E}(F_T) = F_t$,

$$-\mathbb{E}\left(\ln\frac{F_T}{F_t}\right) = e^{r(T-t)}\left[\int_0^{F_t}\frac{P_t(K,T)}{K^2}dK + \int_{F_t}^\infty \frac{C_t(K,T)}{K^2}dK\right]. \tag{11A.13}$$

On the other hand, by Itô's lemma,

$$\mathbb{E}\left(\int_t^T \sigma_u^2 du\right) = -2\mathbb{E}\left(\ln\frac{F_T}{F_t}\right). \tag{11A.14}$$

By replacing Eq. (11A.14) this formula into Eq. (11A.13) yields Eq. (11.90). ■

Remark A1. The previous proof results hold when the short-term rate is constant. The case of stochastic interest rates can actually be dealt with, although with some tools, which will be introduced more systematically in Chapter 13 (Section 13.2). We anticipate how these tools work in the present appendix, as they allow us to solve for the fair price of variance contracts even when interest rates are stochastic. Note that if interest rates are stochastic, Eq. (11A.13) generalizes to:

$$-\mathbb{E}\left(e^{-\int_t^T r_\tau d\tau}\ln\frac{F_T}{F_t}\right) = -\mathbb{E}\left(e^{-\int_t^T r_\tau d\tau}\left(\frac{F_T - F_t}{F_t}\right)\right) + \int_0^{F_t}\frac{P_t(K,T)}{K^2}dK + \int_{F_t}^\infty \frac{C_t(K,T)}{K^2}dK. \tag{11A.15}$$

The left hand side of Eq. (11A.15) can be written as

$$\mathbb{E}\left(e^{-\int_t^T r_\tau d\tau}\ln\frac{F_T}{F_t}\right)=P\left(t,T\right)\mathbb{E}\left(\frac{e^{-\int_t^T r_\tau d\tau}}{P\left(t,T\right)}\ln\frac{F_T}{F_t}\right)=P\left(t,T\right)\mathbb{E}^{Q^F}\left(\ln\frac{F_T}{F_t}\right), \tag{11A.16}$$

where $\mathbb{E}^{Q^F}$ denotes the expectation taken under a new probability, known as the *forward probability*. Naturally, the first term on the right side of Eq. (11A.15) is zero, as a forward has no value at inception. But then, this zero value condition implies that

$$F_t=\mathbb{E}_t\left(\frac{e^{-\int_t^T r_\tau d\tau}}{P\left(t,T\right)}F_T\right)=\mathbb{E}_t^{Q^F}\left(F_T\right).$$

That is, the forward price is a martingale under the forward probability. Therefore, Eq. (11A.14) is replaced with,

$$\mathbb{E}_t^{Q^F}\left(\int_t^T\omega_u^2du\right)=-2\mathbb{E}_t^{Q^F}\left(\ln\frac{F_T}{F_t}\right), \tag{11A.17}$$

where ω_t now denotes the instantaneous volatility of the forward price. By combining Eqs. (11A.15), (11A.16) and (11A.17), we get,

$$\mathbb{E}_t^{Q^F}\left(\int_t^T\omega_u^2du\right)=\mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau}\int_t^T\omega_u^2du\right)=\frac{2}{P\left(t,T\right)}\left[\int_0^{F_t}\frac{P_t\left(K,T\right)}{K^2}dK+\int_{F_t}^{\infty}\frac{C_t\left(K,T\right)}{K^2}dK\right].$$

That is, the fair price of a variance contract for a swap of *forward volatility* can be expressed in a model-free format. Note that it is the price of a variance contract that we can express in a model-free fashion, not the (undiscounted) expected realized variance. Indeed, the payoff of a variance contract for forward realized variance is:

$$\int_t^T\omega_u^2du-\mathbb{P}_{\text{var}}\left(t,T\right),$$

such that the zero value condition at inception,

$$0=\mathbb{E}_t\left[e^{-\int_t^T r_\tau d\tau}\left(\int_t^T\omega_u^2du-\mathbb{P}_{\text{var}}\left(t,T\right)\right)\right],$$

leads to,

$$\mathbb{P}_{\text{var}}\left(t,T\right)=\frac{1}{P\left(t,T\right)}\mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau}\int_t^T\omega_u^2du\right)=\mathbb{E}_t^{Q^F}\left(\int_t^T\omega_u^2du\right).$$

Proof that $(\hat{\theta}_\tau,\hat{\psi}_\tau)$ in Eq. (11.103) is self-financed. For a portfolio strategy to be self-financed, we need to have $\psi_\tau M_\tau=V_\tau-\theta_\tau S_\tau$ and $dV_\tau=\theta_\tau dS_\tau+\psi_\tau dM_\tau$, or:

$$dV_\tau=\theta_\tau S_\tau\frac{dS_\tau}{S_\tau}+\psi_\tau M_\tau\frac{dM_\tau}{M_\tau}=\theta_\tau S_\tau\left(\frac{dS_\tau}{S_\tau}-rd\tau\right)+rV_\tau d\tau, \tag{11A.18}$$

where the second line follows by $\psi_\tau M_\tau=V_\tau-\theta_\tau S_\tau$. With $(\hat{\theta}_\tau,\hat{\psi}_\tau)$, we have that:

$$\begin{aligned}d\hat{V}_\tau&=\hat{\theta}_\tau dS_\tau+\hat{\psi}_\tau dM_\tau\\&=\frac{M_\tau}{M_T}\frac{dS_\tau}{S_\tau}+\hat{\psi}_\tau M_\tau rd\tau\\&=\frac{M_\tau}{M_T}\frac{dS_\tau}{S_\tau}+\left(\hat{V}_\tau-\frac{M_\tau}{M_T}\right)rd\tau\\&=\frac{M_\tau}{M_T}\left(\frac{dS_\tau}{S_\tau}-rd\tau\right)+r\hat{V}_\tau d\tau,\end{aligned} \tag{11A.19}$$

where we have used the portfolio weights in Eq. (11.103) and the expression for the portfolio value $\hat{V}$ in Eq. (11.104). Eq. (11A.19) is the same as Eq. (11A.18), once we use the portfolio weight $\hat{\theta}_\tau$ in Eq. (11.103). Therefore, $(\hat{\theta}_\tau, \hat{\psi}_\tau)$ is self-financed. ■

11.16 Appendix 6: Skewness contracts

By Eq. (11A.11), we have that, for any function f as many times differentiable as we might need,

$$\begin{aligned} f(F_T) &= f(F_t) + f'(F_t)(F_T - F_t) + \int_{F_t}^{F_T} (F_T - t) f''(t)\, dt \\ &= f(F_t) + f'(F_t)(F_T - F_t) + \int_0^{F_t} f''(K)(K - F_T)^+ dK + \int_{F_t}^{\infty} f''(K)(F_T - K)^+ dK, \end{aligned}$$

where $F_\tau = e^{r(T-\tau)} S_\tau$, the forward rate. Multiplying both sides of this equation by $e^{-r(T-t)}$, and taking expectations, yields Eq. (11.105) in the main text.

References

Bakshi, G. and D. Madan (2000): "Spanning and Derivative Security Evaluation." *Journal of Financial Economics* 55, 205-238.

Bakshi, G., N. Kapadia, D. Madan (2003): "Stock Return Characteristics, Skew Laws, and Differential Pricing of Individual Equity Options." *Review of Financial Studies* 16, 101-143.

Ball, C.A. and A. Roma (1994): "Stochastic Volatility Option Pricing." *Journal of Financial and Quantitative Analysis* 29, 589-607.

Bergman, Y. Z., B. D. Grundy, and Z. Wiener (1996): "General Properties of Option Prices." *Journal of Finance* 51, 1573-1610.

Black, F. (1976a): "The Pricing of Commodity Contracts." *Journal of Financial Economics* 3, 167-179.

Black, F. (1976b): "Studies of Stock Price Volatility Changes." *Proceedings of the 1976 Meeting of the American Statistical Association*, 177-81.

Black, F. and M. Scholes (1973): "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81, 637-659.

Bollerslev, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics* 31, 307-327.

Bollerslev, T., Engle, R. and D. Nelson (1994): "ARCH Models." In: McFadden, D. and R. Engle (Editors): *Handbook of Econometrics* (Volume 4), 2959-3038. Amsterdam, North-Holland

Britten-Jones, M. and A. Neuberger (2000): "Option Prices, Implied Price Processes and Stochastic Volatility." *Journal of Finance* 55, 839-866.

Carr, P. and D. Madan (2001): "Optimal Positioning in Derivative Securities." *Quantitative Finance* 1, 19-37.

Christie, A.A. (1982): "The Stochastic Behavior of Common Stock Variances: Value, Leverage, and Interest Rate Effects." *Journal of Financial Economics* 10, 407-432.

Clark, P.K. (1973): "A Subordinated Stochastic Process Model with Fixed Variance for Speculative Prices." *Econometrica* 41, 135-156.

Corradi, V. (2000): "Reconsidering the Continuous Time Limit of the GARCH(1,1) Process." *Journal of Econometrics* 96, 145-153.

Cox, J.C., S.A Ross and M. Rubinstein (1979): "Option Pricing: A Simplified Approach." *Journal of Financial Economics* 7, 229-263.

Cox, J.C., J.E. Ingersoll and S.A. Ross (1985): "A Theory of the Term Structure of Interest Rates." *Econometrica* 53, 385-407.

Demeterfi, K., E. Derman, M. Kamal and J. Zou (1999): "More Than You Ever Wanted To Know About Volatility Swaps." Goldman Sachs Quantitative Strategies Research Notes.

Derman, E. (1998): "Stochastic Implied Trees: Arbitrage Pricing with Stochastic Term and Strike Structure of Volatility." *International Journal of Theoretical and Applied Finance* 1, 61-110.

Derman, E. and J. Kani (1994): "Riding on a Smile." *Risk* 7, 32-39.

Duffie, D. and C-f. Huang (1985): "Implementing Arrow-Debreu Equilibria by Continuous Trading of Few Long-Lived Securities." *Econometrica* 53, 1337-1356.

Dumas, B. (1995): "The Meaning of the Implicit Volatility Function in Case of Stochastic Volatility." Available from:
`http://www.insead.edu/facultyresearch/faculty/personal/bdumas/research/index.cfm`.

Dupire, B. (1994): "Pricing with a Smile." *Risk* 7, 18-20.

El Karoui, N., M. Jeanblanc-Picqué and S. Shreve (1998): "Robustness of the Black and Scholes Formula." *Mathematical Finance* 8, 93-126.

Engle, R.F. (1982): "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica* 50, 987-1008.

Fama, E. (1965): "The Behaviour of Stock Market Prices." *Journal of Business* 38, 34-105.

Fornari, F. and A. Mele (2006): "Approximating Volatility Diffusions with CEV-ARCH Models." *Journal of Economic Dynamics and Control* 30, 931-966.

Gatheral, J. (2006): *The Volatility Surface: A Practioner's Guide*. New York: John Wiley and Sons.

Hakansson, N. (1979): "The Fantastic World of Finance: Progress and the Free Lunch." *Journal of Financial and Quantitative Analysis* (Proceeding Issue) 14, 717-734.

Harrison, J.M. and D.M. Kreps (1979): "Martingales and Arbitrage in Multiperiod Securities Markets." *Journal of Economic Theory* 20, 381-408.

Heston, S.L. (1993a): "Invisible Parameters in Option Prices." *Journal of Finance* 48, 933-947.

Heston, S.L. (1993b): "A Closed Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options." *Review of Financial Studies* 6, 327-344.

Hicks, J.R. (1939): *Value and Capital*. Oxford: Oxford University Press.

Hull, J. and A. White (1987): "The Pricing of Options with Stochastic Volatilities." *Journal of Finance* 42, 281-300.

Keynes, J.M. (1930): *A Treatise on Money: The Applied Theory of Money*. London: MacMillan.

Mandelbrot, B. (1963): "The Variation of Certain Speculative Prices." *Journal of Business* 36, 394-419.

Mele, A. (1998): *Dynamiques non linéaires, volatilité et équilibre*. Paris: Editions Economica.

Mele, A. and F. Fornari (2000): *Stochastic Volatility in Financial Markets. Crossing the Bridge to Continuous Time*. Boston: Kluwer Academic Publishers.

Merton, R. (1973): "Theory of Rational Option Pricing." *Bell Journal of Economics and Management Science* 4, 637-654.

Nelson, D.B. (1990): "ARCH Models as Diffusion Approximations." *Journal of Econometrics* 45, 7-38.

Nelson, D.B. (1991): "Conditional Heteroskedasticity in Asset Returns: A New Approach." *Econometrica* 59, 347-370.

Neuberger, A. (1994): "Hedging Volatility: the Case for a New Contract." *Journal of Portfolio Management* 20, 74-80.

Parkinson, M. (1980): "The Extreme Value Method for Estimating the Variance of the Rate of Returns." *Journal of Business* 53, 61-68.

Renault, E. (1997): "Econometric Models of Option Pricing Errors." In: Kreps, D., Wallis, K. (Editors): *Advances in Economics and Econometrics* (Volume 3), 223-278. Cambridge: Cambridge University Press.

Renault and Touzi.

Romano, M. and N. Touzi (1997): "Contingent Claims and Market Completeness in a Stochastic Volatility Model." *Mathematical Finance* 7, 399-412.

Rubinstein, M. (1994): "Implied Binomial Trees." *Journal of Finance* 49, 771-818.

Scott, L. (1987): "Option Pricing when the Variance Changes Randomly: Theory, Estimation, and an Application." *Journal of Financial and Quantitative Analysis* 22, 419-438.

SEC-CFTC (2010): "Findings Regarding the Market Events of May 6, 2010." A joint report by the Securities and Exchanges Commission & the Commodity Futures Trading Commission, September.

Tauchen, G. and M. Pitts (1983): "The Price Variability-Volume Relationship on Speculative Markets." *Econometrica* 51, 485-505.

Taylor, S. (1986): *Modeling Financial Time Series*. Chichester, UK: Wiley.

Vasicek, O. (1977): "An Equilibrium Characterization of the Term Structure." *Journal of Financial Economics* 5, 177-188.

Wiggins, J. (1987): "Option Values and Stochastic Volatility: Theory and Empirical Estimates." *Journal of Financial Economics* 19, 351-372.

12
Engineering of fixed income securities

12.1 Introduction

This chapter is an introduction to fixed income security pricing. Sometimes, we say that bonds share similarities with derivatives, in that they draw value from the movements in the underlying risks—in the case of default-free bonds, the interest rates the market requires to invest at different horizons safely. However, fixed income securities quite differ from conceptually standard derivatives such as equity options, say. For example, a simple pure discount bond is difficult to price, being tied down to the time value of money. Its value reflects intertemporal preferences and beliefs of market participants, which are unobservable and, importantly, not traded. Thus, we cannot price this bond in a preference-free format, unlike in the instances conducive to the Black and Scholes (1973) formula reviewed in the previous chapter as well as throughout Part I.

This chapter reviews models where we can *still* price derivatives written on bond prices and other securities in a preference-free setting. We rely on "no-arbitrage models," which are the fixed income counterparts to the local volatility models reviewed in Chapter 11. In this framework, we give up modeling the current security prices in the first place. Rather, we take these prices as given, and exploit all the information embedded into them to extract risk-neutral probabilities of future price movements. Once risk-neutral probs are reverse-engineered, we can price any interest rate derivatives in a preference-free format. "No-arbitrage models" means that the only assumption we are really making is absence of arbitrage (in addition to standard statistical laws for price movements).

Ho and Lee (1986) develop the first model that illustrates this approach. This model is solved in closed-form, but relies on unrealistic assumptions, and could lead to negative interest rates. We develop a calibration approach based on the extraction of Arrow-Debreu security prices, which can accommodate for more realistic interest rate developments. Arrow-Debreu securities are abstract securities that pay off in mutually exclusive states of the world, as explained in Part I of the lectures (see, e.g., Chapter 2), thus providing information regarding the risk-neutral probability of the events in which they specifically pay off. While these assets do not obviously trade, we can extract their shadow value from the price of fixed income securities based on a model; we can then use these extracted values to price any interest rate derivative.

We center around these themes, which we illustrate through simple numerical applications including the pricing of interest rate derivatives such as options on bonds, swaps, caps, callable or convertible bonds, emphasizing the joint behavior of derivative prices and the underlying—for example, the derivative price predicted by a given model can tell us about whether the underlying is mispriced. The framework of analysis is in discrete time, and relies on the same "implied binomial trees" utilized in Chapter 11 to model equity derivatives. Implied binomial trees simplify many conceptual intricacies; Chapter 13 deals with more advanced topics in interest rate modeling and derivative evaluation, including the empirical motivation underlying them as well as a systematic analysis through continuous time methods. Finally, this chapter does not cover credit risk, which is the focus of Chapter 14. We now proceed with a number of basic pieces of motivation, explaining more in detail a few of the very issues arising in fixed income markets.

12.1.1 Relative pricing in fixed income markets

While current bond prices cannot be given a preference-free representation, we can still aim to price interest rate derivatives in a preference-free fashion. The keyword is "relative pricing"—the situation in which we price a number of assets given the price of others, while ensuring absence of arbitrage, as explained in many previous junctures of these lectures. Pricing options on traded assets is in general quite a different matter than pricing the underlying assets in the first place. Even in the equity case, we try to evaluate an option, say, in a preference-free fashion, even if the underlying asset cannot really. By "preference-free," we refer to the possibility to extract risk-neutral probabilities, or Arrow-Debreu prices, from the price of already traded assets, and price derivatives written on the states spanned by the extracted Arrow-Debreu prices.

We wish to achieve similar objectives in the fixed income space, although the task is challenging. Consider, for example, the Black & Scholes formula. The logic leading to it cannot exactly be applied to evaluate fixed income securities. Indeed, the Black & Scholes model relies on the assumption of a constant volatility of the underlying price. In the context of interest rate derivatives, the volatility of the underlying asset price depends, instead, on the maturity of the underlying, as it tends to zero as the time-to-maturity goes to zero.

More generally, pricing and hedging interest rate derivatives requires a model describing developments of the whole yield curve, as we shall explain. It goes without saying that the general principles underlying the APT are naturally still the same. The methods described in this and following chapter aim to an objective, where the dynamics of the "pricers" under the risk-neutral probability (e.g., the dynamics of the short-term rate) do not depend on risk-aversion corrections. The reader who is impatient about being referred to a concrete illustration of these somewhat surprising statements is referred to Eq. (12.42) in this chapter and the discussion around it, or in general, to Chapter 13 (see Section 13.6), where we describe the details of the celebrated Heath, Jarrow and Morton (1992) model.

12.1.2 Many evaluation paradigms

After reading this introductory chapter, you might find that so many methods and models might readily become available, which could be used to price interest rate derivatives. Derivatives houses do really have dozens of models, with different houses possibly asking for different prices relating to the same product. This circumstance needs not to be an arbitrage opportunity. Rate markets are typically over-the-counter, of the type studied in Chapter 11 of Part II. Indeed, this chapter relies on stylized examples of markets with different derivative prices predicted by

models that match the same set of initially given market prices. Market incompleteness and segmentation are responsible of these interesting features, as we shall explain.

With dozens of methods available to price fixed income products, we do not see the emergence of a "single" model to price all of the extant fixed income products. Typically, any bank has a battery of different models, with pieces of this battery possibly fighting for different goals. For example, a bank might display a preference for a certain type of models as a result of (i) its culture and history, or (ii) the particular business is pursuing. For example, in the next chapter, we shall see that to price interest rate options such as "caps," we may use the market model, which relies on the "Black 76" formula. However, using this model implies that we do not have a closed-form solution for the price of "swaptions," which can only be solved through numerical methods. If the swaptions business is not important for the bank then, we may safely adopt the market model.

[In progress]

12.1.3 Plan of the chapter

The chapter is organized as follows. Section 12.2 through 12.3 develop the basics underlying fixed income securities, such as interest rate and market conventions, duration, convexity, and an introduction to basic hedging and trading strategies. Section 12.4 is the first section to deal with models that aim to fit the initial yield curve without errors, using binomial trees. We have two fundamentals ways to achieve this goal. As for the first one, developed in Section 12.4, we freeze "scenarios," i.e. the values taken by the short-term rate on the branches of a tree, and search for the risk-neutral probs such that the prices predicted by the model agree with the market. As for the second, we fix risk-neutral probs, and search for "scenarios" such that the model and the market are the same. Section 12.5 deals with the second approach, which is the essence of the Ho and Lee (1986) model. Naturally, we may consider situations where we might simultaneously search for probabilities and short-term rate scenarios that make models consistent with the markets. These situations are quite complex and necessitate a general framework of analysis, developed in Section 12.6, and hinging upon calibration through Arrow-Debreu securities. Section 12.7 concludes this chapter and provides numerical examples of how to evaluate bonds with callability and convertibility features, which will receive a more systematic theoretical treatment in the next chapters.

12.2 Markets and interest rate conventions

12.2.1 Markets for interest rates

There are three main types of markets for interest rates: (i) LIBOR; (ii) Treasure rate; (iii) Repo rate (or repurchase agreement rate). We briefly survey these markets and, then review how to extract potentially useful information from them while constructing spreads.

12.2.1.1 LIBOR (London Interbank Offer Rate) and other interbank rates

Financial institutions trade deposits with each other, which span maturities that range from just overnight to one year at a given currency. The LIBOR rate reflects the rate at which financial institutions are willing to borrow in these markets, on average. It is an average indicative quote of the interbank lending market. It is determined as follows. A panel of major banks exists, where each bank belonging to the panel reports the rate it expects to be charged by its peers in

case of borrowing needs for a given maturity and a given currency. The four highest and the four lowest rates of these banks reports are ignored, and the remaining reports are aggregated by Thomson Reuters for ten currencies, and published daily by the British Bankers Association.[1] The left panel of Figure 12.1 plots the LIBOR referenced to US dollars.

The LIBOR is a fundamental point of reference to financial institutions, which look at it as an opportunity cost of capital. Moreover, many fixed income instruments are indexed to the LIBOR: forward rate agreements, interest rate swaps, or variable mortgage rates (see Chapters 13 and 14). Because the banks submissions over the process of the LIBOR formation do not rely on trades (only on subjective estimates of the cost of capital), there might be incentives for report manipulation biased towards the rates that would make a given bank reap profits over LIBOR-sensitive derivatives, as hypothesized over the LIBOR scandals emerging in June 2012.

While the LIBOR reflects market conditions inherent the banking system for a given currency, the US Federal Funds rate more directly links to the liquidity to be deposited within the Federal Reserve. Banks have to maintain reserves with the Federal Reserve to partially back deposits and to clear financial transactions, as further explained in Section 14.6 of Chapter 14. Transactions involve banks with excess reserves with the Fed, which earn no interest, to banks with reserve deficiencies. The Federal Funds rate is the overnight rate at which banks lend these reserves to each other. It is affected by the FDRBNY, which aims to make it lie within a range of the target rate decided by the governors at Federal Open Market Committee meetings. This range is "maintained" through open market operations.

12.2.1.2 Treasury rate

It is the rate at which a given Government can borrow at a given currency. The left panel of Figure 12.1 depicts the time behavior of the interest rate on short-term and long term US government debt: the 3 month T-bill rate and the 10 year Treasury yield.

12.2.1.3 Repo rate (or repurchase agreement rate)

A Repo agreement is a contract by which one counterparty sells some assets to another, with the obligation to buy these assets back at some future date. The assets act as collateral. The rate at which such a transaction is made is the repo rate. One day repo agreements give rise to *overnight repos*. Longer-term agreements give rise to *term repos*.

12.2.1.4 Interest rate spreads

Interest rate spreads are the difference between interest rates applying to two different markets. Thus, they have the potential to remove components that are incorporated by both markets, thereby isolating interesting pieces of information.

One important example is the LIBOR-OIS spread, which is the difference between the 3-month LIBOR minus the 3-month OIS. The OIS ("overnight indexed swap") rate is the swap rate in a swap agreement of fixed against variable interest rate payments, where the variable interest rate is an overnight reference, typically an average, unsecured interbank overnight rate, such as the Federal Funds rate in the US, SONIA in the UK or EONIA in the Euro area.[2] The LIBOR reflects the monetary policy stance but should also incorporate a premium related to counterparty risk. Instead, the OIS is a mere interest rate swap; as such, it should primarily

[1] Instead, the LIBID (London Interbank Bid Rate) is the rate that these financial institutions are prepared to pay to borrow money, on average, but it does not rely on a formal setting procedure such as that leading to the daily LIBOR. Naturally, the LIBID is less than the LIBOR.

[2] See the next chapter (Section 13.8.5) for extensive discussions regarding interest rate swaps.

reflect the monetary policy stance. Therefore, the LIBOR-OIS spread has the potential to isolate credit views on financial institutions. Historically, the LIBOR-OIS spread has behaved quite flat, although then it reached high record levels during the 2007 subprime crisis (see the right panel of Figure 12.1).

Another example of an interest rate spread commonly used in empirical research is the so-called TED spread, which is the difference between the LIBOR and the Treasury bill rate. The TED spread captures "flight to quality" effects that typically occur during times of crisis, when Treasuries are considered particularly valuable by investors (see the right panel of Figure 12.1). Due to this "flight to quality" reason, the TED spread might fail isolate views about developments in the interbank market.

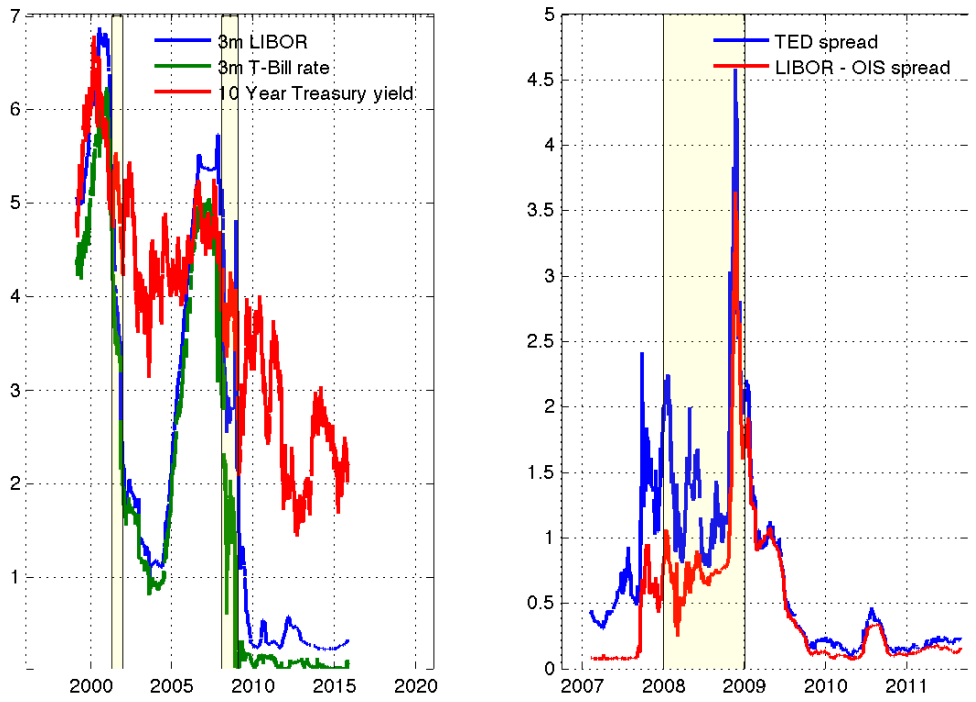


FIGURE 12.1. Left panel: The 3m LIBOR referenced to USD, the 3m T-Bill rate and the 10 year Treasury yield. Right panel: The TED spread (the difference between the 3m LIBOR and the 3m T-Bill rate) and the LIBOR-OIS spread (the difference between the 3m LIBOR and the 3m OIS). The shaded areas mark recession periods identified by the National Bureau of Economic Research.

On a historical note, the Federal Funds rate has been the object of much empirical research. In an attempt to explain how the "credit view" contributes to growth more than Friedman's monetary view, Bernanke and Blinder (1992) show that the Federal Funds rate makes the predicting power of M1 growth insignificant, as we further review in Section 13.6 of Chapter 13. This finding initially spread enthusiasm about the ability of this rate to explain short-run aggregate fluctuations. However, as surveyed for example by Stock and Watson (2003), the

explanatory power of the Federal Funds rate evaporizes, once we condition on the term spread, a fact we comment in Section 13.2.2 of the next chapter.

12.2.2 Mathematical definitions of interest rates

12.2.2.1 Simply compounded interest rates

A simply compounded interest rate at time τ, for the time interval $[\tau, T]$, is defined as the solution L to the following equation:

$$P(\tau, T) = \frac{1}{1 + (T - \tau) L(\tau, T)}. \tag{12.1}$$

This definition is intuitive, and is the most widely used in the market practice. For example, LIBOR rates can be defined consistently with this way, with $P(\tau, T)$ being the initial investment at τ that delivers £1 at T.

12.2.2.2 Yield curves

The *yield-to-maturity,* or *spot rate*, for some maturity date T is the yield on the zero maturing at T, denoted as $r(t, T)$. It is the solution to the following equation,

$$P(t, T) = \frac{1}{(1 + r(t, T))^T}. \tag{12.2}$$

With semi-annual compounding, we have that $P(t, T) = (1 + \frac{r(t,T)}{2})^{-2T}$. In general, we have that if the interest is compounded m times in a year, at the annual rate r, then, investing for T years gives $\left(1 + \frac{r}{m}\right)^{mT}$. Continuous compounding is obtained by letting $m \to \infty$ in the previous expression, leaving e^{rT}. Therefore, the *continuously compounded* spot rate is obtained as:

$$R(t, T) = -\frac{1}{T} \ln P(t, T).$$

It is a sort of "average rate" for investing from time t to time $T > t$. The function, $T \mapsto R(t, T)$, is called the *yield curve*, or the *term structure of interest rates.*

A related and widely used concept is the *par yield curve.* Let $B(t, T)$ be the time t price of a bond that pays off the principal of £1 at expiry T, and a known sequence of constant coupons $C(t, T)$ at $t + 1$, $t + 2, \cdots, T$, such that, in the absence of arbitrage and any other frictions,

$$B(t, T) = C(t, T) \cdot \sum_{i=1}^{T-t} P(t, t + i) + P(t, T).$$

Note that $C(t, T)$ is fixed at time t. A par bond is one that quotes at parity, $B(t, T) = 100\%$. The par yield curve is the sequence of coupon rates $C(t, T)$, for T varying, that correspond to the par:

$$C(t, T) = \frac{B(t, T) - P(t, T)}{\sum_{i=1}^{T-t} P(t, t + i)}, \quad B(t, T) = 1. \tag{12.3}$$

In other words, the coupon rates $C(t, T)$ have to "adjust" to make the market happy to have the coupon bearing bond quote at par, $B(t, T) = 1$. An interesting interpretation of par-yield is the following. Rewrite Eq. (12.3) as follows: $1 - P(t, T) = C(t, T) \cdot \sum_{i=1}^{T-t} P(t, t + i)$. The

right-hand side of this equation is the present value of the flow of known coupons, $C(t,T)$, receivables at the dates $t+1$, $t+2,\cdots,T$. The left-hand side is the present value of the flow of future, and unknown, LIBOR rates, $L(t+1-i,t+i)$ for $i=1,2,\cdots,T-1$, receivable at the dates $t+1$, $t+2,\cdots,T$,

$$\sum_{i=1}^{T-t}\text{Val}\left(\frac{1}{P(t+i-1,t+i)}-1\right)=\sum_{i=1}^{T-t}\left(P(t,t+i-1)-P(t,t+i)\right)=1-P(t,T),$$

where $\text{Val}(x)$ denotes the current price of receiving a random amount of x dollars at the maturity date where these are due, and where we have used (i) the definition of the LIBOR rate, as defined in Eq. (12.1) and, (ii) a no-arbitrage relation, which we shall show in Chapter 13 (see Section 13.1), namely that the present value of $1/P(t+i-1,t+i)$ paid at $t+i$ is simply $P(t,t+i-1)$.

Therefore, we can interpret the par yield as the fixed rate in a swap contract that costs nothing at origination, where one counterparty pays another counterparty, the fixed rate against the variable LIBOR—a *spot* swap rate. These swap contracts, of which Section 12.5.4.2 gives a numerical example, are analyzed in detail in Chapter 13 (see Section 13.7.5).

12.2.2.3 Forward rates

In a *forward rate agreement* (FRA, henceforth), two counterparties agree that the interest rate on a given principal (say £1) in a future time-interval $[T,S]$ will be fixed at some level K. The FRA works as follows: at time T, the first counterparty receives £1 from the second counterparty; at time $S>T$, the first counterparty pays back $£\left[1+1\cdot(S-T)K\right]$ to the second counterparty. The amount K is agreed upon at time t. Therefore, the FRA makes it possible to lock-in future interest rates. We consider simply compounded interest rates because this is the standard market practice.

The amount K for which the current value of the FRA is zero is called the *simply-compounded forward rate* as of time t for the time-interval $[T,S]$, and is usually denoted as $F(t,T,S)$. We can use absence of arbitrage to express $F(t,T,S)$ in terms of bond prices, as follows:

$$\frac{P(t,T)}{P(t,S)}=1+(S-T)F(t,T,S). \tag{12.4}$$

Indeed, an investor in a zero from time t to time S is one who simultaneously makes (i) a *spot* loan from t to T, and (ii) a *forward* loan from T to S. In the absence of arbitrage, it must be the case that,

$$\underbrace{\left[1+r(t,S)\right]^{S-t}}_{\text{zero loan}}=\underbrace{\left[1+r(t,T)\right]^{T-t}}_{\text{spot loan}}\times\underbrace{\left[1+(S-T)F(t,T,S)\right]}_{\text{forward loan}},$$

where $r(t,S)$ is the (discretely compounded) spot rate at time t for maturity S. Eq. (12.4) follows by the definition of $r(t,S)$ in Eq. (12.2).

Alternatively, consider the following portfolio implemented at time t. Go long one bond maturing at T and short $P(t,T)/P(t,S)$ bonds maturing at S, for the time period $[t,S]$. The initial cost of this portfolio is zero because,

$$-P(t,T)+\frac{P(t,T)}{P(t,S)}P(t,S)=0.$$

At time T, the portfolio yields £1, which originates from the bond purchased at time t. At time S, we buy the $P(t,T)/P(t,S)$ bonds shorted at t, and maturing at the very same S, which obviously costs £ $P(t,T)/P(t,S)$. The portfolio, therefore, is acting as a FRA: it pays £1 at time T, and $-$£ $P(t,T)/P(t,S)$ at time S. In addition, the portfolio costs nothing at time t. Therefore, the interest rate implicitly paid in the time-interval $[T,S]$ must be equal to the forward rate $F(t,T,S)$, as stated in Eq. (12.4).

12.2.3 Yields to maturity on coupon bearing bonds

Finally, the yield to maturity $\hat{y}$ (YTM, henceforth) on a bond is simply its rate of return. It is the single discount rate equating the present value of the payoff stream to its market price:

$$\hat{y} : B(0,T) = \sum_{i=1}^{n} \frac{C_{t_i}}{(1+\hat{y})^{t_i}} + \frac{1}{(1+\hat{y})^{T}}, \tag{12.5}$$

where $t \equiv 0 \leq t_1 < \cdots < t_n \equiv T$. Eq. (12.5) differs from the price formula $B(0,T) = \sum_{i=1}^{n} \frac{C_{t_i}}{(1+r(0,t_i))^{t_i}} + \frac{1}{(1+r(0,T))^{T}}$, because the same discount rate $\hat{y}$ is used to discount all future payments. Clearly, spot rates coincide with the YTM on a zero, i.e. $\hat{y} = R(0,T)$.

Next, suppose that coupon payments are the same for each i, $C_{t_i} = \bar{C}$ say, and that the payment dates are set regularly. Eq. (12.5) then collapses to

$$B(0,T) = \left(1 - Z^T\right)\frac{\bar{C}}{\hat{y}} + Z^T, \quad Z^T \equiv (1+\hat{y})^{-T}. \tag{12.6}$$

That is, the price of the coupon bearing bond, $B(0,T)$, is a combination of the value of a perpetuity, $\frac{\bar{C}}{\hat{y}}$, and that of a zero expiring at T, Z^T. For large maturities, the bond value gets closer to that of a perpetuity; for low maturities, the bond value is closer to that of a zero. If $\bar{C} > \hat{y}$, $B(0,T) \in \left(1, \frac{\bar{C}}{\hat{y}}\right)$, and if $\bar{C} < \hat{y}$, $B(0,T) \in \left(\frac{\bar{C}}{\hat{y}}, 1\right)$. In the special case where $\bar{C} = \hat{y}$, the bond would quote at par.

The property to quote at par in this context arises as a special case of a more general characteristics of *floating rate bonds*. Floating rate bonds pay coupons equal to the LIBOR, and would quote at par at their first reset date (see Section 13.7.5 in the next chapter). In the context of this preliminary example, $B(0,T)$ in (12.6) can be understood as the price of a "floating" rate bond in a market without uncertainty: if the coupon it pays is the same as the interest rate we use to discount future cash flows, our coupon bearing bond is, in fact, a "floating" rate bond, and would therefore need to quote at par.

We didn't specify how the relations introduced so far obtain as a result of absence of arbitrage and we didn't cover material regarding how to extract information on the yield curve while solely relying on the quotes of coupon-bearing bonds. Appendix 1 provides standard material on topics such as "bootstrapping" (a way to determine the yield curve based on coupon-bearing bonds) as well as examples of how to implement trades to profit from temporary arbitrage opportunities arising in connection with misaligned coupon-bearing bond prices.

12.2.4 Accruals, invoice, and clean prices on coupon bearing bonds

How do we price coupon bearing bonds that trade at any date after issuance? Naturally, the price to pay at any date is still the present value of future coupons although, then, we still need

to cope with issues relating to the discreteness of coupon payments. By absence of arbitrage, the price of a coupon bearing bond is

$$B_t(T_N) = \sum_{i_t=1}^{N} \frac{C_i}{n} P_t(T_i) + P_t(T_N),$$

where $\frac{C_i}{n}$ is a sequence of coupons paid off over some dates T_i, $i = 1, \cdots, N$, and n is the frequency of coupon payments. For example $n = 2$ for semiannual coupon payments, in which case $T_i - T_{i-1} = \frac{1}{2}$ and in general, $T_i - T_{i-1} = \frac{1}{n}$. Finally, we define T_{i_t} as the first available coupon payment date a bondholder would have access to after time t, i.e. $t : T_{i_t-1} \leq t < T_{i_t}$.

Due to the discreteness of coupon payments, discontinuities arise: at time T_{i_t}, the coupon is paid off, which leads to a discrete drop in the bond price. Precisely, we have:

$$B_t(T_N) = \begin{cases} \frac{C_{i_t}}{n} P_t(T_{i_t}) + \sum_{i=i_t+1}^{N} \frac{C_i}{n} P_t(T_i) + P_t(T_N), & \text{for } t \in [T_{i_t-1}, T_{i_t}) \\ \sum_{i=i_t+1}^{N} \frac{C_i}{n} P_t(T_i) + P_t(T_N), & \text{for } t = T_{i_t} \end{cases} \tag{12.7}$$

such that the discrete price drop at time T_{i_t} is equal to,

$$\frac{C_{i_t}}{n} P_t(T_{i_t}) = \frac{C_{i_t}}{n}, \quad \text{for } t = T_{i_t}.$$

It is market practice to avoid consider the effects relating to this drop in value while quoting bond prices. Define the "accruals" as that idealized portion of the coupon payments that occurs between T_{i_t-1} and t,

$$\text{Accr}_t \equiv \begin{cases} \frac{C_{i_t}}{n} \frac{t - T_{i_t-1}}{T_{i_t} - T_{i_t-1}} & \text{for } t \in [T_{i_t-1}, T_{i_t}) \\ 0, & \text{for } t = T_{i_t} \end{cases}$$

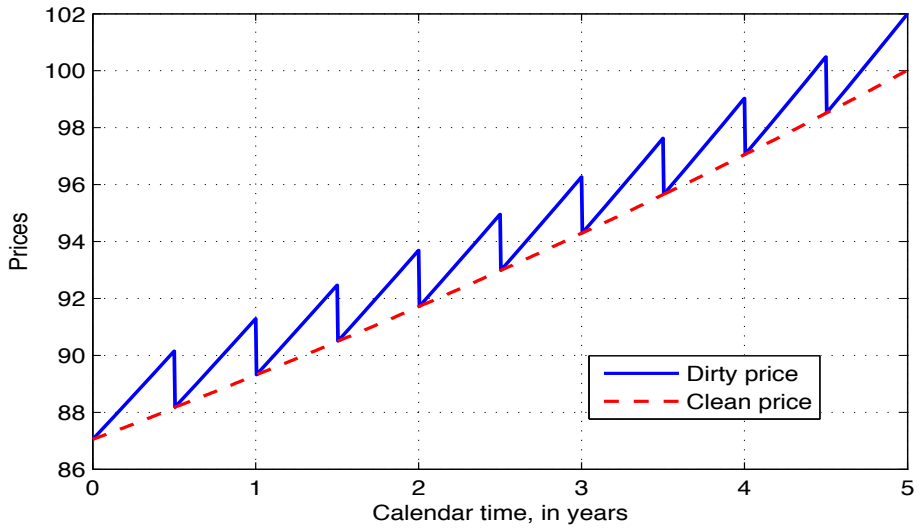
Define a "clean price" as the difference between the "dirty price" $B_t(T_N)$ and the accruals,

$$B_t^c(T_N) \equiv B_t(T_N) - \text{Accr}_t = \frac{C_{i_t}}{n} \left(P_t(T_{i_t}) - \frac{t - T_{i_t-1}}{T_{i_t} - T_{i_t-1}} \right) + \sum_{i=i_t+1}^{N} \frac{C_i}{n} P_t(T_i) + P_t(T_N). \tag{12.8}$$

It is easy to see that the clean price, $B_t^c(T_N)$, does not experience any jumps over the reset dates—the accruals drop by $\frac{C_{i_t}}{n}$, which is precisely the amount the dirty price drops by, such that the first term in the equality in Eq. (12.8) is zero. Therefore, dirty and clean prices are the same at each coupon payment.

The dirty price, $B_t(T_N)$, is sometimes referred to as "full price" or "invoice," reflecting the fact that it is the price at which buyers and sellers have to settle a transaction. US traders typically quote clean prices, because dirty prices mainly reflect mechanical adjustments relating to coupon payments, whereas clean prices are more sensitive to changes in the general macroeconomic or the fixed income markets outlook, as illustrated by a basic example below. However, on a transaction perspective, the real object of interest is the dirty price—$B_t(T_N)$ is the price to be paid originating from the bond transaction.

The following picture depicts the typical shark-tooth pattern of the dirty price in an overly simplified world, where $C_i = 4$, annualized, $n = 2$, the yield curve is flat at the continuously compounded rate $r = 7\%$, and the maturity of the bond equals $T = 5$ years.



The sharp increase in the dirty price is mainly due to the term $2 \times P_t(T_{i_t})$ in Eq. (12.7). This term increases roughly linearly over the payment dates, and overwhelms the increase in value of the second term of Eq. (12.7). The drop is precisely 2 at each payment date, where the dirty price equals the clean price.

The pattern in the picture is of course not unique. For example, the clean price decreases over time when the yield curve is flat at a continuosly compounded $r = 2\%$. When the level of the yield curve is high, future coupons are heavily discounted such that the value of them is low when time to maturity is high. When, instead, the level of the yield curve is low, the current value of future coupons is high, and the clean price converges to 100 from above. We can determine a threshold value for r, say $\bar{r}$, such that the clean price is approximately constant at 100, using Eq. (12.6). Because we assume the yield curve is flat at r continuously compounded, and $n = 2$, we need to find $r : e^{\bar{r}} = (1 + \frac{\hat{y}}{2})^2$, with $\hat{y} = 2\%$, whence $\bar{r} = 3.96\%$.

12.3 Duration and convexity hedging and trading

The risk of a default-free bond is that future spot rate could change. In other words, the bond required return (i.e., the YTM) can obviously change in the future. Consider the definition of the YTM $\hat{y}$ in Eq. (12.5), and introduce the following function,

$$B(y;T) \equiv \sum_{i=1}^{n} \frac{C_{t_i}}{(1+y)^{t_i}} + \frac{1}{(1+y)^{T}}. \tag{12.9}$$

This function "mimics" how the market price $B(T)$ would behave after the initial YTM $\hat{y}$ changed to some value y and, naturally, is such that, $B(\hat{y};T) = B(T)$. The function $B(y;T)$

is thus the simplest bond pricing model we could ever formulate. It is perhaps too simple, as it does not rely on absence of arbitrage. Nevertheless, we can use this very preliminary model to say something about interest rate risk.

We define a measure of risk of the bond based on the *sensitivity* of the hypothetical bond price $B(y;T)$ with respect to changes in y. We answer the following question: What happens to the hypothetical bond price $B(y;T)$, once we perturb the one rate $\hat{y}$ that discounts all the payoffs? The sensitivity we are dealing with is simply the first partial of the "bond-pricing formula" in Eq. (12.9) with respect to y,

$$B_y(y;T) = -\frac{1}{1+y}\left[\sum_{i=1}^{n}\frac{t_i \cdot C_{t_i}}{(1+y)^{t_i}} + \frac{T \cdot 1}{(1+y)^{T}}\right],$$

where the subscript denotes a partial derivative, i.e. $B_y(y;T) = \frac{\partial}{\partial y}B(y;T)$. The sensitivity, $B_y(y;T)$, is the tangent to the price-yield relation, which Figure 12.2 illustrates in the case of a zero-coupon bond with time to maturity equal to 10 years.

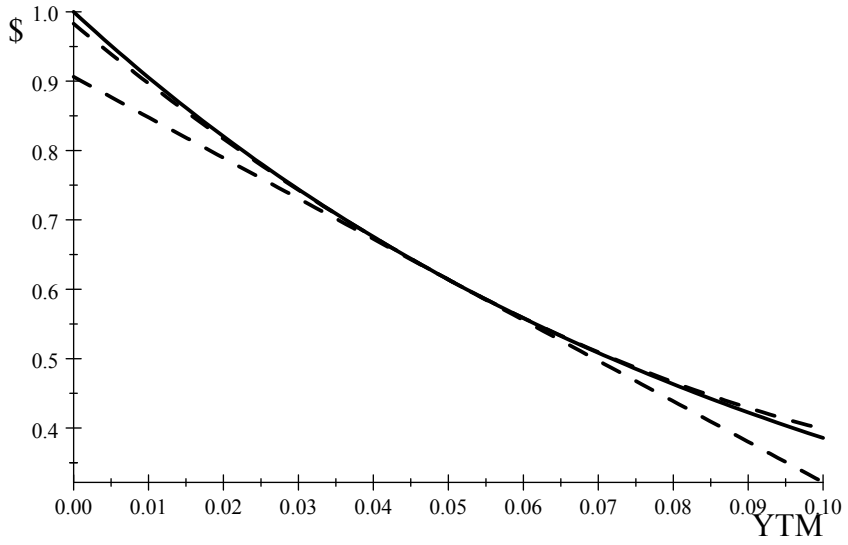


FIGURE 12.2. The relation between the YTM and the bond price, and its first-order (duration) and second-order (convexity) approximations. The solid line depicts the price of a zero coupon bond expiring in 10 years, as a function of the YTM, $(1+\text{YTM})^{-10}$, and the two dashed lines are first-order and second-order Taylor's expansions around YTM $= 5\%$.

12.3.1 Duration

We define the "Macaulay duration" as,

$$\text{D}_{\text{Mac}} \equiv \frac{-B_y(y;T)}{B(y;T)}(1+y) = \sum_{i=1}^{n}\omega_{t_i} \cdot t_i + \hat{\omega}_T \cdot T,$$

where

$$\omega_{t_i} = \frac{C_{t_i}/\left(1+y\right)^{t_i}}{B\left(y;T\right)}, \quad \hat{\omega}_T = \frac{1/\left(1+y\right)^{T}}{B\left(y;T\right)}.$$

In words, the Macaulay duration is a weighted average of the payment dates. The weights ω_{t_i} are the discounted coupons at the various payment dates, $C_{t_i}/\left(1+y\right)^{t_i}$, related to the current market value of these coupons, i.e. the bond price $B\left(y;T\right)$ when the YTM is y. That is, the weights are the proportions of the bond's present value that is attributable to the payoff at date t. The weights satisfy $\sum_{i=1}^{n}\omega_{t_i}+\hat{\omega}_T=1$. Therefore, $\mathrm{D_{Mac}}\leq T$. The Macaulay duration is a measure of how far in the future the bond pays off. For zeros, $\mathrm{D_{Mac}}=T$.

For small y, $\mathrm{D_{Mac}}\left(y\right)$ is simply the semi-elasticity of the bond price with respect to the YTM. This semi-elasticity is also referred to as "modified duration":

$$\mathrm{D} \equiv \frac{-B_y}{B} = \frac{\mathrm{D_{Mac}}}{1+y}.$$

A simple computation reveals that the modified duration, D, satisfies: $\frac{\partial \mathrm{D}}{\partial y} = \frac{-B_{yy}}{B} + \left(\frac{B_y}{B}\right)^2$. Therefore, the modified duration is *decreasing* in the YTM when the bond price is sufficiently convex in the YTM, which is surely the case for long-term maturity dates. Interestingly, the modified duration is *increasing* in the YTM when the bond price is concave in the YTM, a property featured by callable bonds and mortgage-backed securities (MBS, henceforth), as explained in the next chapters (moreover, see Section 12.8.1 below for a basic numerical example). Intuitively, the incentives to proceed to early repayments "kick in" as the YTM decreases, which makes the duration of the MBS decrease.

The Macaulay duration for *continuously compounded rates* is even simpler to calculate. First, define the continuously compounded YTM as the single number $\hat{x}$ such that

$$B(\hat{x};T) = \sum_{i=1}^{n} c_{t_i} e^{-\hat{x}\cdot t_i} + e^{-\hat{x}\cdot T},$$

where $B(\hat{x};T)$ is the market price of a bond paying off the principal of one at maturity and the stream of payoffs c_{t_i}. Next, consider, the function $x \mapsto B\left(x;T\right)$. Define the semi-elasticity of the bond price $B\left(x;T\right)$ with respect to the continuously compounded YTM x,

$$\frac{-B_x\left(x;T\right)}{B\left(x;T\right)} = \frac{\sum_{i=1}^{n} c_{t_i} t_i e^{-x\cdot t_i} + T\cdot e^{-x\cdot T}}{B\left(x;T\right)} = \sum_{i=1}^{n} w_{t_i}\cdot t_i + \hat{w}_T \cdot T,$$

where $B_x\left(x;T\right) = \frac{\partial B(x;T)}{\partial x}$, $w_{t_i} = \frac{c_{t_i} e^{-x\cdot t_i}}{B(x;T)}$ and $\hat{w}_T = \frac{e^{-x\cdot T}}{B(x;T)}$. Note, the weights are such that $\sum_{i=1}^{n} w_{t_i} + \hat{w}_T = 1$. Therefore, the "Macaulay duration" for continuously compounded rates is equal to the semi-elasticity of the bond price with respect to the continuously compounded YTM x.[3] This result may simplify some calculations.

[3]Mathematically, we could have obtained this result in a straightforward manner, as follows. Define the bond price function as $B\left(y\left(x\right)\right)$, where by definition, $y\left(x\right)=e^x-1$. Hence, $B_x\left(y\left(x\right)\right) = B_y\left(y\left(x\right)\right)y'\left(x\right) = B_y\left(y\left(x\right)\right)e^x = B_y\left(y\left(x\right)\right)\left(1+y\right)$. It follows that $\mathrm{D_{Mac}} = \frac{-B_y(1+y)}{B} = \frac{-B_x}{B}$.

12.3.2 Convexity

Convexity measures how the sensitivity, B_y, changes with y. It is the second partial of the bond price with respect to y, B_{yy}. Positive convexity means that the interest rate sensitivity declines as y increases, as in Figure 12.2. This properties arises because $\frac{\partial}{\partial y}(-B_y) = -B_{yy} < 0$. Formally, convexity is defined as,

$$\mathrm{C} \equiv \frac{B_{yy}}{B},$$

such that we can expand the bond price as follows:

$$\frac{\Delta B}{B} \approx -\mathrm{D} \cdot \Delta y + \frac{1}{2}\mathrm{C} \cdot (\Delta y)^2 . \tag{12.10}$$

That is, for very "convex securities", duration may not be a safe measure of return, as Figure 12.2 illustrates.

12.3.3 Asset-liability management

12.3.3.1 Introductory issues

We can use duration to assess how exposed a bond portfolio is to movements in the interest rates and, then, "immunize" this portfolio to interest rates changes. Duration is relevant to asset-liability management. For example, pension funds have known streams of liabilities that must be matched by the assets they hold. In words, the duration of the assets must equal the duration of the liabilities. For example, in the UK, pension funds must mark-to-market the liabilities. Therefore, one objective of these funds is to immunize their liabilities against movements in the interest rates.

Alternatively, consider the following basic example. A bank borrows £100 at 2% for a year and lends this money at 4% for 5 years, where the higher rate compensates for a variety of factors such as business risk or the bank's market power. Assuming that the bank's borrower does not default, the bank generates profits equal to $£(4\% - 2\%) \cdot 100 = 2$ in the first year, and according to its books. However, the appropriate assessment of the current situation should not make reference to past market conditions but, obviously, current. Suppose, for example, that after one year, the interest rate for borrowing increases from 2% to 5%, and remains such for 4 additional years. This assumption is unrealistic, but it gives the idea of where the action is. The market value of the assets is, then, $\frac{100 \cdot 1.04^5}{1.05^4} = 100.09$. Note, we discount at the 5% rate, as this is the cost of capital for the bank.[4] The market value of the liabilities is, instead, $100 \cdot 1.02 = 102$. The bank's problem is a duration mismatch.

Let us return to the pension fund example, and consider the following extreme case. In 30 years from now, a pension fund is due to deliver £100,000 to some future retiree. Suppose the current market situation is such that the yield curve is flat at 4%, such that the market value of this liability is $£100,000 \cdot (1.04)^{-30} = £30,832$. Accordingly, the would-be retiree invests £30.832 in the pension fund. So we have the following situation:

Cash	Pensions
£30,832	£30,832

[4]Suppose, for example, that the bank wants to borrow £102 to pay off its liabilities, and for 4 additional years, then the profit at time 1 is $\frac{100(1.04)^5 - 102(1.05)^4}{(1.05)^4} = -1.9057$. Alternatively, the 5% interest rate is just an opportunity cost of capital, defined as max {borrowing cost, lending rate}, where the borrowing cost is that the bank might obtain from other banks, for example.

Suppose, now, that the pension fund does not invest this cash. This strategy is of course inefficient, but it is precisely the point of this exercise to see why it is so.

Consider two extreme cases, occurring under two scenarios underlying developments in the fixed income market. In one week,

(i) Scenario ↑: the yield curve shifts up parallely to 5%. Accordingly, the value of the liability for the pension fund becomes: $£100,000 \cdot (1.05)^{-30} = 23,138$.

Cash	Profit
$£30,832$	$£7,694$
	Pensions
	$£23,138$

(ii) Scenario ↓: the yield curve shifts down parallely to 3%. Accordingly, the value of the liability for the pension fund is: $£100,000 \cdot (1.03)^{-30} = 41,199$.

Cash	Loss
$£30,832$	$-£10,367$
	Pensions
	$£41,199$

A drop in the yield curve results in a loss for the pension fund: when interest rates go down, the pension fund faces a challenging situation as it has to honour its obligations in 30 years, but the financial market "yields less" than it promised one week earlier. Naturally, the pension fund would face the opposite situation were interest rates to go up, as in the first scenario above.

There are many good ethical reasons we might dislike pension funds have to experience interest rate volatility. The volatility in this very basic example stems from the simple fact the pension fund receives $£30,832$, which it then puts "under the pillow." The most efficient way to erase volatility could have been to invest $£30,832$ in a 30 year bond—at the market conditions of 4%. This is perfect hedging, which relies on the assumption we have access to such a long-term bond. How do we deal with situations in which we do not have access to such a bond? The next sections illustrate these cases.

12.3.3.2 Hedging

Let us consider a portfolio of two bonds with different durations. Its value is given by,

$$V = \theta_1 B_1(\hat{y}_1) + \theta_2 B_2(\hat{y}_2),$$

where $B_1(\hat{y}_1)$ and $B_2(\hat{y}_2)$ are the market value of the bonds, $\hat{y}_1$ and $\hat{y}_2$ are the YTM on the bonds and, finally, θ_1 and θ_2 are the quantities of bonds in the portfolio. Let us consider a small change in the two YTM $\hat{y}_1$ and $\hat{y}_2$. We have,

$$dV = -\left[\theta_1 \mathrm{D}(\hat{y}_1) B_1(\hat{y}_1) d\hat{y}_1 + \theta_2 \mathrm{D}(\hat{y}_2) B_2(\hat{y}_2) d\hat{y}_2\right].$$

The question is: How should we choose θ_1 and θ_2 such that the value of the portfolio remains the same, even after a change in $\hat{y}_1$ and $\hat{y}_2$?

Let us assume a *parallel shift* in the term structure of interest rates. In this case, $d\hat{y}_1 = d\hat{y}_2$. The portfolio is said to be *immunized* if its value V does not change when $\hat{y}_1$ and $\hat{y}_2$ change, i.e. $dV = 0$, which is true when,

$$\theta_1 = -\frac{\mathrm{D}\left(\hat{y}_2\right) B_2\left(\hat{y}_2\right)}{\mathrm{D}\left(\hat{y}_1\right) B_1\left(\hat{y}_1\right)}\theta_2. \tag{12.11}$$

A useful interpretation of this portfolio is that we may be holding a bond with some duration, say we hold θ_2 units of the second bond. Given these holdings, we may wish to sell another bond, possibly with a lower duration, to hedge against movements in the price of the bond we hold.

Alternatively, we can think of the second asset as a *liability*, with a value that fluctuates after interest rates change. Then, we may wish to purchase some *asset* to hedge against the liability. Mathematically, $\theta_2 < 0$ and $\theta_1 > 0$. Moreover, Eq. (12.11) reveals that the number of assets to hold to hedge against the liability is high if the ratio of the two durations of the assets, $\mathrm{D}\left(\hat{y}_2\right)/\mathrm{D}\left(\hat{y}_1\right)$, is large. In this case, the hedging position is obviously inefficient. Asset-liability management, and "immunization", is costly when we hedge high-duration liabilities with low duration assets. We now illustrate these claims through a few basic examples.

12.3.3.3 A first example: hedging zeros with zeros

Suppose that we hold one bond, a zero with maturity equal to 5 years. We want to hedge this risk through another bond, a zero with maturity equal to 1 year. Let us assume that the term-structure is flat at 5%, discretely compounded. Then,

$$\begin{aligned} B_1\left(\hat{y}_1\right) &= \frac{1}{1+\hat{y}_1} = \frac{1}{1+0.05} = 0.95238, \quad \mathrm{D}\left(\hat{y}_1\right) = \frac{\mathrm{D}_{\mathrm{Mac}}\left(\hat{y}_1\right)}{1+\hat{y}_1} = \frac{1}{1+0.05} = 0.95238 \\ B_2\left(\hat{y}_2\right) &= \frac{1}{\left(1+\hat{y}_2\right)^5} = \frac{1}{\left(1+0.05\right)^5} = 0.78353, \quad \mathrm{D}\left(\hat{y}_2\right) = \frac{\mathrm{D}_{\mathrm{Mac}}\left(\hat{y}_2\right)}{1+\hat{y}_2} = \frac{5}{1+0.05} = 4.7619 \end{aligned}$$

and:

$$\theta_1 = -\frac{\mathrm{D}\left(\hat{y}_2\right) B_2\left(\hat{y}_2\right)}{\mathrm{D}\left(\hat{y}_1\right) B_1\left(\hat{y}_1\right)}\theta_2 = -\frac{4.7619 \cdot 0.78353}{0.95238 \cdot 0.95238} \cdot 1 = -4.1135.$$

That is, to hedge the 5Y zero, we need to short-sell approximately four 1Y zeros. The balance of this hedging position is,

$$B_1\left(\hat{y}_1\right)\theta_1 + B_2\left(\hat{y}_2\right)\theta_2 = (-4.1135) \cdot 0.95238 + 0.78353 = -3.1341. \tag{12.12}$$

This is quite an inefficient hedging position. One reason it is inefficient is that hedging long-term bonds with short implies we should rebalance too often. Moreover, as time goes on, the sensitivity of the short-term bonds to changes in the YTM is very small (eventually, the price equals face value plus coupon, at maturity), compared to that of long-term bonds. Therefore, rebalancing becomes increasingly severe as time unfolds.

Next, we study how the value of this portfolio changes after large changes in the YTM. By the assumption that the initial term-structure is flat at 5%, $\hat{y}_1 = \hat{y}_2 = 5\%$. Moreover, by rearranging Eq. (12.12),

$$B_2\left(y = 5\%\right) = 4.1135 \cdot B_1\left(y = 5\%\right) - 3.1341. \tag{12.13}$$

The left hand side of Eq. (12.13) is the price of the 5Y bond. The right hand side is the value of a "replicating" portfolio, which consists of (i) approximately 4 units of the 1Y bond, and (ii) the balance of the hedging position. Precisely, the right hand side is a net obligation: the value of the assets we need to purchase back (approximately 4 units of the 1Y bond), net of some cash we already have (£3.1341), which we can use to partially purchase these assets. In other words, we can interpret this position as a trade, where we buy the 5Y bond and sell approximately four 1Y bonds at some initial time t_0. Come some time $t_1 > t_0$, we liquidate, by reversing the position.

If interest rates do not change, then, approximately, and abstracting from passage of time, there will be no profits or losses, once we liquidate, or mark-to-market, this position. If interest rates change, $y \neq 5\%$, Eq. (12.13) can only approximately hold,

$$B_2(y) \approx 4.1135 \cdot B_1(y) - 3.1341.$$

Figure 12.3 plots the left hand side and the right hand side of this relation.

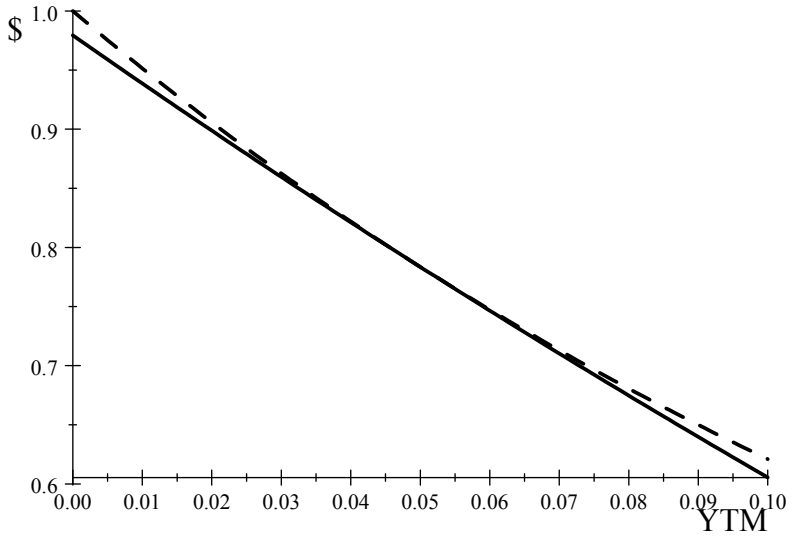


FIGURE 12.3. Dashed line (top): The price of the 5Y zero, $B_2(y) = \frac{1}{(1+y)^5}$, where y is the YTM. Solid line (bottom): The value of the "replicating" portfolio consisting of (i) 4.1135 units of the 1Y zero, and (ii) the balance of the hedging position, which is equal to $-$£3.1341, i.e. $4.1135 \cdot B_1(y) - 3.1341$, where $B_1(y) = \frac{1}{1+y}$ is the 1Y zero price.

What is going on? We are hedging the 5Y zero by selling approximately four 1Y zeros. In a neighborhood of $y = 5\%$, the value of the "synthetic" 5Y zero we sold, $4.1135 \cdot B_1(y) - 3.1341$, behaves as $B_2(y)$. However, the 5Y zero displays more convexity than the "synthetic" bond. This larger convexity implies that:

- If interest rates go down, the price of the 5Y zero bond we hold increases more than the value of the "synthetic" bond we sold.

- If interest rates go up, the price of the 5Y zero bond we hold decreases less than the value of the "synthetic" bond we sold.

In both cases, we make profits. Profits are, indeed, equal to $(B_2^+ - \theta_1 B_1^+) - (B_2 - \theta_1 B_1) \equiv \Delta B_2 - \theta_1 \Delta B_1$, where B_1^+ and B_2^+ are the prices of the 1Y and 5Y bonds. But by convexity, B_2 increases more than $\theta_1 B_1$ when interest rates go down, and B_2 decreases less than $\theta_1 B_1$ when interest rates go up, or $\Delta B_2 - \theta_1 \Delta B_1 \geq 0$.

Note that this is not an arbitrage opportunity! The previous reasoning hinges on the assumption of a parallel shift in the term-structure of interest rates, that is $d\hat{y}_1 = d\hat{y}_2$, where $\hat{y}_1 =$ spot rate for 1 year, and $\hat{y}_2 =$ spot rate for 5 years. While parallel shifts in the term-structure seem empirically relevant, they are not the only shifts that are likely to occur, as explained in the next chapter.

To sum up, duration hedging is a useful tool, but with quite important limitations. As Eq. (12.10) makes clear, duration is only a first-order approximation to the price of a bond. Moreover, duration hedging obviously requires rebalancing, which might be substantial. As we know, a conventional bond is strictly convex in the YTM. Therefore, for large changes in the YTM, the duration-based hedging ratios should be updated. Re-adjustments are in order anyway, independently of whether YTM change or not, as the duration of conventional fixed income securities obviously decreases over time.

12.3.3.4 Duration trading: Barbell and bullet hedges

As a second example of duration hedging, consider the "barbell" trading, which is a way to hedge some liability (a "bullet") with duration D_2 through two assets with durations D_1 and D_3, where $D_1 < D_2 < D_3$—that is a trade where we sell D_2 and buy D_1 & D_3. This trade is expected to work as soon as the yield curve flattens, with its short-end part not going high too much. Moreover, investing in the short-term segment of the yield curve, allows one to invest elsewhere relatively rapidly once the first asset expires, were the bond market to go down.

[*Mention differences with flatteners and steepeners*]

To illustrate a barbell trade, consider the example in the previous section, and suppose that another bond is available for trading, a zero with maturity equal to 10 years. We aim to hedge against movements in the price of the 5Y zero with a portfolio consisting of (i) one 1Y zero and (ii) the 10Y zero. We keep on assuming that the yield-curve is flat at 5%, and only consider parallel shifts in the term-structure of interest rates. We consider extensions below.

Such a "butterfly" trade can be implemented as follows. We look for a portfolio of the 1Y and 10Y zero with the following properties: (i) the market value of the portfolio equals the market price of the 5Y zero,

$$B_2(\hat{y}_2) = B_1(\hat{y}_1)\theta_1 + B_3(\hat{y}_3)\theta_3; \tag{12.14}$$

and (ii) the local risk of the portfolio equals the local risk of the 5Y zero, $\partial B_2(\hat{y}_2)/\partial \hat{y}_2 = -\mathrm{D}(\hat{y}_2) B_2(\hat{y}_2)$, i.e.:

$$\mathrm{D}(\hat{y}_2) B_2(\hat{y}_2) = \mathrm{D}(\hat{y}_1) B_1(\hat{y}_1)\theta_1 + \mathrm{D}(\hat{y}_3) B_3(\hat{y}_3)\theta_3. \tag{12.15}$$

The solution to Eqs. (12.14) and (12.15) is given by,

$$\theta_1 = \frac{\mathrm{D}(\hat{y}_3) - \mathrm{D}(\hat{y}_2)}{\mathrm{D}(\hat{y}_3) - \mathrm{D}(\hat{y}_1)} \frac{B_2(\hat{y}_2)}{B_1(\hat{y}_1)}, \quad \theta_3 = \frac{\mathrm{D}(\hat{y}_2) - \mathrm{D}(\hat{y}_1)}{\mathrm{D}(\hat{y}_3) - \mathrm{D}(\hat{y}_1)} \frac{B_2(\hat{y}_2)}{B_3(\hat{y}_3)}. \tag{12.16}$$

By the same calculation in the example of the previous section, we have that $B_3(\hat{y}_3) = 0.61391$ and $\mathrm{D}(\hat{y}_3) = 9.5238$. Using the figures in the previous example, we calculate θ_1 and θ_3 in Eqs. (12.16),

$$\theta_1 = \frac{9.5238 - 4.7619}{9.5238 - 0.95238} \frac{0.78353}{0.95238} = 0.45706, \quad \theta_3 = \frac{4.7619 - 0.95238}{9.5238 - 0.95238} \frac{0.78353}{0.61391} = 0.56724.$$

Figure 12.4 depicts the behavior of the bullet price and the market value of the barbell as we change the YTM. Note that the barbell portfolio is more convex than the bullet. Moreover, the barbell trade is "self-financed." By construction, the value of the bullet we sell equals the value of the barbell portfolio. Therefore, large movements in the YTM lead to profits under the assumption of parallel shifts in the yield curve.

Note that in this example, as in that of the previous section, the direction of interest rate movements does not matter for value creation. A "convexity trading" such as this might be the basis of a standard non-directional strategy, resembling one where, say, we go long a number of "undervalued" stocks and short a number of "overvalued" stocks such that the initial value of the portfolio is zero. Then, we likely make profits: in good times, the undervalued stock should increase in value more than the overvalued, and in bad times, the drop in value of the undervalued stock should be less severe than that of the overvalued. For the barbell, the driver of value is convexity: as Eq. (12.10) illustrates, the convexity term, C, is, trivially, always positive, independently of the sign of Δy. Therefore, as soon as we hedge a bond with a portfolio that has the same duration as the given bond, but higher convexity, the position leads to profits, given the assumptions made so far.

Naturally, a barbell trade does not lead to an arbitrage. The P&L summarized by Figure 12.4 relies on the assumption of parallel shifts in the yield curve. However, and as explained in the next chapter (see Section 13.3), it is not realistic to assume that large *and* parallel movements in the yield curve. Historically, large shifts, occurring over long horizons, are accompanied by changes in the yield curve shape. In other words, factors affecting parallel movements in the yield curve are frequent, albeit not the only ones. At least three factors are needed to explain the entire variation of the yield curve.

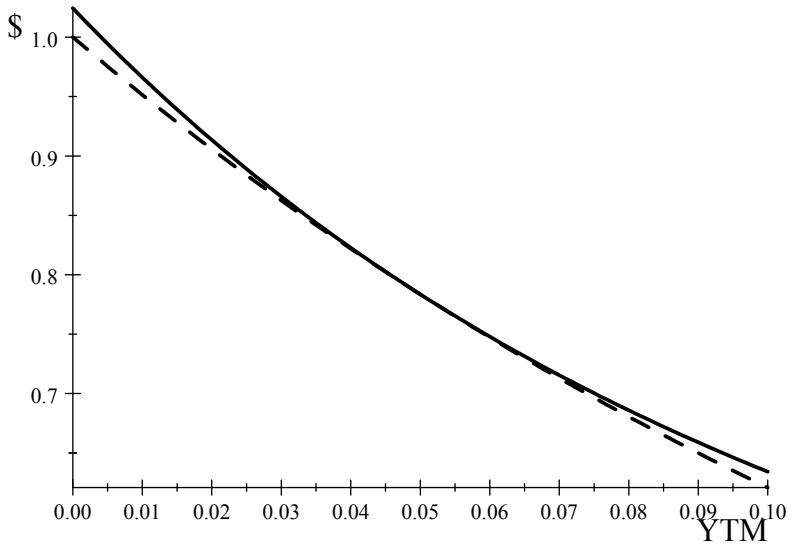


FIGURE 12.4. "Barbell trading." Dashed line (bottom): The price of the 5Y zero, $B_2(y) = \frac{1}{(1+y)^5}$, where y is the YTM. Solid line (top): The value of the "barbell" portfolio consisting of (i) 0.45706 units of the 1Y zero and (ii) 0.56724 of the 10Y zero, i.e. $B_1(y_1) \cdot 0.45706 + B_3(y_3) \cdot 0.56724$, where $B_1(y) = \frac{1}{1+y}$ is the 1Y zero price and $B_3(y) = \frac{1}{(1+y)^{10}}$ is the 10Y zero price.

Table 12.4 considers the case of *non-parallel* shifts in the term-structure. We assume that the initial term-structure is not flat. Then, we consider two scenarios: (i) A "twist" in the term-structure, i.e. long-term rates lower than short-term; (ii) a "steepening" of the term-structure.

TABLE 12.4.

	YTM	Bullet price	Mod. dur.	Barbell value = $\theta_1 B_1(\hat{y}_1) + \theta_3 B_3(\hat{y}_3)$
Initial term-structure				
1Y	$\hat{y}_1 = 4\%$	$B_1(\hat{y}_1) = 0.961$	$\mathrm{D}(\hat{y}_1) = 0.961$	
5Y	$\hat{y}_2 = 5\%$	$B_2(\hat{y}_2) = 0.783$	$\mathrm{D}(\hat{y}_2) = 4.762$	
10Y	$\hat{y}_3 = 6\%$	$B_3(\hat{y}_3) = 0.558$	$\mathrm{D}(\hat{y}_3) = 9.434$	
				Barbell value = 0.783
"Twist"				
1Y	$\hat{y}_1 = 6\%$	$B_1(\hat{y}_1) = 0.943$	$\mathrm{D}(\hat{y}_1) = 0.943$	
5Y	$\hat{y}_2 = 5\%$	$B_2(\hat{y}_2) = 0.783$	$\mathrm{D}(\hat{y}_2) = 4.762$	
10Y	$\hat{y}_3 = 4\%$	$B_3(\hat{y}_3) = 0.675$	$\mathrm{D}(\hat{y}_3) = 9.615$	
				Barbell value = 0.847
"Steepening"				
1Y	$\hat{y}_1 = 4\%$	$B_1(\hat{y}_1) = 0.961$	$\mathrm{D}(\hat{y}_1) = 0.961$	
5Y	$\hat{y}_2 = 5\%$	$B_2(\hat{y}_2) = 0.783$	$\mathrm{D}(\hat{y}_2) = 4.762$	
10Y	$\hat{y}_3 = 7\%$	$B_3(\hat{y}_3) = 0.508$	$\mathrm{D}(\hat{y}_3) = 9.346$	
				Barbell value = 0.751

We use the portfolio in Eq. (12.16). Indeed, note that Eq. (12.16) does not rely on the assumption that the yield curve is flat. It only relies on the assumption that subsequent movements in the yield curve are parallel. The exercise we are about to perform aims to explore the implications of a barbell trade once the assumption of parallel movements is incorrect.

Eq. (12.16), we find that in correspondence of the initial term-structure ($\hat{y}_1 = 4\%, \hat{y}_2 = 5\%, \hat{y}_3 = 6\%$), $\theta_1 = 0.449$ and $\theta_3 = 0.629$. We keep this portfolio fixed, and determine the barbell value, $\theta_1 B_1(\hat{y}_1) + \theta_3 B_3(\hat{y}_3)$, occurring at the two scenarios "twist" and "steepening," with $B_2(\hat{y}_2) = 0.783$ in all cases. The trade is as follows: at time zero, we sale short the five year bond, which we hedge through the barbell portfolio (θ_1, θ_3), using the proceeds of the short-sale. Then, at some future date, we purchase back the five year bond and sell back the portfolio (θ_1, θ_3). The convexity of the barbell trade is, in fact, a view about movements of long-term bond prices, and leads to profits in the "twist" scenario. That is, by convexity, the price B_3 varies more than the price of shorter maturity zeros, thus leading to profits. Note, however, that this strategy leads to losses in the "steepening" scenario.

We need to state an important caveat. The previous conclusions rely on a comparative static analysis, and abstract from the principle that term-structure movements occur under no-arbitrage. For example, the value of the zeros changes over the horizons we are designing scenarios for, even without any changes in the yield curve. Whether this effect is minor depends on the horizon and the model we use to generate scenarios. In Section 12.6.6, we shall revisit the example of this section and illustrate how passage of time and absence of arbitrage can be factored into the analysis, and change some results emanating from Table 12.4.

12.3.3.5 Fixed income arbitrage strategies

The previous "convexity trades" are examples of *yield curve arbitrage* strategies. They may purely rely on convexity or, as discussed in the previous section, on directional views about interest rate movements. For example, we have explained, we may short five year bonds, and go long two- and ten-year bonds, as we view that short-term interest raise will raise and medium term interest rates will lower. This "butterfly" strategy is somehow cheap, intellectually, and not necessarily rewarding, and will be further analyzed in Section 12.6.6.

Swap spread arbitrage is a popular strategy. It was responsible of leading LTCM to a loss of about $1.6 billion in 1997. The strategy works as follows: (i) enter a swap paying the floating LIBOR, L_t, and receiving a fixed rate $\bar{C}$; (ii) short a par Treasury with the same maturity as the swap, thus paying the fixed coupon rate $\overline{CT}$, and invest the proceeds at the repo rate r_t. Thus, the payoff of the strategy is the fixed spread to be received, $F = \bar{C} - \overline{CT}$, and the floating spread to be paid, $S_t = L_t - r_t$. So we go long or short this strategy according to whether we view F to be larger or smaller than the average floating spread S_t over the strategy horizon. Historically, the spread S_t has certainly been volatile, but quite stable, so it is a reasonable strategy. The problem occasionally, though, S_t can attain quite large values.

Trading strategies more sophisticated than the previous ones rely on *models*, aiming to identify points of the yield curve that are misaligned from those predicted by the models. All in all, the strategy is to buy the cheap and short the model-based rich, where the model-based rich is replicated through a portfolio with cash and the bonds that are well-priced by the model, weighted with model-based delta, as in the derivation of the bond pricing formula in the next chapter (see Section 13.4.2.2).

12.3.3.6 Negative convexity, and market volatility

What happens when bond prices have "negative convexity"? In the next chapter, we shall see that the value of a callable bond can be *concave* in the short-term rate. A similar feature is displayed by mortgage-backed-securities (MBS, henceforth), which can now be concave in the YTM! The reason for this negative convexity is that early repayments are likely to occur as the YTM decreases, which entails two inextricable consequences: (i) the price of the MBS "increases less" than a conventional bond price after a decline in the YTM, especially when the YTM is low; (ii) the duration of the MBS decreases as the YTM decreases.

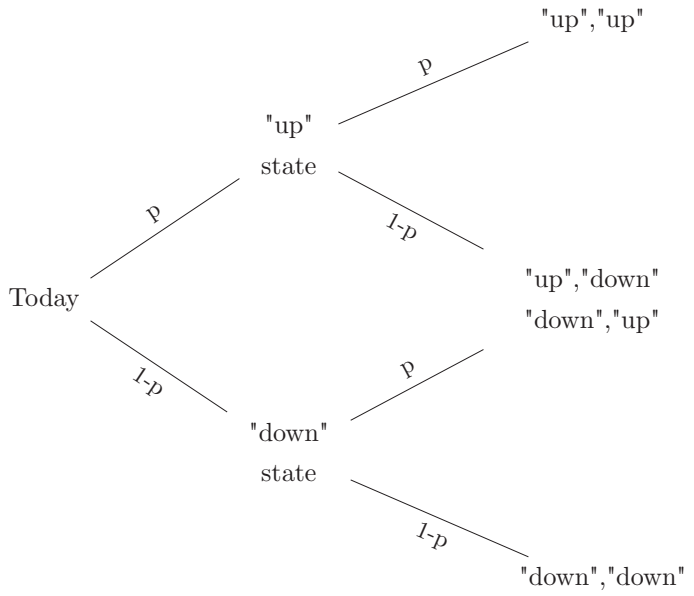
Hedging against MBS might lead to an increased volatility in rate markets. The mechanism is the following. Institutions that are long MBS would typically short conventional bonds for hedging purposes, consistently with the prediction of Eq. (12.11). However, the duration of MBS increases as interest rates increase, due to negative convexity: $\frac{\partial \text{Duration}}{\partial r} = -\text{Convexity}$. Therefore, an interest rate increase can lead these institutions to short additional conventional bonds, which worsens liquidity and leads to a further increase in the interest rates, thereby feeding a vicious circle. Perli and Sack (2003) estimate that in 2002 and 2003, this mechanism may have amplified the volatility of long-term US rates by a factor between 15% and 30%. It is an instance of what is sometimes defined as "endogenous risk," the circumstance that the trend of a certain economic variable triggers actions from market participants that in turn, reinforce the initial trend, as in the case of the 1987 crash discussed in Chapter 12 (see Section 12.4.5), or in the case of assets sell-off in times of crisis, discussed in Chapter 14 (see Section 14.5.3).

12.4 Foundational issues in interest rate modeling

This section is an introduction to the design of binomial trees to price and fit fixed income securities. Such a design requires revising a few details pertaining the equity case examined in Chapter 11. It is instructive to review how generally binomial trees are to be used to model price movements:

(i) At the time of evaluation, we observe the state. In the next period, there can be two mutually exclusive states of the world: (a) the state "up," occurring with probability p; and (b) the state "down," occurring with probability $1-p$.

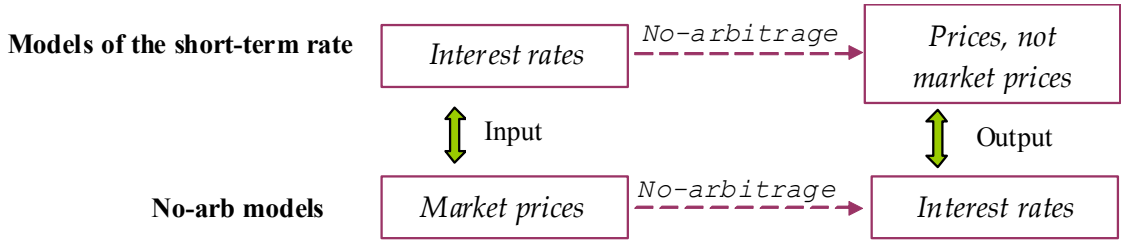
(ii) After two periods, there can be three mutually exclusive states of the world, as in the following diagram. We label the tree in this diagram a "recombining" tree, to emphasize that the "up & down" and the "down & up" nodes are the same.



The previous diagram can be used to price options written on stocks. The stock price unfolds through the branches of the tree. Then, we figure out the no-arbitrage movements of the option price along the tree. Suppose, however, we wish to price an option written on a zero, a 3 Year zero say. Can we apply the same methodology to price the option? The answer is no, and the reason is that we cannot exogenously "track" the movements of the prices of the zero, as in the case of the stock price. Instead, after one year, the 3 Year zero becomes a 2 Year zero, i.e. quite a different asset.

These issues can be mitigated by modeling the movements of the entire yield curve. There are two approaches, as in the diagram below. In the first, we model the dynamics of the short-term rate, defined as the interest rate on a loan with maturity equal to the time intervals in the tree. The resulting model, referred to as *model of the short-term rate*, has implications in terms of the movements of the entire term-structure. This approach, developed in the next section, leads to evaluation formulae in which the current price of the zeros predicted by the model are

not necessarily equal to the *market* prices. A second approach, based on *calibration*, leads to the so-called *no-arbitrage* models, where we model the dynamics of the entire term-structure. This approach gives rise to option evaluation formulae in which the current prices of the zeros predicted by the *model* are equal to the *market* prices. We describe this approach in the last sections of this chapter, using binomial trees, with the next chapter developing their continuous time counterparts.



12.4.1 Tree representation of the short-term rate

12.4.1.1 Recursive evaluation

Consider a two-period, two-state tree, where the current short-term rate is r. The development of the short-term rate is uncertain. That is, the future short-term rate, $\tilde{r}$, is random, and can take two values: either r^+ with (physical) probability p, or r^- with probability $1-p$. We assume that $r^+ > r^-$.

$$\begin{array}{lcl} & r^+ \Rightarrow P(r^+, T) \\ {\scriptstyle p}\ \nearrow & & \\ r & & \\ {\scriptstyle 1-p}\ \searrow & & \\ & r^- \Rightarrow P(r^-, T) \end{array}$$

Let us explain. We are looking for a pricing function of r, say $P(x, T)$, such that it is $P(x = r^+, T)$ in the bad state of the world, and $P(x = r^-, T)$ in the good. The very same function is $P(x = r, T)$, today. When it comes to price options in a complete market setting such as Black & Scholes, we take the stock price as exogeneous and, then, search for an endogenous call pricing function, as explained in the previous chapter, achieving a preference-free evaluation. The usual argument relies on market completeness. In the Black & Scholes market, the stock price is driven by one source of randomness, such that markets are completed by the very same asset, and the option price can be expressed in a preference-free format, through a replication argument.

In the bond pricing case of this section, we shall also price a bond through a replication argument. However, the resulting evaluation formula will not be preference-free, because the risks affecting the asset are not traded—the short-term rate is not traded. While we can still determine the price of the bonds through no-arbitrage, the resulting evaluation formula is not preference-free.

Let us elaborate. Suppose that two zeros with distinct maturities are available for trading. A money market accounting technology is also available (MMA, in the sequel). Investing £1 in the MMA generates £$1 \cdot (1 + r)$ in the second period. We derive an evaluation formula for the zero based on the previous model of the short-term rate. The idea is to build up a portfolio that contains one zero and the MMA. We, then, make sure the value of this portfolio in the second period replicates the value of the zero we wish to price. By no-arbitrage, the value of

the portfolio in the first period equals the value of the zero we wish to price. The appendix develops the arguments, and shows that in the absence of arbitrage, there is a function of r at most, say λ, such that the expected excess return on the bond satisfies:

$$E_p\left[P(\tilde{r},T)\right]-(1+r)\,P\left(r,T\right)=\underbrace{\frac{\Delta P(\tilde{r},T)}{\Delta\tilde{r}}\cdot\mathrm{Vol}\left(\tilde{r}-r\right)}_{=\text{ price volatility}}\cdot\underbrace{\lambda}_{=\text{ unit risk premium}}, \tag{12.17}$$

where $E_p\left[P(\tilde{r},T)\right]$ denotes the expectation of the bond price under p, and $\mathrm{Vol}(\tilde{r}-r)\equiv\Delta\tilde{r}=r^{+}-r^{-}$, which is interpreted as the volatility of the short-term rate.

The meaning of Eq. (12.17) is quite simple. Suppose you invest $£P\left(r,T\right)$ into a MMA. Its gross return is then $(1+r)\,P\left(r,T\right)$. Alternatively, suppose you invest $£P\left(r,T\right)$ into a bond, the expected gross return of which is obviously $E_p\left[P(\tilde{r},T)\right]$. The left-hand side of Eq. (12.17) is, then, the expected excess return on the bond, and the right-hand side the prediction of the model. Eq. (12.17) is, in fact, an APT relation, where λ is interpreted as a unit premium related to the risk of holding the bond over one period of time. It says that the expected excess return on the zero equals the volatility of its price multiplied by the unit price of risk. We call the term,

$$\frac{\Delta P(\tilde{r},T)}{\Delta\tilde{r}}\cdot\mathrm{Vol}\left(\tilde{r}-r\right),$$

"price volatility," because it measures the amplitude of the price variation due to changes in the short-term rate in the future, $\frac{\Delta P(\tilde{r},T)}{\Delta\tilde{r}}$, i.e. the "price-sensitivity", where this price sensitivity is normalized by the volatility of the short-term rate, $\mathrm{Vol}(\tilde{r}-r)$. We can elaborate on Eq. (12.17), and bridge to the continuous time APT relations of Chapter 5. The key observation is that Eq. (12.17) relies on a tree with a trading period normalized to unity. When the trading period is equal to some Δt, the interest earned over that period is $r\cdot\Delta t$, and accordingly, $\lambda\cdot\Delta t$ is the premium to compensate for the risk of holding the bond over an amount of time equal to Δt. With these changes, Eq. (12.17) can be cast as,

$$E_p\left[P(\tilde{r},T)\right]-(1+r\cdot\Delta t)\,P\left(r,T\right)=\frac{\Delta P(\tilde{r},T)}{\Delta\tilde{r}}\mathrm{Vol}\left(\tilde{r}-r\right)\lambda\cdot\Delta t,$$

such that, by considering small Δt, and rearranging terms,

$$\frac{E_p\left(dP(r,T)\right)}{dt}-rP\left(r,T\right)=\frac{\partial P(r,T)}{\partial r}\mathrm{Vol}\left(dr\right)\lambda.$$

Once we assume r is solution to a stochastic differential equation, we can use Itô's lemma, to turn the previous equation into a partial differential equation subject to a boundary condition that states the the bond price is one at expiration. Chapter 13 contains a rigorous discussion of these topics.

Eq. (12.17) can be cast in an alternative format that it equally easy to interpret. Rearranging terms leaves:

$$P\left(r,T\right)=\frac{(p-\lambda)\,P\left(r^{+},T\right)+\left[1-(p-\lambda)\right]P\left(r^{-},T\right)}{1+r}=\frac{E_q\left[P\left(\tilde{r},T\right)\right]}{1+r} \tag{12.18}$$

where $q\equiv p-\lambda$ is the risk-neutral probability.

Let us add a few considerations. We "expect" that $\lambda<0$ because bond prices are decreasing in the short-term rate here. Then, $q\equiv p-\lambda>p$.[5] Hence, the risk-neutral probability of an

[5]To ensure q is a probability, we need to have that (i) $q\equiv p-\lambda>0\Leftrightarrow-\lambda>-p$ and (ii) $q\equiv p-\lambda<1\Leftrightarrow-\lambda<1-p$. That is, $-\lambda\in(-p,1-p)$

upward movement of the short-term rate, q, is higher than the true probability, p. An investor who goes long a bond, is concerned by an increase of the short-term rate in the future and, hence, "corrects" the true probability p by assigning a higher risk-adjusted probability to the "upward" state.

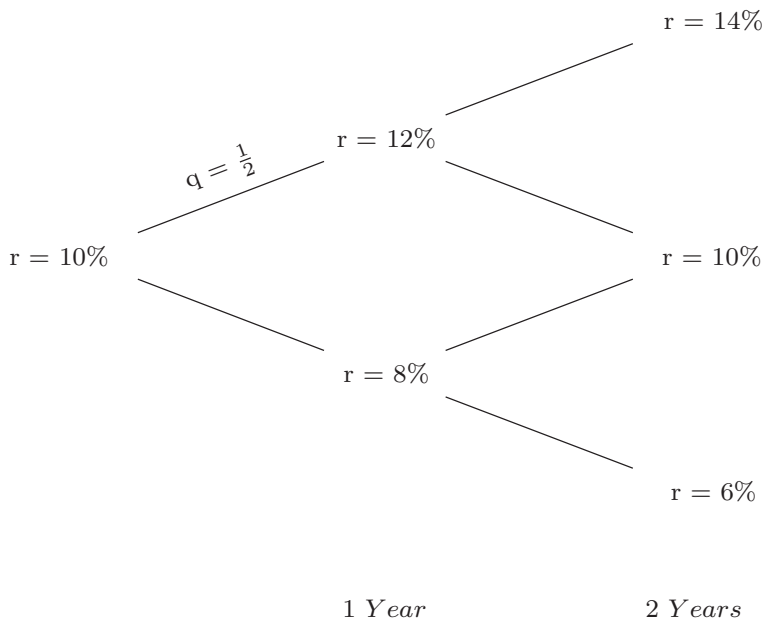
It is the FTAP again. There is not arbitrage if and only if there is a risk-neutral probability such that prices earn an expected return equal to the short-term rate. As explained in Part I (see, e.g., Chapter 2), it does not mean that the market is risk-neutral. The market, instead, prices assets by discounting the assets future payoffs in a generous way, through r. To compensate for this, the expectation in the numerator is taken under a probability which over-weights the bad states of nature, so as to lower back asset prices. It is like we wished to price assets in Planet Earth, a planet full of frictions, risk-aversion and the like, which are reflected in relatively poor asset evaluation. Then, we search for another planet, which has no frictions or risk-aversion, where the asset is evaluated through actuarial methods. To reconcile prices in these two planets, we need the idealized and peaceful planet to display a distorted probability, which assigns relatively higher chances to bad events than Planet Earth does.

12.4.1.2 One example

Assume the current short-term rate equals 10%. We know that in one year, and with (*physical*) probability p, r will increase by 2 percentage points, and with probability $1-p$, it will decrease by 2 percentage points. Finally, with the same probability p, the short-term rate prevaling from the next year to two years time, will increase by 2 further percentage points from its previous value in one year time. We take the probability of an upward movement to be 20% and the absolute value of the Sharpe ratio to be 30%. Given these data, we use the formula, and obtain an estimate of the *risk-neutral probability* of an upward movement of the short-term rate, equal to $q = p - \lambda = 20\% - (-30\%) = 50\%$.

Pricing zeros

Let us price a zero maturing in two years, hinging upon the following tree:



We can use Eq. (12.18) to "fill-in" each node of the tree. We start from the end of the tree, where the price of the two year zero is £1, and then use Eq. (12.18) to fill every node, as illustrated in Figure 12.5. In one year time, the price of the zero is simply one divided by the discount factor prevailing at the beginning of the next year. The price we are looking for is obtained by applying Eq. (12.18) yielding

$$\frac{E_q\left[P\left(\tilde{r},2\right)\right]}{1+r}=\frac{qP(r^+,2)+(1-q)\,P(r^-,2)}{1+r}=\frac{\frac{1}{2}\left(0.8928\right)+\frac{1}{2}\left(0.9259\right)}{1.10}=0.8267.$$

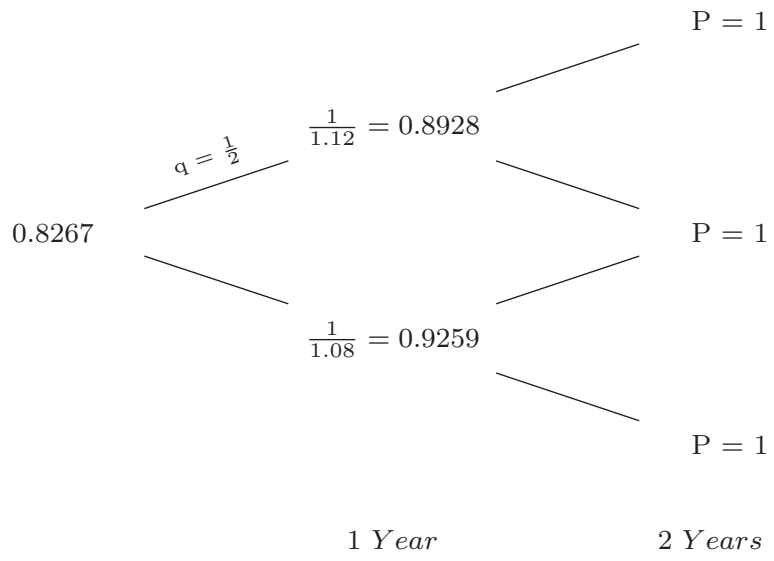


FIGURE 12.5.

Convexity effects

Does the two-year spot rate equal to 10%? It is a natural and, as we shall see, interesting question, as the short-term rate is a martingale under q, in this particular example. However, the answer to the previous question is in the negative. Let us elaborate. The two-year spot rate, $r\left(0,2\right)$, satisfies, $0.8266=\left(1+r\left(0,2\right)\right)^{-2}$, or $r\left(0,2\right)=9.98\%$. That is,

$$0.8267=\frac{1}{1+r}E_q\left(\frac{1}{1+\tilde{r}}\right)>\frac{1}{1+r}\frac{1}{1+E_q(\tilde{r})}=0.8264.$$

In other words, suppose the interest rate is known with certainty to be 10% in the second period. Then, the price should be 0.8264, because $1+E_q(\tilde{r})$. Price, then, increase upon activation of uncertainty. It's a "convexity effect," which we shall explain in deeper detail the next chapter (Section 13.4.5.1).

12.4.2 Tree pricing

We can generalize the tree to a multiperiod case. We use Eq. (12.18) to evaluate zeros at all nodes of the tree and maturities. Given q, which can be estimated once we estimate p and λ, we use recursively Eq. (12.18), and calculate prices of zeros. We can, then, price any derivative written on these zeros. The drawback of this approach is that the initial term structure is predicted

with error! Let us illustrate with a concrete numerical example. Consider the tree in Figure 12.6, where the current short-term rate for one year is $r = 4\%$. Also shown in this tree is the price of a hypothetical 3 Year zero at the expiration date, $t = 3$, and at $t = 2$. At $t = 3$, $P = 1$ in all states of nature. At $t = 2$, the price P is $P(r, T) = E_q[P(\tilde{r}, T)]/(1 + r) = 1/(1 + r)$, for $r = 6\%$, 4% and 2%. The issue, now, is how to determine the price of the zero in correspondence of the remaining nodes. We should use the formula, $P(r, T) = E_q[P(\tilde{r}, T)]/(1 + r)$ to populate the tree, but we do not know p, λ, and q. Suppose we "estimate" p and λ. In this case, we determine q simply as $q = p - \lambda$, as in Eq. (12.18), and we come up with $q = \frac{1}{2}$. The following diagram gives the price of the zero in all the nodes at time $t = 1$, and at the evaluation time $t = 0$, yielding a price of the 3 Year zero equal to 0.8893.

Next, consider a European call option written on the 3 Year zero, with expiration date equal to 2 and strike price $K = 0.95$. The following diagram gives the value of the option predicted by the model at each node of the tree. The model predicts that the current price of the call option is 0.0124.

12.4.3 Introduction to calibration

Calibration is a procedure by which we search for a given model's parameter values, such that the model's predictions coincide with selected empirical counterparts. For example, in the real business cycle literature reviewed in Chapter 3, we might calibrate a model's parameters to ensure that the correlation of asset returns with output is the same as that in the data. In this chapter, calibration is a procedure by which we search for a given model's parameters that make the model price equal to the market.

12.4.3.1 Motivation

The *model* we are dealing with in the previous section *predicts* that the price of the 3 Year zero is equal to 0.8893. However, there is no guarantee that this model-implied price *equals* the *market price* of the 3 Year zero. Suppose, instead, that the market price of the 3 Year zero, $P_{\$}$ say, equals 0.8700. What should we do to make the model-implied price of the 3 Year zero equal to the market price? The question is important: how can we trust an option pricing model that is not even able to pin down the initial market value of the asset underlying the option contract? Alternatively, suppose the option is a bespoke product of a bank. The bank's client might question why the bank's evaluation model predicts a price of the bond to be so high (0.8893), compared to the observable and cheap market price (0.8700).

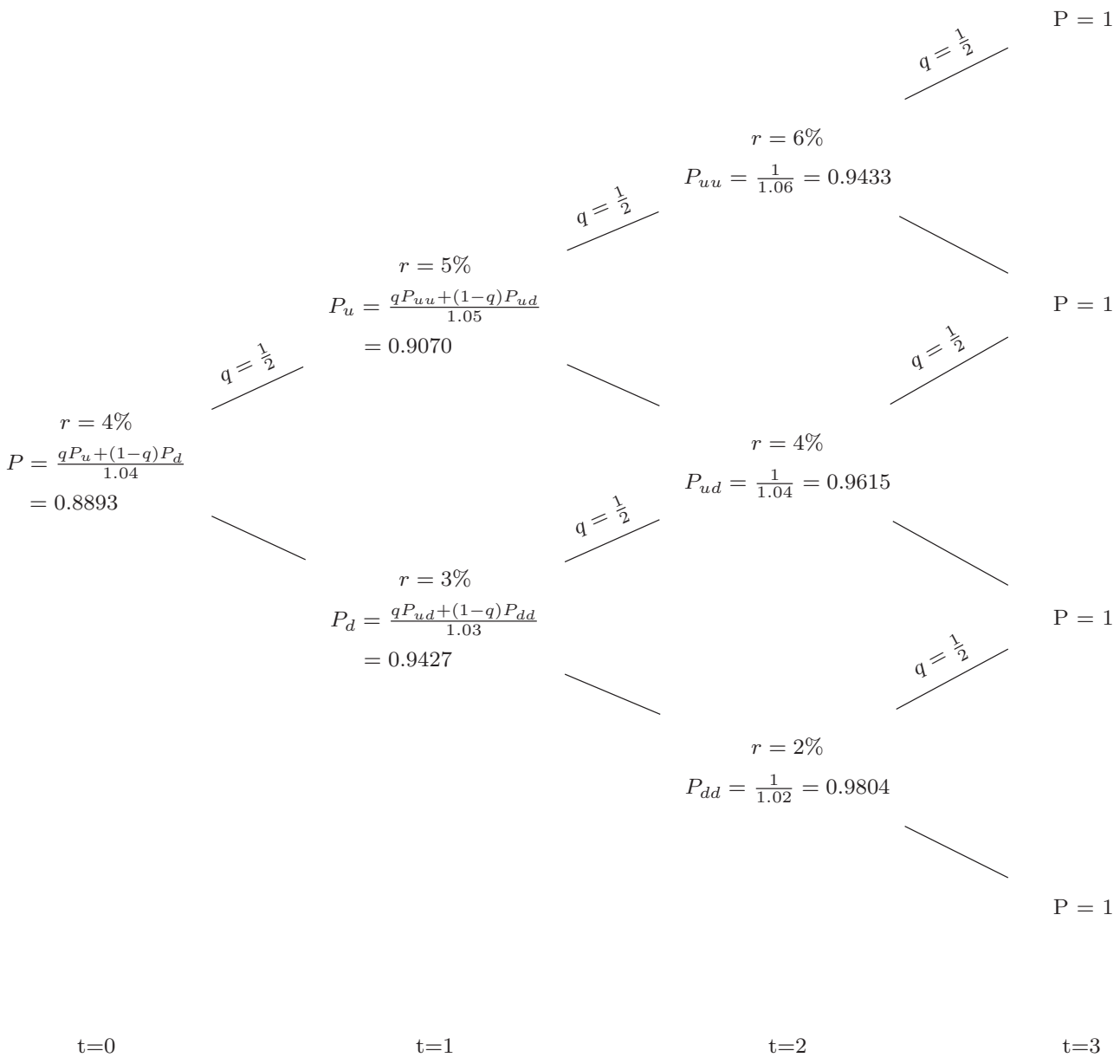
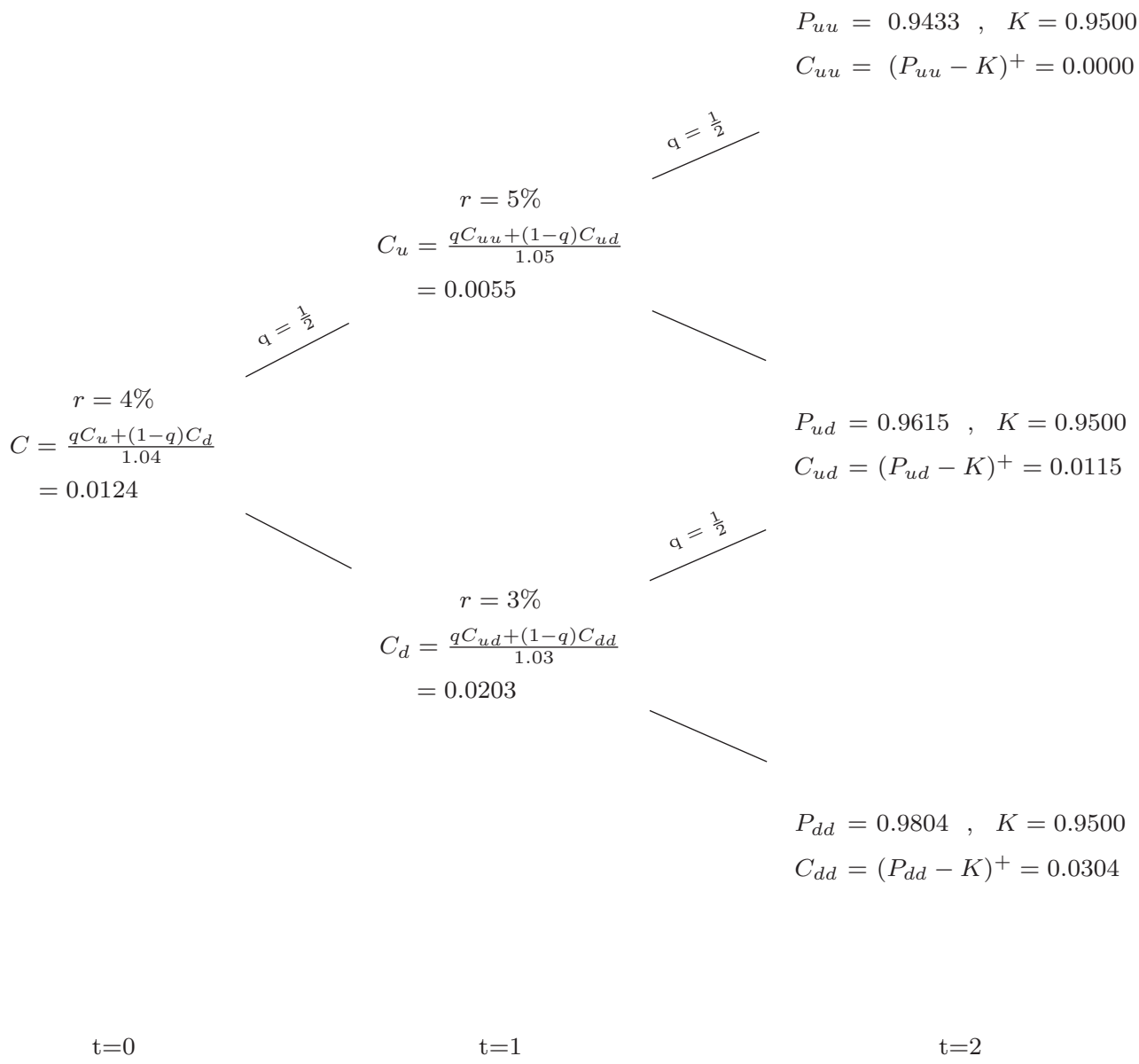


FIGURE 12.6. The dynamics of the short-term rate

12.4.3.2 Perfectly fitting trees

We look for perfectly fitting trees, that is, those trees where risk-neutral probabilities and/or the values of the short-term rate in each node are not given in advance but, rather, are such that the initial yield curve is fitted without error. These trees are called implied binomial trees—"implied" by the market prices. Let us consider the example in the previous section. To make the model-implied price of the 3 Year zero equal to the market price, $P_{\$} = 0.8700$, we cannot take the risk-neutral probability q as given, i.e. independent of the observed price $P_{\$} = 0.8700$, as we did before. Rather, we should *calibrate* the probability q, as follows,

$$P_{\$} = 0.8700 = \frac{1}{1.04}\left[q \cdot P_1\left(5\%\right) + (1-q) \cdot P_1\left(3\%\right)\right] \tag{12.19}$$



where $P_1(5\%)$ and $P_1(3\%)$ are the prices of the zero at time $t = 1$, in the events that the short-term rate is up to 5% or down to 3%.

The previous equation follows, again, by Eq. (12.18). Note, now, that the unknown is not the price, which is instead given by the market price. Rather, we are looking for, or calibrating, the probability q that makes the RHS of Eq. (12.19) equal to its LHS. Naturally, we need to calculate the prices of the zeros $P_1(5\%)$ and $P_1(3\%)$. These prices can be found by another application of Eq. (12.18), as follows,

$$P_1(5\%) = \frac{q \cdot 0.9433 + (1-q) \cdot 0.9615}{1.05}, \quad P_1(3\%) = \frac{q \cdot 0.9615 + (1-q) \cdot 0.9804}{1.03}.$$

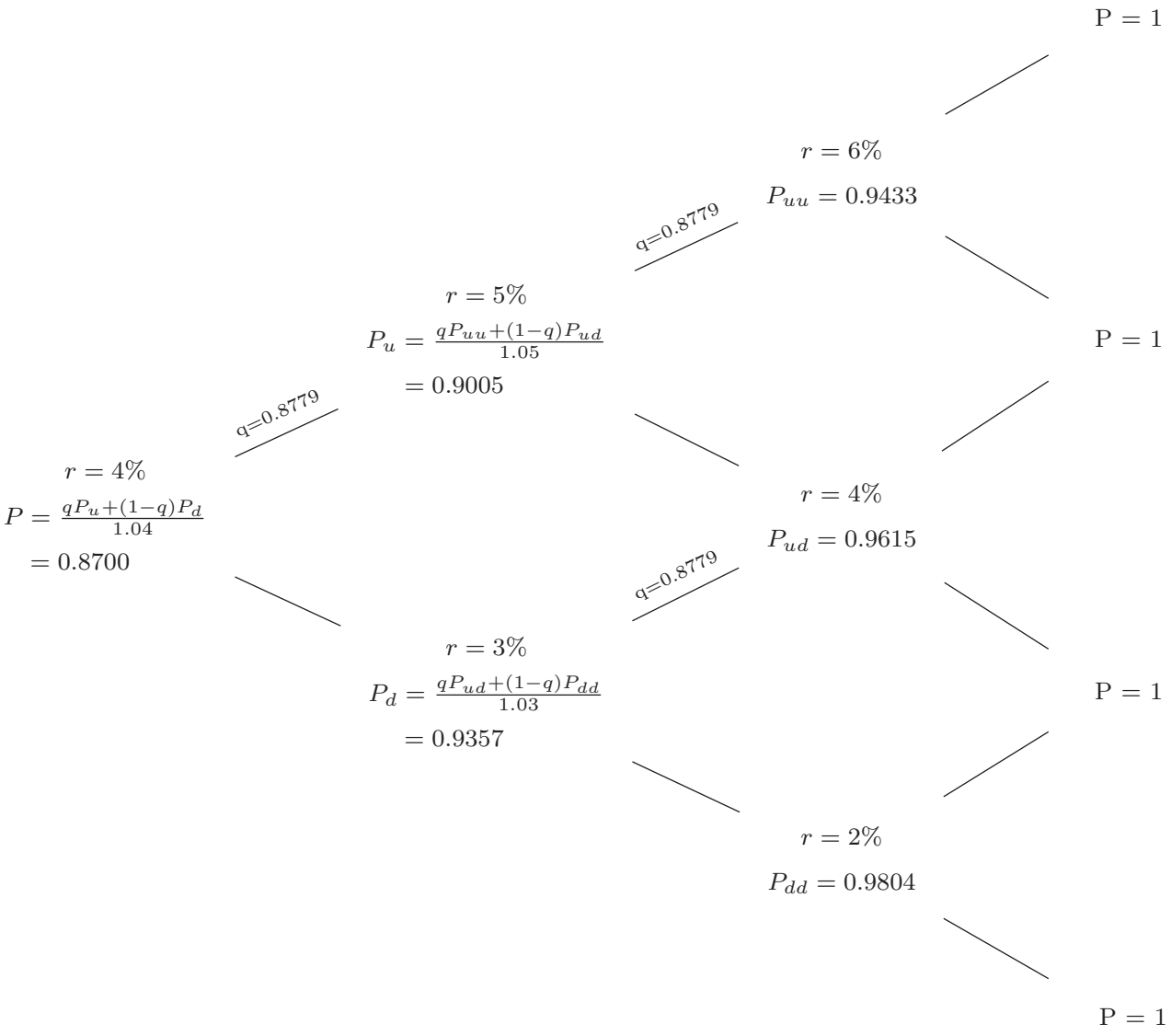
By replacing the previous expressions for $P_1(5\%)$ and $P_1(3\%)$ into Eq. (12.19), we obtain,

$$P_\$ = 0.8700 = \frac{1}{1.04}\left(q \cdot \frac{q \cdot 0.9433 + (1-q) \cdot 0.9615}{1.05} + (1-q) \cdot \frac{q \cdot 0.9615 + (1-q) \cdot 0.9804}{1.03}\right).$$

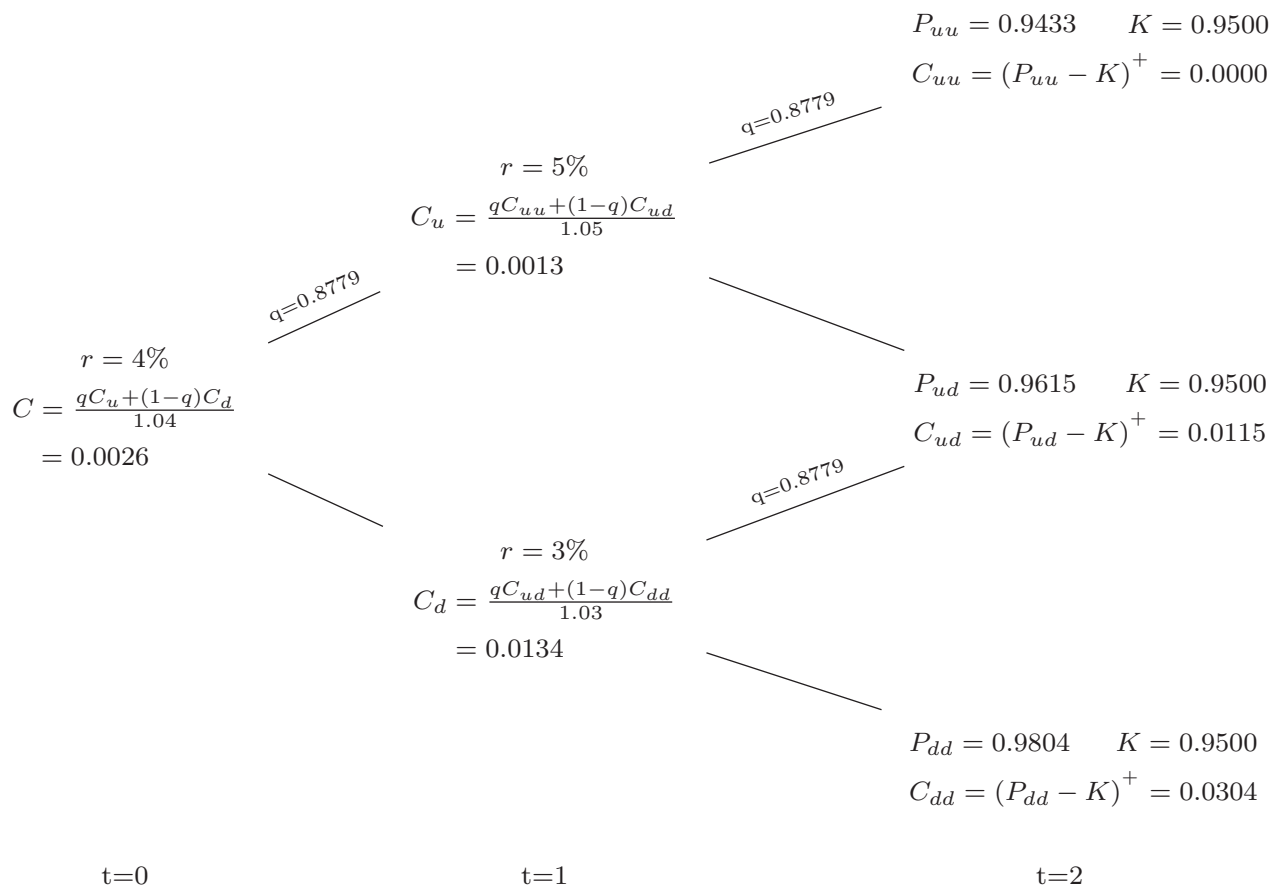
This is a nonlinear equation in q, which we can easily solve, to obtain, $q = 0.8779$. Hence, we find:

$$P_1(5\%) = 0.9005 \text{ and } P_1(3\%) = 0.9357.$$

The next diagram depicts the *implied binomial tree*, i.e. the tree that we obtain after we match the model-implied price of the 3 Year zero to the market price, $P_\$ = 0.8700$.



Note that $P_1(5\%)$ and $P_1(3\%)$ are quite away from the values we found earlier whilst imposing that $q = \frac{1}{2}$. In the "implied" tree, they are smaller than those obtained with $q = \frac{1}{2}$, state by state. This is because in the implied tree, $q = 0.8779$, such that the model can match a lower initial price, 0.8700. The implied tree puts more weight on those states of nature where the short-term rate is high or, equivalently, bond prices are low. We expect the price of the option in the implied binomial tree to be lower than that we found earlier, because call option prices decrease with the underlying. Let us perform the calculations, by relying on the *implied binomial tree* depicted in the next diagram.

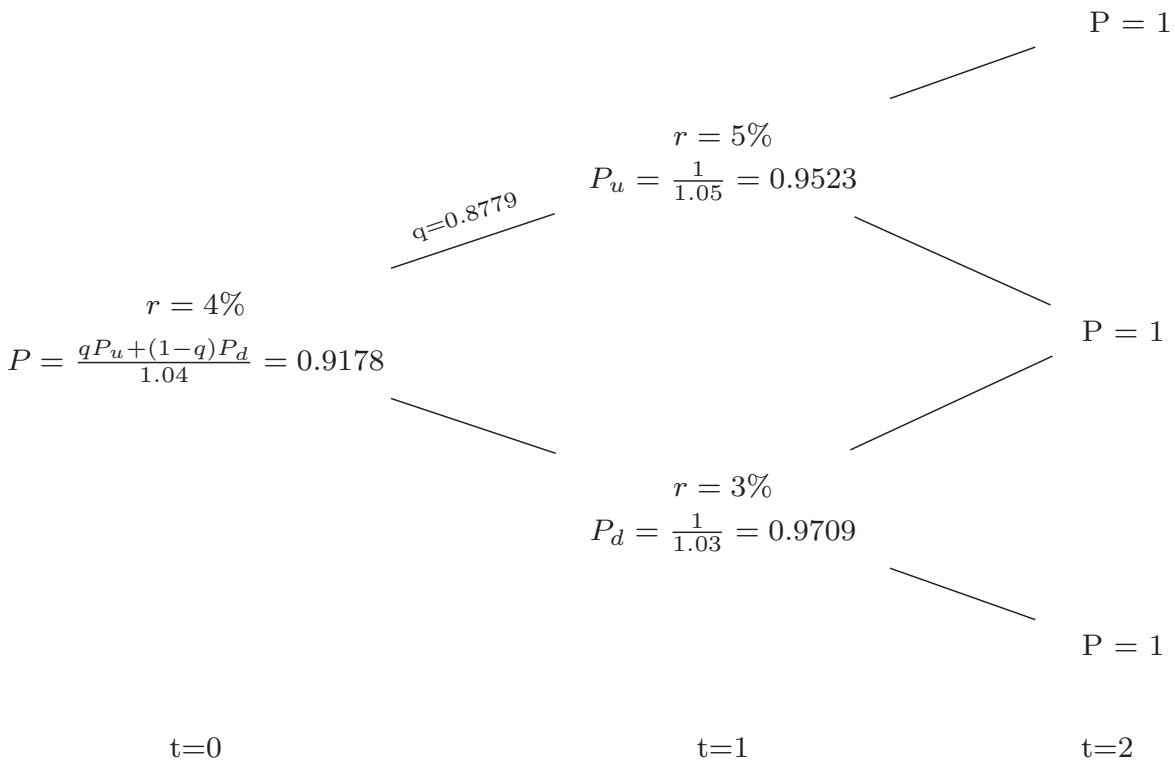


The calculations in the previous diagram reveal indeed that the option price predicted by the implied binomial tree is 0.0026, which is one order of magnitude less than the option price we find earlier, 0.0124! The interpretation of this result relates, again, to the implied risk-neutral probability, which is much larger than $q = \frac{1}{2}$. The implied tree puts a relatively large weight on events where the short-term rate is high or bond prices are low, which reduces the likelihood the option will be exercized leading then to a small option price.

12.4.3.3 Another zero

We might not be done yet. Let us go back to the zero pricing problem, and assume we observe the price of a 2 Year zero, and that this price equals 0.9200, a reasonable figure. Is there any chance that the inputs to the pricing problem for the 3 Year zero could also lead to fit the 2 Year zero without errors? Of course there isn't. Indeed, in the next diagram, we use the inputs to the 3 Year zero, and Eq. (12.18), and find that the price of the 2 Year zero implied by the price of the 3 Year zero is equal to 0.9178. Unless the market price happens, by chance, to equal 0.9178, we cannot simultaneously fit the price of the 3 Year and the 2 Year zeros.

To simultaneously fit the price of the 3 Year and the 2 Year zeros, we should implement at least one of the two strategies: (i) to make the probabilities q time-varying; (ii) to calibrate the entire structure of the short-term movements in Figure 12.6. We implement the first of these two strategies in the next subsection. We develop the second strategy in Section 12.5.



12.4.3.4 Implementing implied binomial trees

We build up implied binomial trees in more general cases, arising in the presence of several bond prices to be matched. Suppose the time interval is six months, such that the short-term rate is for six months. The current short-term rate is 3.99%, annualized. It can change to either 4.50% or to 4.00%, with equal (physical) probability. Suppose that two zeros are available for trading: a 6M zero and a 1Y zero, where the current price of the 1Y zero is 0.95974. What is the risk-neutral probability implied by this tree? This probability must be such that, the price of all the zeros are matched exactly. Figure 12.7 depicts the tree corresponding to this situation.

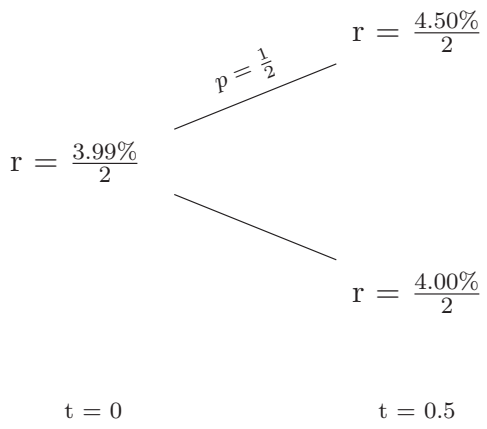
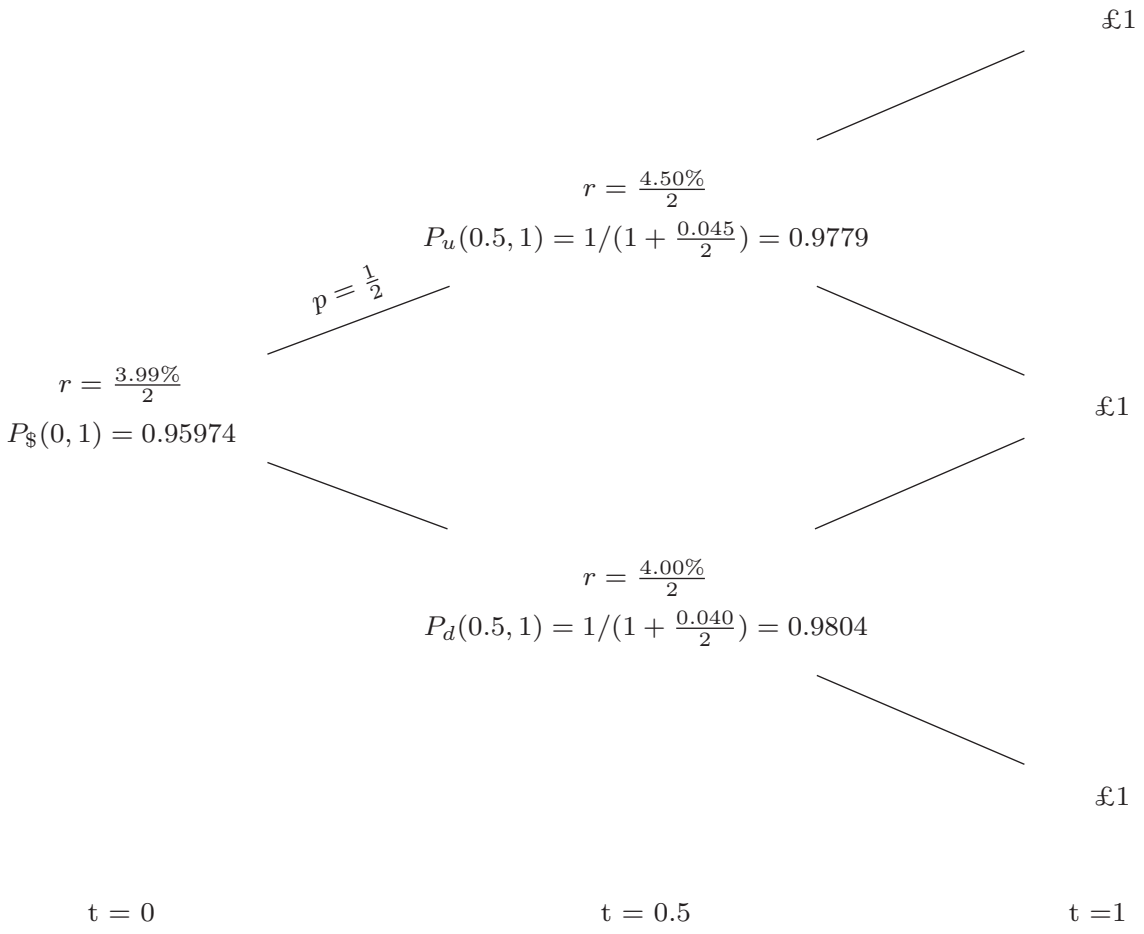


FIGURE 12.7. The dynamics of the short-term rate: high interest rate scenario

In this tree, $p = \frac{1}{2}$ denotes the *physical* probability. Naturally, the price of a 6M zero at $t = 0$, equals, $P_{\$}(0, 0.5) = 1/\left(1 + \frac{0.0399}{2}\right) = 0.9804$. This price is actually observed. That is, the current short-term rate, 3.99%, is a mere definition. Next, we proceed to find the no-arbitrage movements of the 1Y zero, which are displayed below.



Note, the current market price, $P_{\$}(0,1) = 0.95974$, is less than the *expected* price to prevail tomorrow, discounted at the current interest rate,

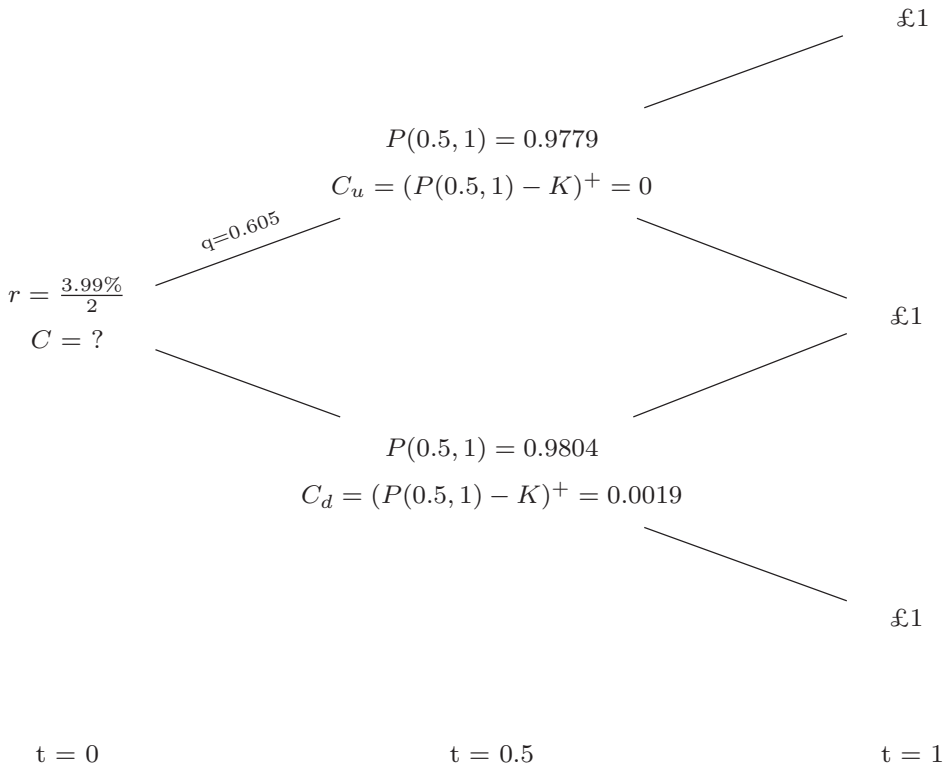
$$\frac{1}{1+r}E_p\left[P\left(0.5,1\right)\right] = \frac{1}{1+\frac{0.0399}{2}}\left(\frac{1}{2}0.9779 + \frac{1}{2}0.9804\right) = 0.9599.$$

Hence, $p = \frac{1}{2}$ cannot be the risk-neutral probability. To find out the risk-neutral probability, we proceed as follows. In the absence of arbitrage,

$$P_{\$}(0,1) = 0.95974 = \frac{1}{1+\frac{0.0399}{2}}\left[q \cdot 0.9779 + (1-q)\cdot 0.9804\right],$$

with obvious notation. This is one equation with one unknown, q, which is solved by $q = 0.605$.

We may now proceed with pricing derivatives. Consider a European call option on the 1Y zero, with expiration date in six months and exercise price equal to 0.9785. Its payoff is as depicted below:



So the option price is, by risk-neutral evaluation,

$$C = \frac{1}{1+\frac{0.0399}{2}}\left[q\cdot 0 + (1-q)\cdot 0.0019\right] = 0.9804\left[0.395\cdot 0.0019\right] = 7.3579\times 10^{-4}. \qquad (12.20)$$

What happens when the short-term rate does not evolve as in the diagram of Figure 12.7 but, instead, as in Figure 12.8?

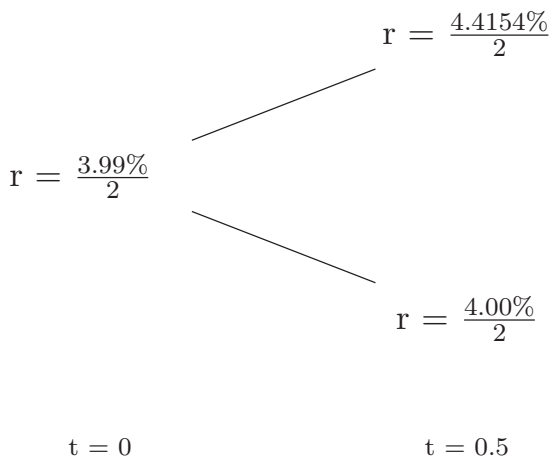


FIGURE 12.8. The dynamics of the short-term rate: low interest rate scenario

The previous tree is one where the short-term rate in the upper state of the world equals $r = 4.4154\%$, not 4.50%, as in Figure 12.7. It implies that the price of the 1Y bond in 6 months in this state is:

$$P_{\text{up}}(0.5, 1) \equiv \frac{1}{1 + \frac{r}{2}} = \frac{1}{1 + \frac{4.4154\%}{2}} = 0.9784.$$

The risk-neutral probability, q, solves:

$$\begin{aligned} P_{\$}(0, 1) &= 0.95974 \\ &= \frac{1}{1 + r} \left[q P_{\text{up}}(0.5, 1) + (1 - q) P_{\text{down}}(0.5, 1) \right] \\ &= \frac{1}{1 + \frac{0.0399}{2}} \left[q \cdot 0.9784 + (1 - q) \cdot 0.9804 \right]. \end{aligned}$$

The solution is, $q = 0.756$, which is higher than the solution we found earlier using the tree in Figure 12.7 (i.e., $q = 0.605$). The option price is, now,

$$C = \frac{1}{1 + \frac{0.0399}{2}} \left[q \cdot 0 + (1 - q) \cdot 0.0019 \right] = 0.9804 \left[0.244 \cdot 0.0019 \right] = 4.5451 \times 10^{-4}. \tag{12.21}$$

The up-state of the world in Figure 12.8 is less severe than that in Figure 12.7. Why then is the price in Eq. (12.21) smaller than that in Eq. (12.20)? To match the initial price $P_{\$}(0, 1) = 0.95974$, the model in Figure 12.8 must put more weight on the up-state of the world, i.e. a larger implied risk-neutral probability. This implies a larger risk-neutral probability that low bond prices will arise in the future and, hence, a lower option price.[6]

In a segmented market, two investment banks might have different views about developments in the short-term rate—the two alternative views illustrated by Figures 12.7 and 12.8. The first bank favours a "high" interest rate scenario, but it is not too risk-averse to that scenario ($r_{\text{up}} = 4.5\%$, $q = 0.605$). The second bank favours a "mild" interest rate scenario, although

[6]Mathematically, we have that $P_{\$}(0, 1) = \frac{1}{1+r}(P_{\text{down}} - q\Delta P)$, where $\Delta P \equiv P_{\text{down}} - P_{\text{up}} > 0$. While P_{down} predicted by Figure 12.7 is the same as that in Figure 12.8, the "bond price volatility," i.e. the difference ΔP, is lower in Figure 12.8 than in Figure 12.7. Therefore, the tree in Figure 12.8 is consistent with the given market price, $P_{\$}(0, 1)$, only when q increases from 0.605 to 0.756.

it assigns a greater chance of this scenario to arise ($r_{\text{up}} = 4.4154\%$, $q = 0.756$). But then, naturally, both institutions need to agree on the initial bond price, $P_{\$}(0,1) = 0.95974$. (The first bank might have, then, a quite conservative risk-management system although then its option prices are higher than the second bank.) The segmentation could arise, for example, because the clientèle of the first bank and that of the second bank are unlikely to meet and, the prices for the option charged by the banks are not publicly known. In the absence of market imperfections (and arbitrage), however, the investment banks should agree on the option price too. Note, finally, that the price in Eq. (12.21) is almost half of that in Eq. (12.20). Derivatives can be quite nonlinear object, due to their optionality. A small deviation in the assumptions on the short-term rate developments can lead to dramatic option pricing implications.

Let us add a period in the tree of Figure 12.7, assuming that the short-term rate is as in the following diagram:

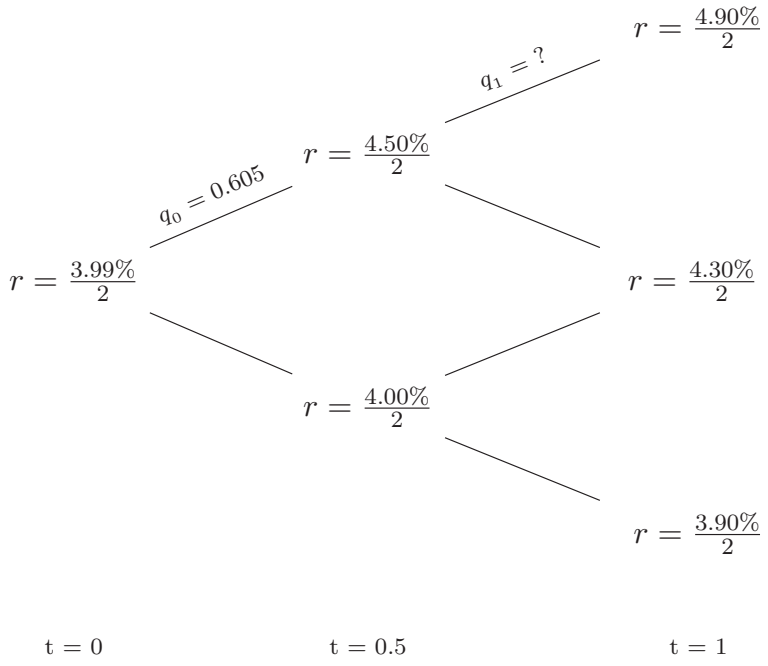
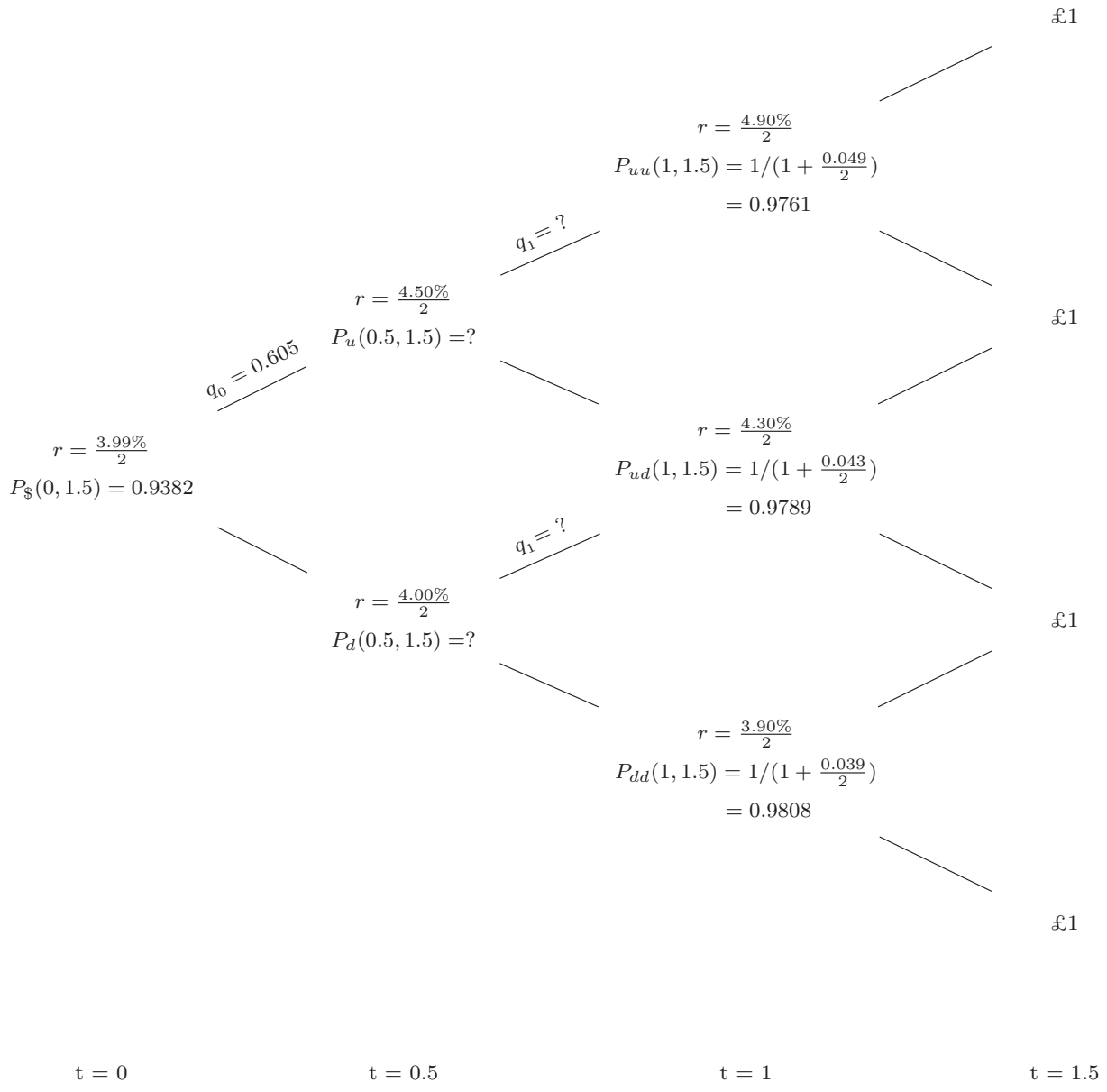


FIGURE 12.9.

In this tree, q_0 is the risk-neutral probability for the first period, and q_1 is the risk-neutral probability for the second period. We already know that $q_0 = 0.605$. The probability q_1 is the risk-neutral probability for the time-period $(0.5, 1)$, and can differ from q_0. Suppose, also, that an additional zero is available for trading, a 1.5Y zero. The current price of this zero is $P_{\$}(0, 1.5) = 0.9382$. To derive the risk-neutral probability q_1, we calibrate the implied tree for the 1.5Y zero, as follows.



We need to calculate the prices $P_u(0.5, 1.5)$ and $P_d(0.5, 1.5)$, which shall be used to receover q_1, throughout the no-arbitrage property of the zero, and the previously calculated $q_0 = 0.605$. By no-arbitrage, we have, as usual, that:

$$P_u(0.5, 1.5) = \frac{1}{1 + \frac{0.045}{2}} [q_1 \cdot 0.9761 + (1 - q_1) \cdot 0.9789] \tag{12.22}$$

$$P_d(0.5, 1.5) = \frac{1}{1 + \frac{0.040}{2}} [q_1 \cdot 0.9789 + (1 - q_1) \cdot 0.9808] \tag{12.23}$$

The problem, q_1 is not known. Therefore, Eqs. (12.22)-(12.23) do not allow us to pin down the prices $P_u(0.5, 1.5)$ and $P_d(0.5, 1.5)$. But here is where calibration comes in. We know the current price of the 1.5Y zero, which is, $P_{\$}(0, 1.5) = 0.9382$. In the absence of arbitrage,

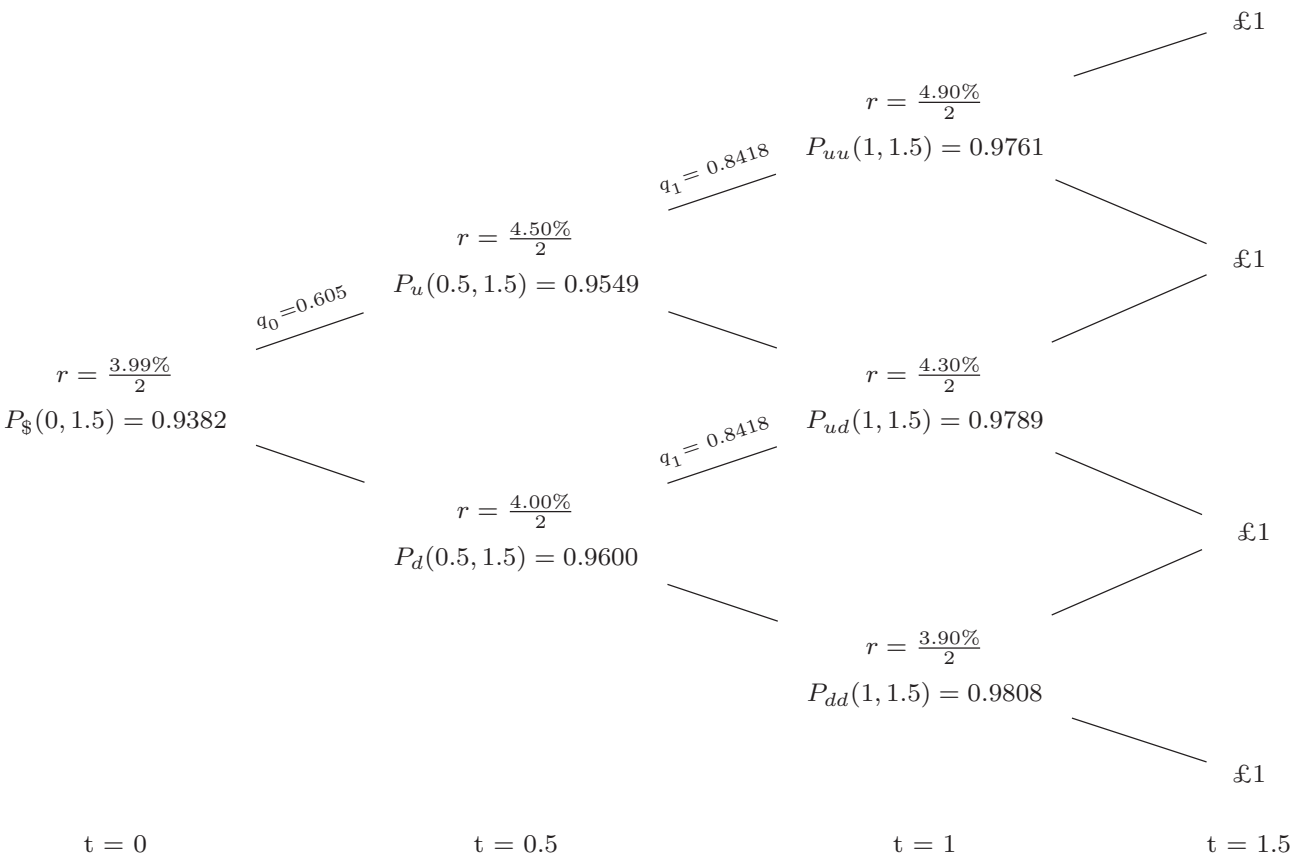
$$P_{\$}(0, 1.5) = 0.9382 = \frac{1}{1 + \frac{0.0399}{2}} [q_0 \cdot P_u(0.5, 1.5) + (1 - q_0) \cdot P_d(0.5, 1.5)],$$

where $P_u(0.5, 1.5)$ and $P_d(0.5, 1.5)$ are as in Eqs. (12.22)-(12.23), and where $q_0 = 0.605$. So we have,

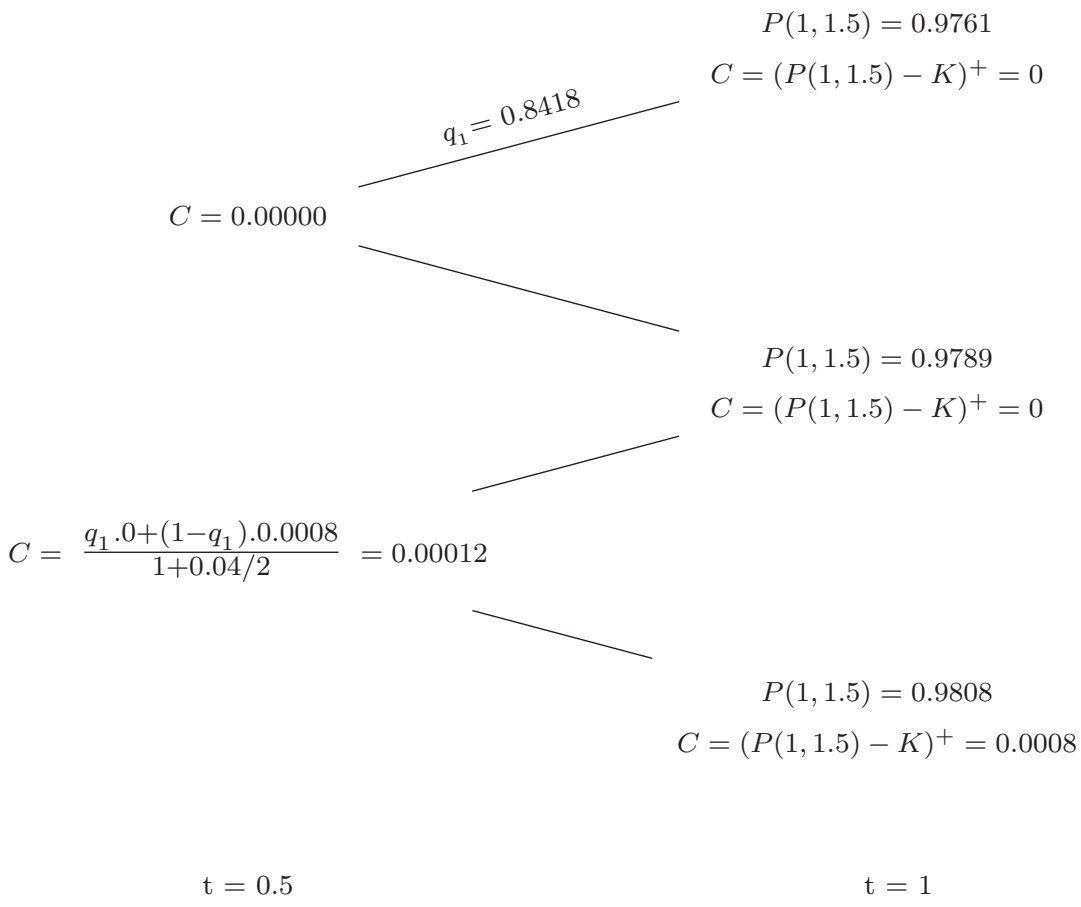
$$0.9382 = \frac{1}{1 + \frac{0.0399}{2}} \left[0.605 \cdot P_u(0.5, 1.5) + 0.395 \cdot P_d(0.5, 1.5)\right], \tag{12.24}$$

where $P_u(0.5, 1.5)$ and $P_d(0.5, 1.5)$ are as in Eqs. (12.22)-(12.23). Hence, by replacing Eqs. (12.22)-(12.23) into Eq. (12.24) leaves one equation with one unknown, q_1. Solving, yields, $q_1 = 0.8412$, which implies that, $P_u(0.5, 1.5) = 0.9549$, and $P_d(0.5, 1.5) = 0.9600$.

To sum up, we have the tree below.



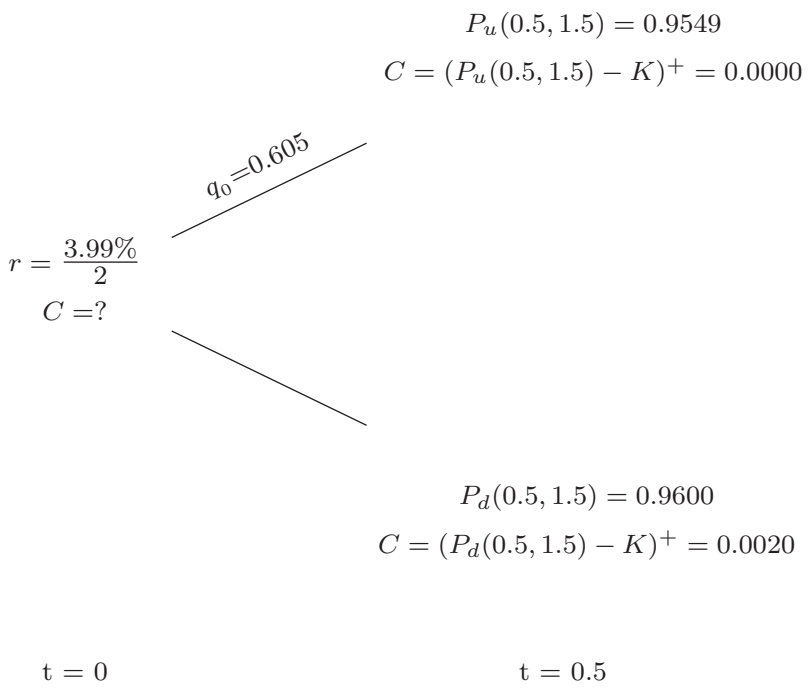
We are ready to evaluate derivatives written on these zeros. Consider, for example, a call option on the 1.5Y zero, with expiration date in 1Y and exercise price equal to 0.9800. The price of the option at time $t = 0.5$, is either zero or $C = 0.00012$, as illustrated below.



The no-arbitrage price of the 1Y call option on the 1.5Y zero, struck at $K = 0.9800$, is:

$$C = \frac{1}{1 + \frac{0.0399}{2}} [0 \cdot q_0 + 0.00012 \cdot (1 - q_0)] = 0.9804 [0.00012 \cdot (1 - 0.605)] = 4.647 \times 10^{-5}.$$

We can use the tree in Figure 12.9 to price additional derivatives, such as, say, a call option on the 1.5Y zero, with expiration date in six months, and exercise price equal to 0.9580. We have the following tree.

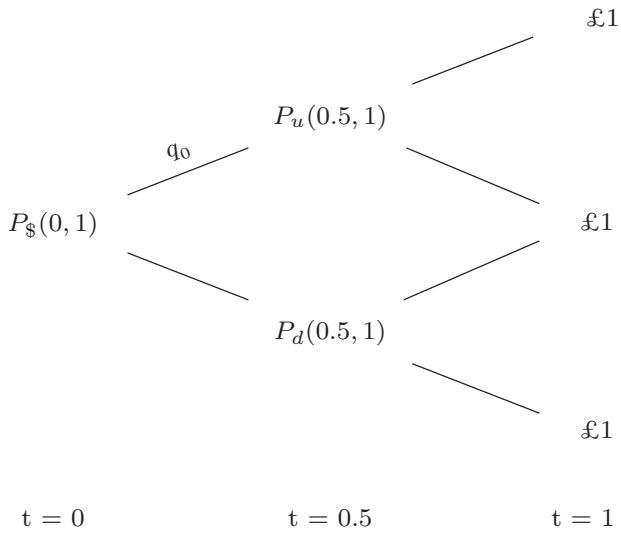


Therefore, the no-arbitrage price of the option is,

$$C = \frac{1}{1 + \frac{0.039}{2}} \left[q_0 \cdot 0 + (1 - q_0) \cdot 0.0020\right] = 0.9804 \left[0.395 \cdot 0.0020\right] = 7.745 \times 10^{-4}.$$

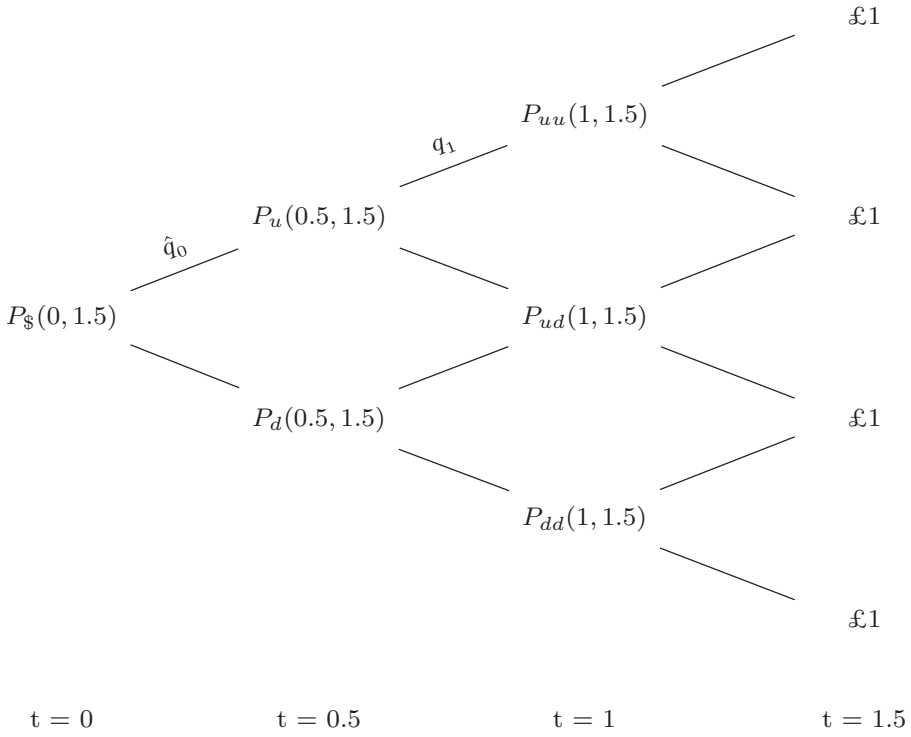
12.4.3.5 Summing up

What have we done? Our starting point is the tree in Figure 12.9, which we use to recover the two risk-neutral probabilities q_0 (for the time span $(0, 0.5)$) and q_1 (for the time span $(0.5, 1)$), using the information about the market price of two zeros, the 1Y and the 1.5Y. Precisely, given $P_{\$}(0, 1)$, the price of the 1Y zero, we recover q_0, as illustrated below:



This is possible as $P_u(0.5, 1)$ and $P_d(0.5, 1)$ do not "depend" on q_0 and so they are obtained in a straightforward manner.

Next, and given q_0, we determine q_1, using $P_{\$}(0, 1.5)$, the price of the 1.5Y zero, as illustrated below:



Again, the risk-neutral probability, q_1, can be recovered because $P_{uu}(1, 1.5)$, $P_{ud}(1, 1.5)$ and $P_{dd}(1, 1.5)$ do not "depend" on q_1, as they are next to expiration, and are thus easily obtained. Given $P_{uu}(1, 1.5)$, $P_{ud}(1, 1.5)$ and $P_{dd}(1, 1.5)$, we can express $P_u(0.5, 1.5)$ and $P_d(0.5, 1.5)$ as two linear functions of q_1. Finally, no-arbitrage forces the market price, $P_{\$}(0, 1.5)$, to be linear in $P_u(0.5, 1.5)$ and $P_d(0.5, 1.5)$ and, hence, q_1, thereby allowing us to recover q_1.

We can continue, by adding one time period, as in the tree in Figure 12.10 below. We can recover q_2, once we are given the market price of a 2Y zero, $P_{\$}(0, 2)$, as follows:

- The price of the 2Y zero at time $t = 1.5$ (the filled nodes in Figure 12.10) (say $P(1.5, 2)$) can now be determined, given an assumption about the numerical values of the short-term rate in those nodes.
- Given the prices $P(1.5, 2)$ at time $t = 1.5$, and the previously calibrated probabilities $\hat{q}_0$ and $\hat{q}_1$, we impose no-arbitrage thereby expressing the current market price $P_{\$}(0, 2)$ as a linear function of q_2. Then, we solve for q_2.

The calibration can continue. We extend the tree, by adding more and more periods. Then, we use the price of one additional zero to "recover" time varying risk-neutral probabilities. An alternative procedure consists in: (i) *fixing* the risk-neutral probabilities q to some value at all times (e.g., $q = \frac{1}{2}$), and (ii) figuring out the "implied" values for the short-term rate in each node of the tree. Section 12.6 develops a systematic approach that allows to implement this procedure.

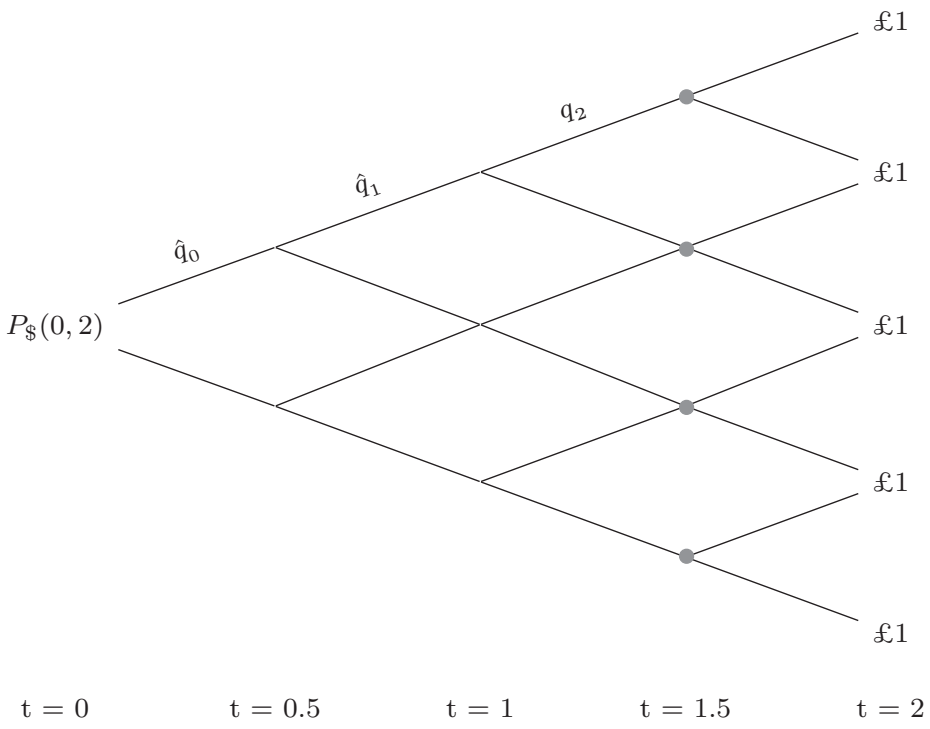


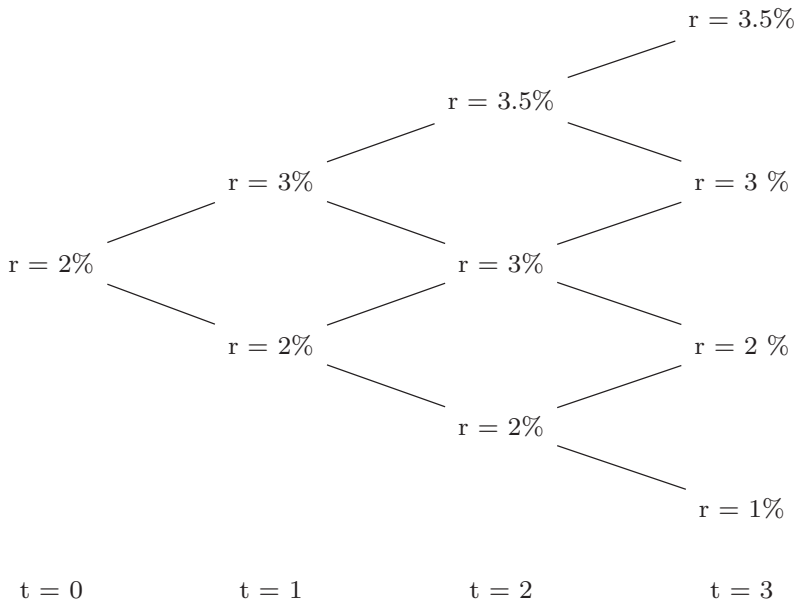
FIGURE 12.10.

12.4.4 Calibrating probabilities throught derivative data

This section deals with two numerical examples where we exploit information from derivative data to say something about the assets underlying the very same derivative contracts. Namely, we shall use derivative data to calibrate risk-neutral probabilities.

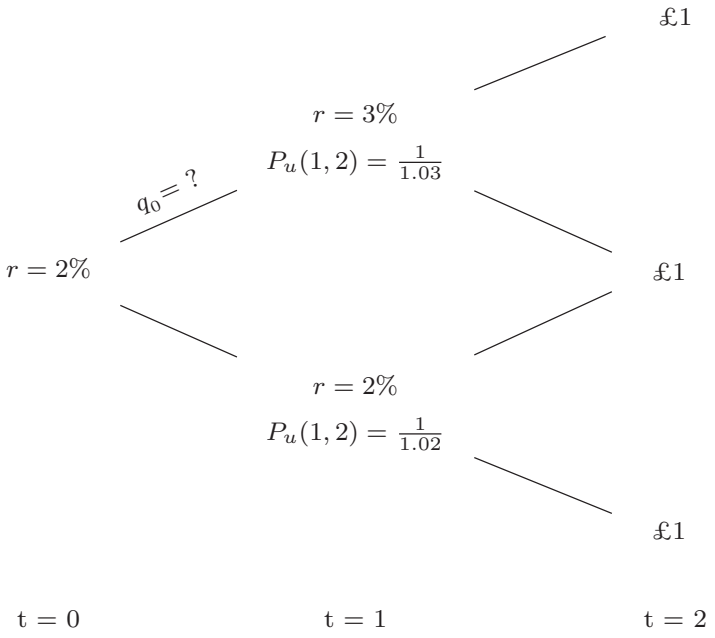
12.4.4.1 Options

Suppose that a two year zero coupon bond is traded for a price equal to $P_\$(0,2) = 0.95500$. We assume that the short-term rate evolves over time according to the tree described in the following diagram.



Suppose that a European call option written on a three year zero coupon bond is traded. This option has a strike price equal to 0.97000, expires in two years, and quotes for $C_{\$}(0,2) = 1.0141 \cdot 10^{-3}$. We can use the price of this derivative, to find the no-arbitrage price of a three year bond which, every year, pays off 3% of the principal of 1.00. Precisely, we use the price of the two year zero coupon bond to recover the risk-neutral probability applying to the first year, and the price of the option to recover the risk-neutral probability applying to the second year. With these probabilities, we determine the no-arbitrage price of the three year bond. (We assume the two probabilities are state-independent, for otherwise we would need the price of additional assets to reverse-engineer state independent risk-neutral probabilities.)

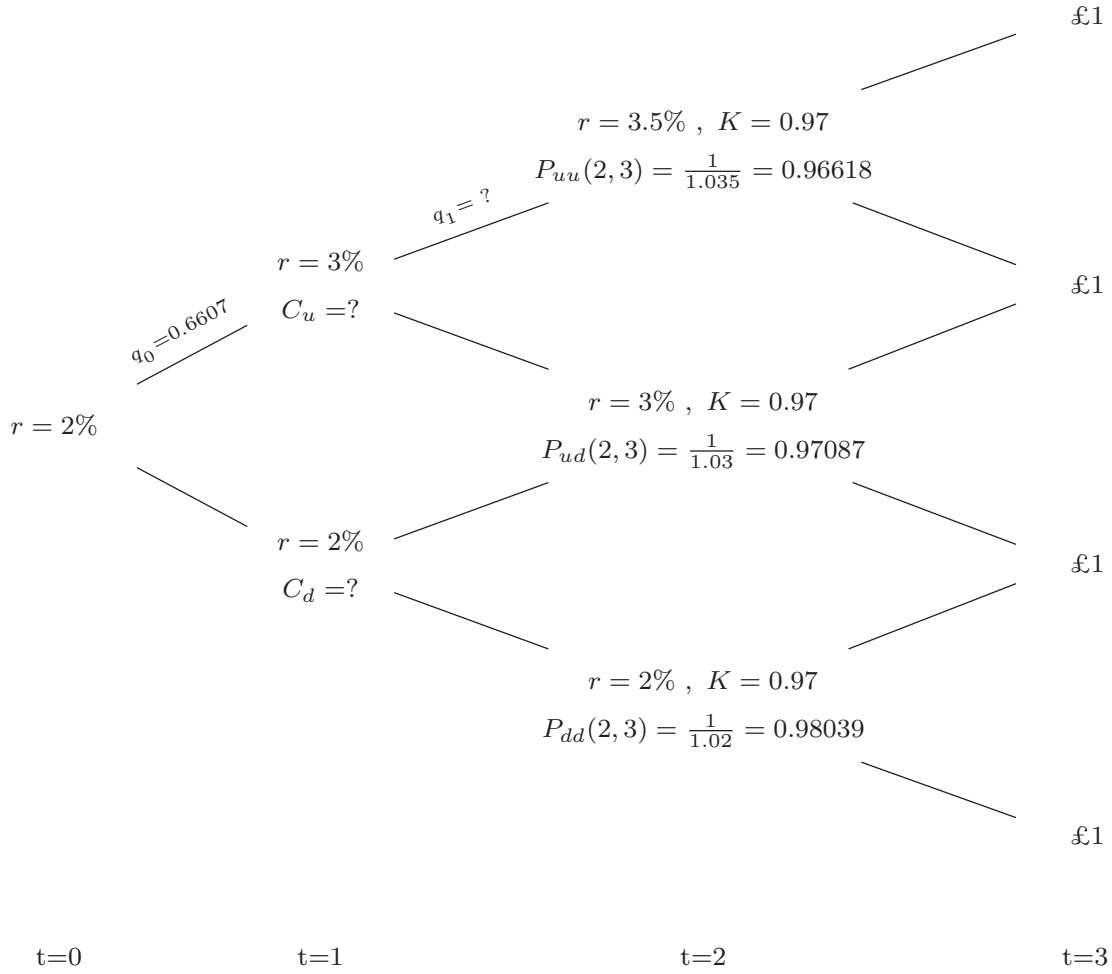
So we know that $P_{\$}(0,2) = 0.95500$. Moreover, as illustrated below, we can extract the price of the 2Y bond in the up- and down- states of the world at time $t = 1$.



We have $P_u(1,2) = \frac{1}{1.03} = 0.97087$, and $P_d(1,2) = \frac{1}{1.02} = 0.98039$. We can now solve for the risk-neutral probability. We have,

$$\begin{aligned} P_{\$}(0,2) &= 0.95500 \\ &= \frac{1}{1.02}\left(q_0 \cdot P_u(1,2) + (1-q_0) P_d(1,2)\right) \\ &= \frac{1}{1.02}\left(q_0 \cdot 0.97087 + (1-q_0) 0.98039\right). \end{aligned}$$

Solving for q_0, yields, $q_0 = 0.6607$. We use this probability, and the price of the option, $C_{\$}(0,2)$, to solve for the risk-neutral probability relevant to the second period, as illustrated below.



In this tree, $K = 0.97000$ is the strike price of the option. The option price at time $t = 1$, in the two states, can be either C_u or C_d, where:

$$\begin{aligned} C_u &= \frac{1}{1.03}\left[q_1 \max\left\{P_{uu}(2,3) - K, 0\right\} + (1-q_1)\max\left\{P_{ud}(2,3) - K, 0\right\}\right] \\ &= \frac{1}{1.03}(1-q_1) \cdot 0.00087. \\ C_d &= \frac{1}{1.02}\left[q_1 \max\left\{P_{ud}(2,3) - K, 0\right\} + (1-q_1)\max\left\{P_{dd}(2,3) - K, 0\right\}\right] \\ &= \frac{1}{1.02}\left[q_1 \cdot 0.00087 + (1-q_1) \cdot 0.01039\right]. \end{aligned}$$

Hence, the option price satisfies,

$$\begin{aligned} C_{\$}(0,2) &= 1.0141 \cdot 10^{-3} \\ &= \frac{1}{1.02}\left(q_0 C_u + (1-q_0) C_d\right) \\ &= \frac{1}{1.02}\left(0.6607 C_u + 0.3393 C_d\right) \\ &= \frac{1}{1.02}\left(0.6607 \frac{1}{1.03}(1-q_1)\cdot 0.00087 + 0.3393 \frac{1}{1.02}\left[q_1 \cdot 0.00087 + (1-q_1)\cdot 0.01039\right]\right). \end{aligned}$$

Solving for q_1 yields, $q_1 = 0.8000$.

Next, we determine the price of the zero maturing at time 3, $P(0,3)$. We use the diagram in Figure 12.11.

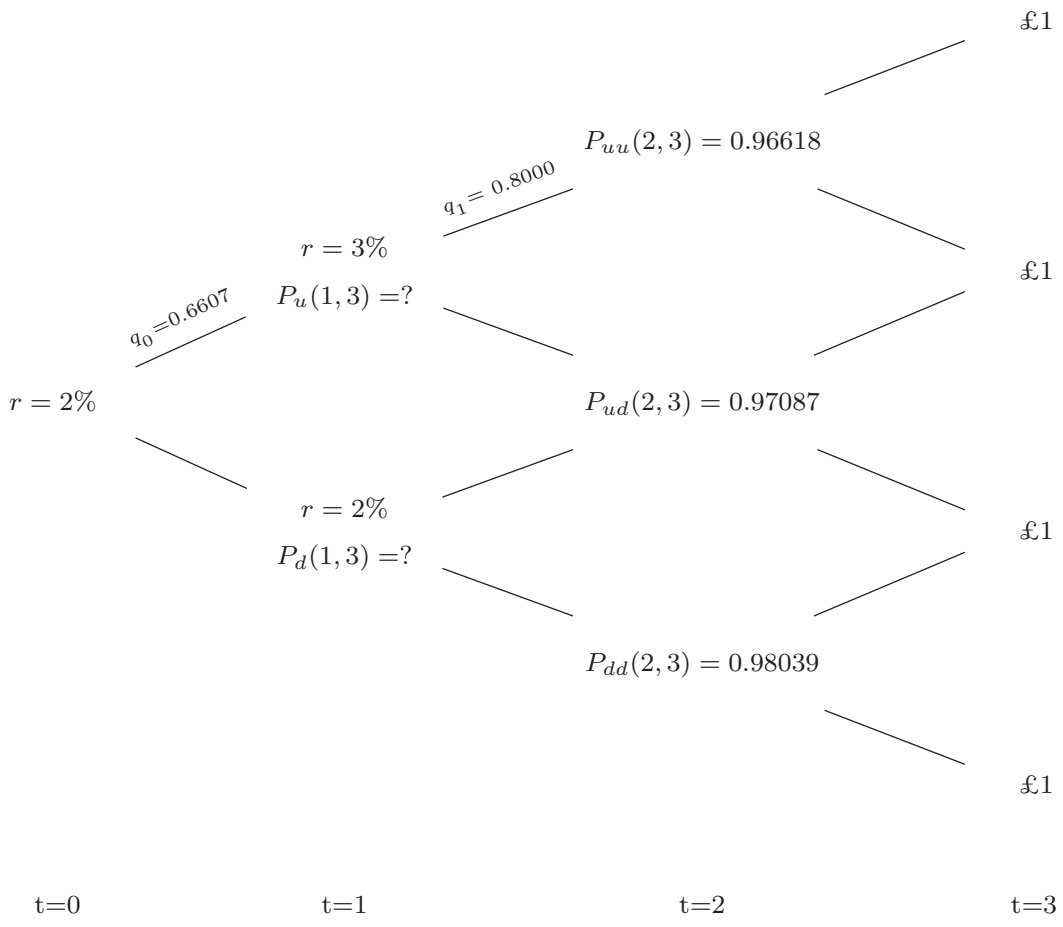


FIGURE 12.11.

We have,

$$\begin{aligned}
P_u(1,3) &= \frac{1}{1.03}\left[q_1 P_{uu}(2,3) + (1-q_1) P_{ud}(2,3)\right] \\
&= \frac{1}{1.03}\left(0.80 \cdot 0.96618 + 0.20 \cdot 0.97087\right) = 0.93895. \\
P_d(1,3) &= \frac{1}{1.02}\left[q_1 P_{ud}(2,3) + (1-q_1) P_{dd}(2,3)\right] \\
&= \frac{1}{1.02}\left(0.80 \cdot 0.97087 + 0.20 \cdot 0.98039\right) = 0.95370.
\end{aligned}$$

The price of a 3Y zero coupon bond, embedded in the market prices, $P_{\$}(0,2)$ and $C_{\$}(0,2)$, is therefore:

$$\begin{aligned}
P(0,3) &= \frac{1}{1.02}\left[q_0 P_u(1,3) + (1-q_0) P_d(1,3)\right] \\
&= \frac{1}{1.02}\left(0.6607 \cdot 0.93895 + 0.3393 \cdot 0.95370\right) = 0.92545.
\end{aligned}$$

We are now ready to evaluate the 3Y bond with 3% coupon rate. It is,

$$\begin{aligned}
P_{\text{coupon=3\%}}(0,3) &= 0.03 \cdot \left[P_{\$}(0,1) + P_{\$}(0,2) + P(0,3)\right] + P(0,3) \\
&= 0.03 \cdot (0.98039 + 0.95500 + 0.92545) + 0.92545 = 1.0113.
\end{aligned}$$

The discretely compounded yield curve implied by the previous calculations is given by $r_{0,1} = 2.00\%$ (1Y); $r_{0,2} : 0.95500 = (1 + r_{0,2})^{-2}$, or $r_{0,2} = 2.328\%$ (2Y); and $r_{0,3} : 0.92545 = (1 + r_{0,3})^{-3}$, or $r_{0,3} = 2.616\%$ (3Y). Note, we are capable of computing the yield curve, without knowing all bond data, but inverting some of them from the price of an option! We can go further. Suppose the price of the "missing" bond becomes available, so to speak. We want to make sure this price is consistent with absence of arbitrage. Suppose, for example, that the market price is $P_{\$}(0,3) > P(0,3) = 0.92545$, say. Then, we can sell short the 3Y zero, and set up a dynamic, self-financing strategy aiming to replicate the 3Y zero, i.e. capable of delivering £1 at maturity.

We would proceed as follows. Consider the tree in Figure 12.11. We build up a portfolio, which is long the option and a MMA. We assume the 3Y bond "converges" to the values P_{uu}, P_{ud} and P_{dd} in Figure 12.11, for otherwise we might implement a trivial arbitrage from time $t = 2$ to $t = 3$. At time $t = 0$, we go long Δ_0 options and M_0 units of the MMA, so as to make sure that the portfolio delivers $P_u(1,3)$ in the upstate of $P_d(1,3)$ in the downstate—thereby ensuring that the price of the bond is replicated at time $t = 1$. The value of this replicating strategy is, of course, $P(0,3)$, so by short-selling the 3Y bond at $t = 0$, we obtain an initial profit, equal to $P_{\$}(0,3) - P(0,3)$. Suppose, then, that at time $t = 1$, we are in the up-node, such that the bond price is $P_u(1,3)$. In this node, we can build up another portfolio long Δ_1 options and M_1 units of the MMA, aiming to replicate the price of the bond at time $t = 2$—either $P_{uu}(2,3)$ or $P_{ud}(2,3)$. The value of this replicating portfolio would be just $P_u(1,3)$, which is what is obtained by the replicating strategy at implemented at time $t = 0$. The strategy is clearly self-financed, as the following calculations reveal. By construction, $\Delta_0 C_u(1) + M_0(1+r) = P_u(1,3) = \Delta_1 C_u(1) + M_1$ (with $r = 2\%$), and $\Delta_1 \tilde{C}_{u\cdot}(2) + M_1(1+r) = P_{u\cdot}(2,3)$, where $r = 3\%$, and: $\tilde{C}_{u\cdot}(2), \tilde{P}_{u\cdot}(2,3)$ are either $C_{uu}(2), P_{uu}(2,3)$, or $C_{ud}(2), P_{ud}(2,3)$, at time $t = 2$,

with straight forward notation. Therefore, we have that,

$$\begin{aligned}\tilde{P}_{u\cdot}(2,3) - P_u(1,3) &= \Delta_1\left(\tilde{C}_{u\cdot}(2) - C_u(1)\right) + rM_1 \\ &= \Delta_1\left(\tilde{C}_{u\cdot}(2) - C_u(1) - rC_u(1)\right) + rP_u(1,3).\end{aligned}$$

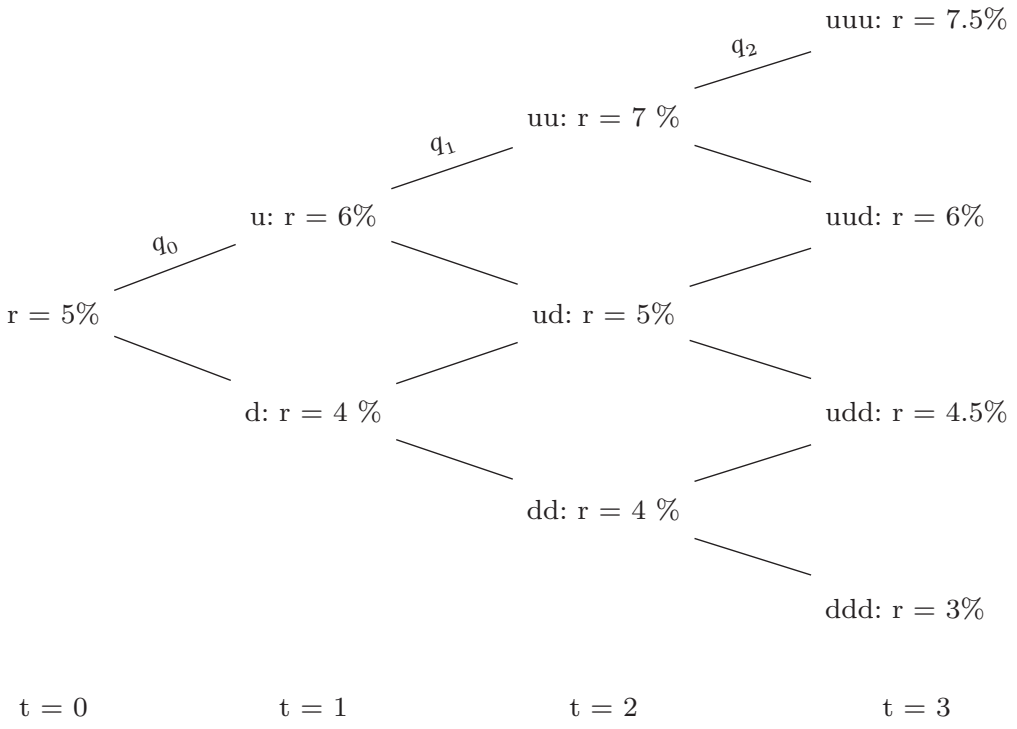
Likewise, if, instead, at time $t = 1$, we end up in the down-node, where the bond price (and the value of the strategy implemented at time $t = 0$) is $P_d(1,3)$, we can invest $P_d(1,3)$ in options and MMA so as to replicate the price of the bond at time $t = 2$—either $P_{ud}(2,3)$ or $P_{dd}(2,3)$. The presence of dynamically complete markets allows us to implement an arbitrage.

12.4.4.2 Arrow-Debreu securities, and the pricing of interest rate derivatives

Arrow-Debreu securities are assets that only pay over a specific state of the world, as explained in Chapter 2. We shall deal with these securities in more detail in Section 12.7, because they allow us to implement perfectly fitting models quite elegantly. This section is a first introduction to them. We make use of Arrow-Debreu securities, firstly, to extract risk-neutral probabilities and, secondly, to price quite basic interest rate derivatives, such as a "caplet" or a "forward rate agreement." Likewise, the pricing of interest rate derivatives covered in this section is a preliminary introduction, as the next chapter will systematically deal with it, within a continuous time setting.

Extracting risk-neutral probabilities from Arrow-Debreu securities

Assume that the discretely compounded one-year rate, or the "short-term rate," evolves over time as described by the following tree:



Assume three securities are available for trading: (i) a zero coupon bond expiring in two years, quoting for 0.91000; (ii) a zero coupon bond expiring in three years, quoting for 0.86500;

(iii) an Arrow-Debreu security, paying off £1 only at time 3 in the state "uuu" of the previous diagram, where the short-term rate equals 7.5%, quoting for 0.10000.

Assume that the risk-neutral probabilities of upward movements in the short-term rate change over time and take three values, q_0, q_1 and q_2, but are independent of the state of nature, as illustrated in the previous diagram. We can calibrate these probabilities, through the previously given available market data. First, we derive q_0 using the price of the two year bond, $P_{\$}(0,2)$ say, as follows:

$$P_{\$}(0,2) = 0.91000 = \frac{1}{1.05}\left(q_0 P_u(1,2) + (1-q_0) P_d(1,2)\right),$$

where

$$P_u(1,2) = \frac{1}{1.06} = 0.94340, \quad P_d(1,2) = \frac{1}{1.04} = 0.96154.$$

Solving for q_0 yields $q_0 = 0.33284$. Next, we calibrate q_1 to match the price of the three year bond, $P_{\$}(0,3)$ say. We have,

$$\begin{aligned} P_{\$}(0,3) = 0.86500 &= \frac{1}{1.05}\left(q_0 P_u(1,3) + (1-q_0) P_d(1,3)\right) \\ &= \frac{1}{1.05}\left(0.33284 \cdot P_u(1,3) + 0.66716 \cdot P_d(1,3)\right), \end{aligned}$$

where

$$\begin{aligned} P_u(1,3) &= \frac{1}{1.06}\left(q_1 P_{uu}(2,3) + (1-q_1) P_{ud}(2,3)\right), \\ P_d(1,3) &= \frac{1}{1.04}\left(q_1 P_{ud}(2,3) + (1-q_1) P_{dd}(2,3)\right), \end{aligned}$$

and

$$P_{uu}(2,3) = \frac{1}{1.07} = 0.93458, \quad P_{ud}(2,3) = \frac{1}{1.05} = 0.95238, \quad P_{dd}(2,3) = \frac{1}{1.04} = 0.96154.$$

Solving for q_1 yields $q_1 = 0.66507$. Finally, we calibrate q_3 through the price of the Arrow-Debreu security paying in state "uuu." This price, denoted as $p^{\$}_{uuU}(3)$, is given by:

$$p^{\$}_{uuu}(3) = 0.10000 = \frac{1}{1.05}\frac{1}{1.06}\frac{1}{1.07} q_0 q_1 q_2 = \frac{1}{1.05}\frac{1}{1.06}\frac{1}{1.07} 0.33284 \times 0.66507 \times q_2,$$

i.e. $q_2 = 0.53798$.

Pricing interest rate derivatives

Next, we use the previously calibrated probabilities to price some interest rate derivatives. First, consider a "caplet" contingent on the rates prevailing at time $t = 3$, paying off at time $t = 4$, with strike rate equal to 5%, and notional value equal to £100. The payoff of this derivative instrument at $t = 4$ is $\max\{L - 5\%, 0\}$, where L is the rate at $t = 3$. Therefore, the discounted payoffs at time $t = 3$ are:

uuu: $\frac{1}{1.075}\max\{7.5 - 5, 0\} = 2.32560$

uud: $\frac{1}{1.06}\max\{6 - 5, 0\} = 0.94340$

udd: $\frac{1}{1.045}\max\{4.5-5,0\}=0$

ddd: $\frac{1}{1.03}\max\{3-5,0\}=0$

As for time $t=2$ and $t=1$, we have:

$t=2$ uu: $A_1 \equiv \frac{1}{1.07}\left(q_2 2.32560+(1-q_2)\,0.94340\right)=1.5766$

ud: $A_2 \equiv \frac{1}{1.05}\left(q_2 0.94340+(1-q_2)\,0\right)=0.48336$

dd: $A_3 \equiv 0$

$t=1$ u: $B_1 \equiv \frac{1}{1.06}\left(q_1 A_1+(1-q_1)\,A_2\right)=1.1419$

d: $B_2 \equiv \frac{1}{1.04}\left(q_1 A_2+(1-q_1)\,A_3\right)=0.30910$

Therefore, the price of the caplet is:

$$\text{Caplet Price}=\frac{1}{1.05}\left(q_0 B_1+(1-q_0)\,B_2\right)=0.55837.$$

Next, consider a forward rate agreement, whereby at time $t=0$, two counterparties agree that at time $t=4$, they will exchange with each other the variable short-term rate prevailing at time $t=3$, against a fixed interest rate equal to K. We can use the previously calibrated probabilities to determine the *forward rate*, i.e. the level of K that makes the value of this agreement equal to zero at time $t=0$. Take the case of a "payer" forward agreement, one for which the discounted payoffs at time $t=3$ are:

uuu: $\pi_1 \equiv \frac{1}{1.075}\left(7.5-K\right)$

uud: $\pi_2 \equiv \frac{1}{1.06}\left(6-K\right)$

udd: $\pi_3 \equiv \frac{1}{1.045}\left(4.5-K\right)$

ddd: $\pi_4 \equiv \frac{1}{1.03}\left(3-K\right)$

At time $t=2$ and $t=1$, the payoffs are:

$t=2$ uu: $\tilde{A}_1 \equiv \frac{1}{1.07}\left(q_2\pi_1+(1-q_2)\,\pi_2\right)=5.9519-0.87506\times K$

ud: $\tilde{A}_2 \equiv \frac{1}{1.05}\left(q_2\pi_2+(1-q_2)\,\pi_3\right)=4.7950-0.90443\times K$

dd: $\tilde{A}_3 \equiv \frac{1}{1.04}\left(q_2\pi_3+(1-q_2)\,\pi_4\right)=3.5215-0.92632\times K$

$t=1$ u: $\tilde{B}_1 \equiv \frac{1}{1.06}\left(q_1\tilde{A}_1+(1-q_1)\,\tilde{A}_2\right)=5.2495-0.83481\times K$

d: $\tilde{B}_2 \equiv \frac{1}{1.04}\left(q_1\tilde{A}_2+(1-q_1)\,\tilde{A}_3\right)=4.2004-0.87669\times K$

We can now express the value of the contract as a function of the fixed rate K, as follows:

$$\text{Fwd}\left(K\right)\equiv\frac{1}{1.05}\left(q_0\tilde{B}_1+(1-q_0)\,\tilde{B}_2\right)=4.3329-0.82167\times K. \tag{12.25}$$

The forward rate is, simply, the value of K such that $\text{Fwd}\left(K\right)=0$, i.e. $K=5.27330\%$. More generally, we can determine the value of the forward rate agreement in Eq. (12.25), for any

value of K. For example, we have that $\text{Fwd}\,(K=6) = 4.3329 - 0.82167 \times 6 = -0.59712$, in percentage terms.

Finally, we can derive the price of a bond expiring at time $t=4$, $P\,(0,4)$, through the relation, $\frac{P_{\$}(0,3)-P(0,4)}{P(0,4)} = \frac{P_{\$}(0,3)}{P(0,4)} - 1 = 5.27330\%$, leading to $P\,(0,4) = 0.82170$, which is indeed the same figure we can obtain by solving the tree for $P\,(0,4)$, as we now show. We have, for time $t=3$,

uuu: $P_{uuu}\,(3,4) = \frac{1}{1.075} = 0.93023$

uud: $P_{uud}\,(3,4) = \frac{1}{1.06} = 0.94340$

udd: $P_{udd}\,(3,4) = \frac{1}{1.045} = 0.95694$

ddd: $P_{ddd}\,(3,4) = \frac{1}{1.03} = 0.97087$

Then, we can solve, recursively, as usual:

$t=2$ uu: $P_{uu}\,(2,4) = \frac{1}{1.07}\left(q_2 P_{uuu}\,(3,4) + (1-q_2)\,P_{uud}\,(3,4)\right) = 0.87506$

ud: $P_{ud}\,(2,4) = \frac{1}{1.05}\left(q_2 P_{uud}\,(3,4) + (1-q_2)\,P_{udd}\,(3,4)\right) = 0.90443$

dd: $P_{dd}\,(2,4) = \frac{1}{1.04}\left(q_2 P_{udd}\,(3,4) + (1-q_2)\,P_{ddd}\,(3,4)\right) = 0.92632$

$t=1$ u: $P_{u}\,(1,4) = \frac{1}{1.06}\left(q_1 P_{uu}\,(2,4) + (1-q_1)\,P_{ud}\,(2,4)\right) = 0.83481$

d: $P_{d}\,(1,4) = \frac{1}{1.04}\left(q_1 P_{ud}\,(2,4) + (1-q_1)\,P_{dd}\,(2,4)\right) = 0.87669$

Finally, we calculate the price $P\,(0,4) = \frac{1}{1.05}\left(q_0 P_u\,(1,4) + (1-q_0)\,P_d\,(1,4)\right) = 0.82167 \approx 0.82170$, by rounding.

12.4.5 Extensions to trinomial trees

Trinomial trees might be useful to model mean-reversion. They might, also, be considered as approximation to partial differential equations.

12.5 The Ho and Lee model

Ho and Lee (1986) develop a revolutionary approach to modeling yield curve movements. This approach does not aim to an economic theory to *explain* the yield curve that we observe. Rather, it takes the yield curve as given, and shifts the focus towards the modeling of no-arbitrage *movements* of the entire yield curve. As explained, we need to "match" model prices to market prices, to avoid having derivatives with underlyings deviating from market prices. The next chapter derives the Ho & Lee model in continuous time, as this derivation allows us to illustrate a general approach to interest rate modeling, developed later by Heath, Jarrow and Morton (1992). The original derivation of the model is, however, in discrete time, and it is quite instructive to follow this approach here, and to compare with the alternative calibration methods of this chapter.

The main idea underlying Ho and Lee is that the movements of the yield curve can be modeled through a binomial tree, much in the spirit of the Cox, Ross and Rubenstein (1979) tree representation of Black and Scholes (1973). However, in Black and Scholes (1973) and Cox, Ross and Rubenstein (1979), the asset underlying the option contract is a traded risk, such

that the underlying price satisfies the martingale condition. Instead, interest rate derivatives generally depend on non-traded risks, which are not martingales. Moreover, the mere presence of boundary conditions induce bond return volatility to be time-varying.

Ho and Lee address these issues by modeling the movements of the entire collection of bond prices. We have three ways to achieve this task: (i) by making risk-neutral probs time-varying, for a given tree with predetermined values of the short-term rate (as in the previous sections); (ii) by assuming a constant risk-neutral probability, and searching for the values of the short-term rate on the tree; (iii) by a combination of (i) and (ii) as in the implied binomial tree of Chapter 11. The Ho and Lee model relies on the second way. The key element of this model is the determination of the no-arbitrage ups and downs of the entire yield curve, through modeling bond prices.

12.5.1 The tree

The key element of the model is the determination of the no-arbitrage ups and downs of the entire yield curve, obtained by directly modeling bond prices of arbitrary expiration. Note that once bond prices are obtained, forward rates are obtained as a result. Therefore, the Ho and Lee model is a model of forward rate movements. It is a simple but powerful remark, because the key point of the model is, then, to re-express bond prices again, as a function of future forward rates. In this sense, the Ho and Lee model is a representation of current bond prices (in terms of forward rates), rather than a model of current bond prices.

Assume that the price of any zero evolves according to a binomial tree. Let $P_j(t,T)$ be the price of a pure discount bond as of time t, with time to maturity $T-t$, after j *upstate bond price movements*. Let $j \sim B(t,1-q)$, a binomial random variable, meaning that $\Pr(j) = \binom{t}{j}(1-q)^j q^{t-j}$, such that

$$E(j) = (1-q)t, \quad Var(j) = q(1-q)t,$$

where $1-q$ is the risk-neutral probability of a single upstate movement *of the price*. Therefore we have,

$$\begin{array}{lll} & & P_{j+1}(t+1,T) \\ & \overset{1-q}{\nearrow} & \\ P_j(t,T) & & \\ & \underset{q}{\searrow} & \\ & & P_j(t+1,T) \end{array}$$

That is, if at time t, the number of upstate movements is equal to j, then, at time $t+1$, the number of upstate movements can either jump to $j+1$, with probability $1-q$, or stay at j, with probability q. (Therefore, we are now following the convention to have high values of bond prices in the upper parts of the tree.) Note, further, that after one period, any zero is one period closer to maturity. At maturity, $t=T$, the price of any zero is worth one unit of *numéraire*, viz

$$P_j(T,T) = 1, \quad \text{for all } j \text{ and } T.$$

Note, in the previous tree, it shall not necessarily hold that $P_j(t+1,T) < P_j(t,T)$. On the contrary, we would expect that especially when time-to-maturity approaches zero, $P_j(t+1,T) > P_j(t,T)$, as the price of the zero needs to converge to par.

12.5.2 Price movements and the martingale restriction

In the absence of arbitrage opportunities, the expected return on the zero at t must equal the short-term rate, viz $P_j(t,T) = e^{-r_j(t)} E_q(P_\cdot(t+1,T))$, or

$$P_j(t,T) = P_j(t,t+1)\left[(1-q)\,P_{j+1}(t+1,T) + qP_j(t+1,T)\right], \tag{12.26}$$

where $P_j(t,t+1) = e^{-r_j(t)}$, and $r_j(t)$ is the continuously compounded short-term rate at time t after j upward movements. We call this condition the *martingale restriction*.

Let us introduce notation for the movements of the price of any zero along the tree,

$$\underbrace{\frac{P_{j+1}(t+1,T)}{P_j(t,T)} = u(T-t)\,\frac{1}{P_j(t,t+1)}}_{\text{up at } t} \quad \text{and} \quad \underbrace{\frac{P_j(t+1,T)}{P_j(t,T)} = d(T-t)\,\frac{1}{P_j(t,t+1)}}_{\text{down at } t}. \tag{12.27}$$

The two functions, $u(\cdot)$ and $d(\cdot)$, also called "perturbation functions," are taken to be state-independent. They capture the fact that in the case of uncertainty, the price of the zero can either go up or down with respect to the risk-free of return. In other words, Eqs. (12.27) tell us that the discounted gross return from going long a bond is:

$$\underbrace{\frac{P(t+1,T)}{P(t,T)}}_{\text{Gross return}} \cdot \underbrace{P(t,t+1)}_{\text{Discount}} = \begin{cases} u(T-t) & \text{with probability } 1-q \\ d(T-t) & \text{with probability } q \end{cases}$$

where the two functions $u(T-t)$ and $d(T-t)$ have to be endogenously determined. If there was no uncertainty, we would have $u(T-t) = d(T-t) = 1$, for all $t \leq T$. In general, we have that $d(T-t) \leq 1 \leq u(T-t)$, as we shall now demonstrate.

One period before the expiration date, i.e. at $t = T-1$, our price is certain to jump to one, with jump size equal to the short-term rate $r_j(t)$. Hence, the following boundary condition for the two functions $u(\cdot)$ and $d(\cdot)$ holds:

$$u(1) = d(1) = 1. \tag{12.28}$$

In terms of the two functions $u(\cdot)$ and $d(\cdot)$, the martingale restriction in Eq. (12.26) is,

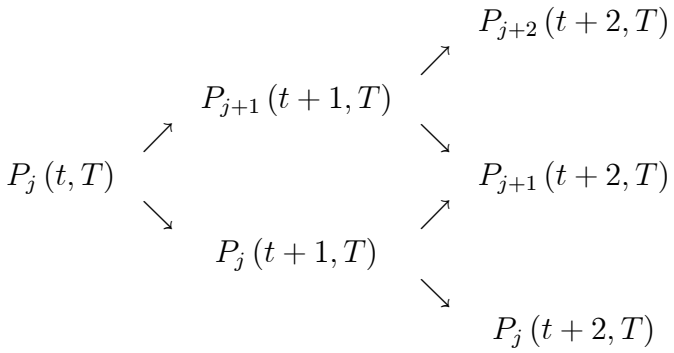
$$1 = (1-q)\,u(T-t) + qd(T-t), \quad t \leq T. \tag{12.29}$$

This relation matches the "martingale restriction" introduced in Chapter 11, and applying to stocks, the current values of which are tied down to their future up and down movements through the risk-neutral probability. The difference in this context is that the up and down movements of the prices depend on the asset maturity, through the two functions $u(T-t)$ and $d(T-t)$, which are endogenous, which makes the evaluation problem more intricate.

12.5.3 The recombining condition and interest rate volatility

Ho and Lee consider a recombining tree: the price $P_j(t,T)$ we are looking for depends only on j, not on the exact sequence of up and down movements leading to j upstate movements. To summarize, we are looking for two functions $u(T-t)$ and $d(T-t)$ such that (i) the no-arbitrage condition in Eq. (12.29) holds true and (ii) the tree is recombining. We now elaborate

on the arguments that lead to the recombining property of the tree.



The recombining property of the tree implies that the bond price at time $t+2$ in the event of $j+1$ jumps, i.e. $P_{j+1}(t+2,T)$, can be generated by one of the two paths, which we track by using the two Eqs. (12.27):

(i) The "up & down" path, $P_j(t,T) \to P_{j+1}(t+1,T) \to P_{j+1}(t+2,T)$, where,

$$P_{j+1}(t+2,T) = \underbrace{\overbrace{u(T-t)\frac{P_j(t,T)}{P_j(t,t+1)}}^{\equiv P_{j+1}(t+1,T),\text{ up at } t} \times d(T-t-1)\frac{1}{P_{j+1}(t+1,t+2)}}_{\text{down at } t+1}. \tag{12.30}$$

(ii) The "down & up" path $P_j(t,T) \to P_j(t+1,T) \to P_{j+1}(t+2,T)$, where

$$P_{j+1}(t+2,T) = \underbrace{\overbrace{d(T-t)\frac{P_j(t,T)}{P_j(t,t+1)}}^{\equiv P_j(t+1,T),\text{ down at } t} \times u(T-t-1)\frac{1}{P_j(t+1,t+2)}}_{\text{up at } t+1}. \tag{12.31}$$

By equating Eq. (12.30) and (12.31), we obtain,

$$\frac{u(T-t)}{d(T-t)} = \frac{u(T-t-1)}{d(T-t-1)}\delta^{-1}, \quad \delta^{-1} \equiv \frac{P_{j+1}(t+1,t+2)}{P_j(t+1,t+2)} \in (1,\infty), \tag{12.32}$$

where we take the ratio δ to be constant, and return to its interpretation in a moment. Eq. (12.32) is a finite-difference equation for the ratio $u(T-t)/d(T-t)$. Given the boundary conditions in Eqs. (12.28), the solution for this ratio is:

$$\frac{u(T-t)}{d(T-t)} = \delta^{-(T-t-1)}. \tag{12.33}$$

We claim that δ in Eq. (12.32) is a contant relating to the *volatility of the short-term rate*. Indeed, by taking logs, we obtain:

$$\ln \delta^{-1} = -\left[r_{j+1}(t+1) - r_j(t+1)\right]. \tag{12.34}$$

Moreover, conditionally upon time t and price jumps equal to $j \leq t$, the short-term rate is binomially distributed, and can take on two values: (i) $r_{j+1}(t+1)$ with probability $1-q$ and

$r_j(t+1)$ with probability q. Then, the conditional variance of the short-term rate, $var_t[\tilde{r}(t+1)]$ say, satisfies,

$$\sqrt{var_t[\tilde{r}(t+1)]} = \sqrt{q(1-q)} \cdot |r_{j+1}(t+1) - r_j(t+1)| = \sqrt{q(1-q)} \cdot \ln \delta^{-1},$$

where the second equality follows by Eq. (12.34).

There are no economic reasons to assume that δ is constant. Assuming δ to be time-varying and possibly state dependent, might actually help match more than the initial yield curve (say European options), similarly as for the implied binomial trees for equity reviewed in Chapter 11. At the same time, assuming δ to be time-varying would likely lead to models without closed-form solutions. We shall return to the topic of models without closed-form solution in Section 12.7. The next section shows that in the context where δ is constant, the bond price has a solution expressed in closed-form.

12.5.4 Model's solution

Eq. (12.33) yields the condition under which the tree is recombining. On the other hand, the martingale restriction in Eq. (12.29) is needed to prevent arbitrage. Therefore, we are left with the following system of two equations, Eq. (12.33) and Eq. (12.29), and two unknowns, $u(\cdot)$ and $d(\cdot)$,

$$\begin{cases} u(T-t) = \delta^{-(T-t-1)} d(T-t) \\ (1-q)\, u(T-t) + q d(T-t) = 1 \end{cases}$$

The solution to this system is,

$$u(T-t) = \frac{1}{(1-q) + q\delta^{T-t-1}}, \quad d(T-t) = \frac{\delta^{T-t-1}}{(1-q) + q\delta^{T-t-1}}. \tag{12.35}$$

So the problem is now solved. Once we assign values to q and δ, and an initial bond price $P(t,T)$, we plug Eqs. (12.35) into Eqs. (12.27), and populate the tree—we plug bond prices on each node of the tree. Once this is done, we can price interest rate derivatives, i.e. assets with payoffs indexed to bond prices or interest rates over a given set of nodes. Note that Eqs. (12.27) dictate the law of motion of the entire structure of bond prices along the trees, and they are silent regarding the initial condition, i.e. at time $t = 0$. The natural initial condition is the market price of the bond for each maturity, say $P_0(0,T) = P_\$(0,T)$ for all T.

We can actually do more, and develop a closed-form solution for the bond price. Let $F_S^j(t)$ be the forward rate as of time t after the occurrence of j upward price movements, and let the continuously compounded forward rate $\hat{F}_S^j(t)$ be defined as,

$$\hat{F}_S^j(t) \equiv \ln\left(1 + F_S^j(t)\right), \quad j \leq t.$$

In Appendix 2, we show that,

$$P_j(t,T) = \frac{P(0,T)}{P(0,t)} e^{-\sum_{S=t}^{T-1}\left(\hat{F}_S^j(t) - \hat{F}_S(0)\right)}. \tag{12.36}$$

That is, we can express the price in closed-form, once we are able to do the same with the forward rate changes, $\hat{F}_S^j(t) - \hat{F}_S(0)$. In Appendix 2, we show that,

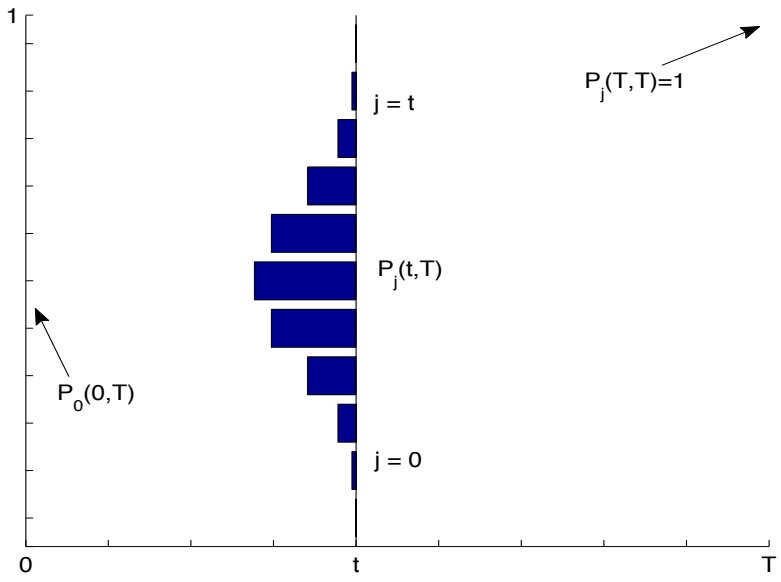
$$\hat{F}_S^j(t) = \hat{F}_S(0) + \ln \frac{u(S+1-t)}{u(S+1)} - (t-j)\ln\delta, \quad j \leq t. \tag{12.37}$$

We replace Eq. (12.37) into Eq. (12.36), use the solution for the perturbation function $u(\cdot)$ in Eqs. (12.35), and using the condition that initial node of the tree is the same as the market price for each T, we find that

$$P_j(t,T) = \frac{P_{\$}(0,T)}{P_{\$}(0,t)} \delta^{(T-t)(t-j)} \prod_{S=t}^{T-1} \frac{(1-q)+q\delta^{S-t}}{(1-q)+q\delta^{S}}. \tag{12.38}$$

From the perspective of time 0, the price of the zero at t, in each state j, is only a function of the initial yield curve, the volatility parameter δ, and the risk-neutral probability q. Note in particular how important this volatility is. It is, in practice, the only parameter to be determined once we fix q, which leads to an interesting parallel between this model and that of Black and Scholes (1973). In both models, the inputs are the volatilities of the fundamentals (the short-term interest rate, in Ho & Lee, and stock returns, in Black & Scholes). However, in Ho & Lee, the input does not link to the short-term rate, but the entire yield curve, and the output relates to future bond price movements. In Black & Scholes, the input is that of the fundamental, the stock price, and the output is the initial option price.

Finally, the model displays the properties we would require from a perfectly fitting one, and illustrated by the next picture. First, it matches each price at time zero, as it can be verified by collapsing $t = 0$. Second, it bridges to one when $t = T$. Third, it is random at any time $0 < t < T$, $P_j(t,T)$ with probability $\binom{t}{j}(1-q)^j q^{t-j}$. Naturally, the values taken by $P_j(t,T)$ are determined by the equilibrium of the model, i.e. Eq. (12.38). All in all, the model predicts random outcomes at time t, ensuring at the same time that the initial yield curve is fitted without errors. The reason we might be interested in random outcomes at time t, is that we might have to evaluate options expiring at that date, through a model that predicts the underlying to be pinned down without errors as explained.



12.5.5 Calibration of the model

We need to "estimate" the value of δ. We can proceed as follows. Consider Eq. (12.38), and let $T = t + 1$. We have,

$$P_j(t, t+1) = \frac{P(0, t+1)}{P(0,t)} \delta^{t-j} \frac{1}{(1-q) + q\delta^t}.$$

The continuously compounded short-term rate predicted by the model is,

$$r_j(t) \equiv -\ln P_j(t, t+1) = \hat{F}_t(0) + \ln\left((1-q) + q\delta^t\right) - (t-j)\ln\delta, \quad j \leq t, \tag{12.39}$$

where $\hat{F}_t(0) \equiv \ln P(0,t) - \ln P(0, t+1)$. We also have,

$$r_j(1) - r(0) = \hat{F}_1(0) - \hat{F}_0(0) + \ln((1-q) + q\delta) + \ln\delta^{-1} \cdot (1-j).$$

Hence, the parameter δ can be chosen such that the volatility of the short-term rate predicted by the model matches exactly the volatility of the short-term rate that we see in the data. Concretely, we can take $\hat{\delta} = \exp(-\operatorname{Std}(\Delta r)/\sqrt{q(1-q)})$, where $\operatorname{Std}(\Delta r)$ is the standard deviation of the short-term rate in the data.

Note, then, the interesting feature of the model. The Ho and Lee model doesn't take any a priori stance on the dynamics of the short-term rate. Rather, it imposes: (i) the martingale restriction on bond prices, an economic restriction, Eq. (12.29); and (ii) the simplifying assumption the tree is recombining, a technical condition, Eq. (12.27). These two conditions suffice to to tell what to expect from the dynamics of the short-term rate. While deliberately simple, the Ho and Lee model is quite powerful. The modern approach to interest rate modeling simply aims to make the Ho and Lee methodology more accurate for practical purposes.

12.5.6 An example

Assume that three zero coupon bonds are available for trading, with current market prices: (i) $P_\$(0,1) = 0.9851$ (the price of a 6M zero), (ii) $P_\$(0,2) = 0.9685$ (the price of a 1Y zero), and (iii) $P_\$(0,3) = 0.9445$ (the price of the 1.5Y zero). We know that the price of one-period zero at time t, in the event of j upward *price*-jumps from the current date to t, is:

$$P_j(t, t+1) = \frac{P_\$(0, t+1)}{P_\$(0,t)} \delta^{t-j} \frac{1}{(1-q) + q\delta^t}, \quad j \leq t, \tag{12.40}$$

where $P_\$(0,t)$ is the current market price of a zero expiring at time t, with t equal to six months, one year and eighteen months, in this example. We assume that $q = \frac{1}{2}$ and $\delta = 0.9802$.

12.5.6.1 The dynamics of the short-term rate

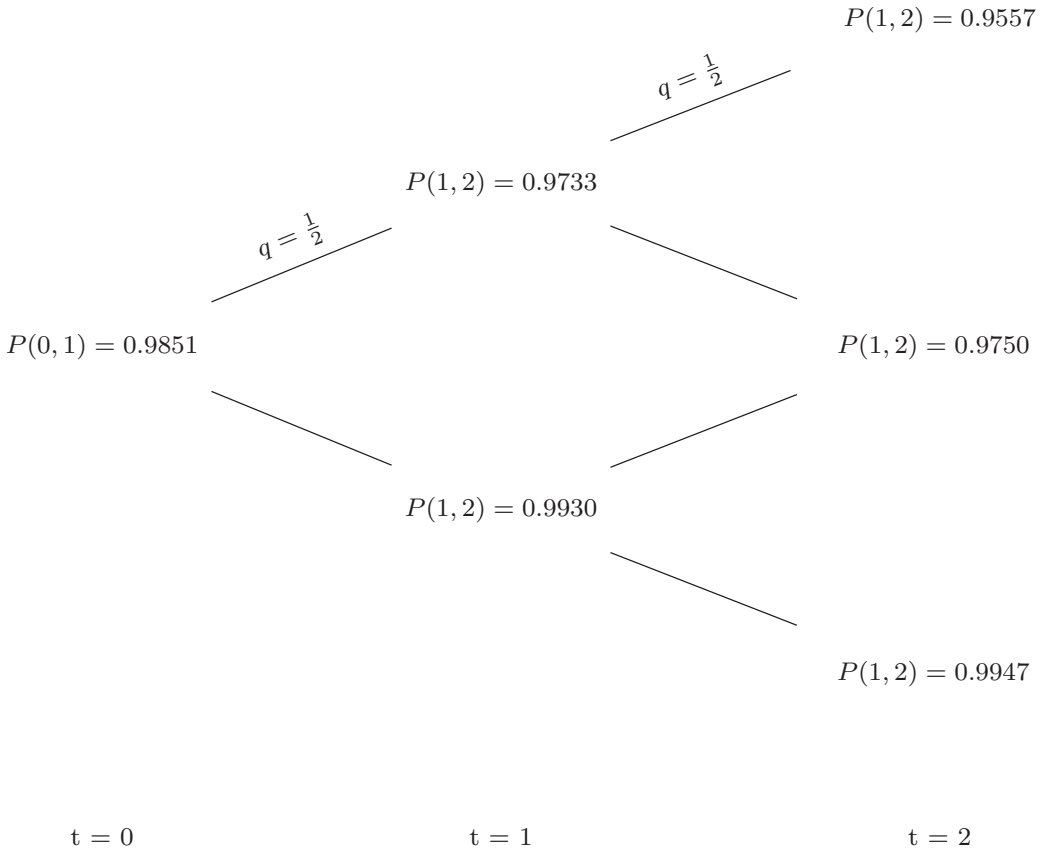
We want to determine the developments of the short-term rate on a recombining tree for as many periods as we can, given the market price of the zeros we observe. We use Eq. (12.40) to find the one-period zeros in each node.

- $t = 0$. We have, trivially, $P(0,1) = P_\$(0,1) = 0.9851$.
- $t = 1$. We have three cases:
  - $j = 0$: $P_0(1,2) = 2\frac{P_\$(0,2)}{P_\$(0,1)}\delta\frac{1}{1+\delta} = 0.9733$

- $j = 1$: $P_1(1,2) = 2\frac{P_\$(0,2)}{P_\$(0,1)}\frac{1}{1+\delta} = 0.9930$

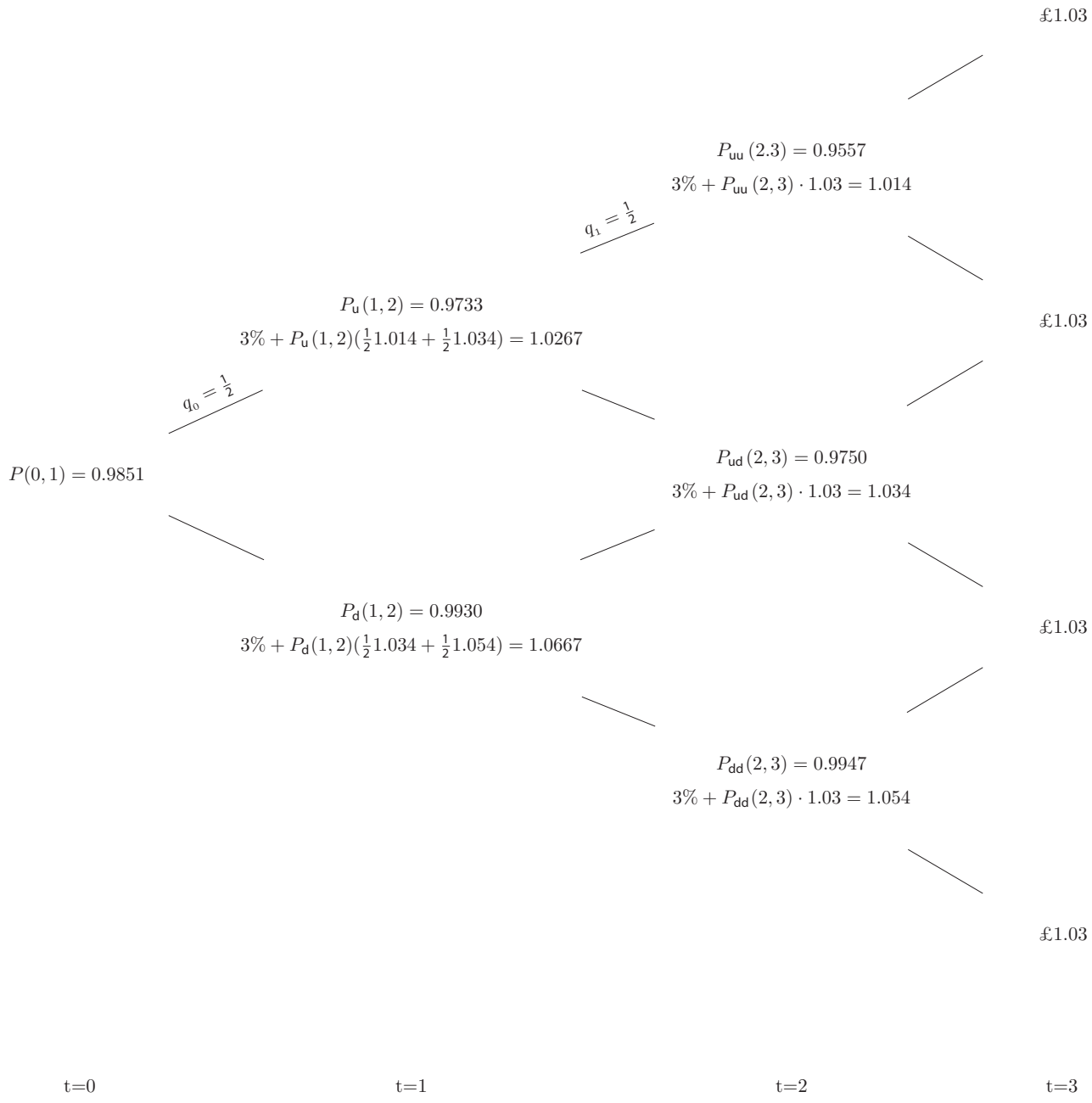
- $t = 2$. We have three cases:
  - $j = 0$: $P_0(2,3) = 2\frac{P_\$(0,3)}{P_\$(0,2)}\delta^2\frac{1}{1+\delta^2} = 0.9557$
  - $j = 1$: $P_1(2,3) = 2\frac{P_\$(0,3)}{P_\$(0,2)}\delta\frac{1}{1+\delta^2} = 0.9750$
  - $j = 2$: $P_2(2,3) = 2\frac{P_\$(0,3)}{P_\$(0,2)}\frac{1}{1+\delta^2} = 0.9947$

So we face the tree below.



12.5.6.2 Pricing a coupon bearing bond

Suppose, now, that we want to find the price of some additional bond, e.g., a 1.5Y bond which pays (semiannually) coupons at 3% of the principal of £1. First, we need to find the value of this bond in each node of the tree. Note, at each node, the price equals (i) the discounted expectation of its future value (including coupons), and (ii) the current coupons, as illustrated in the tree below. That is, the convention, here, is that the bond purchased at time t doesn't give the owner the right to receive any coupon at time t, only from time $t+1$ onwards.



Since the bond does not pay coupons at time zero, its current price is,

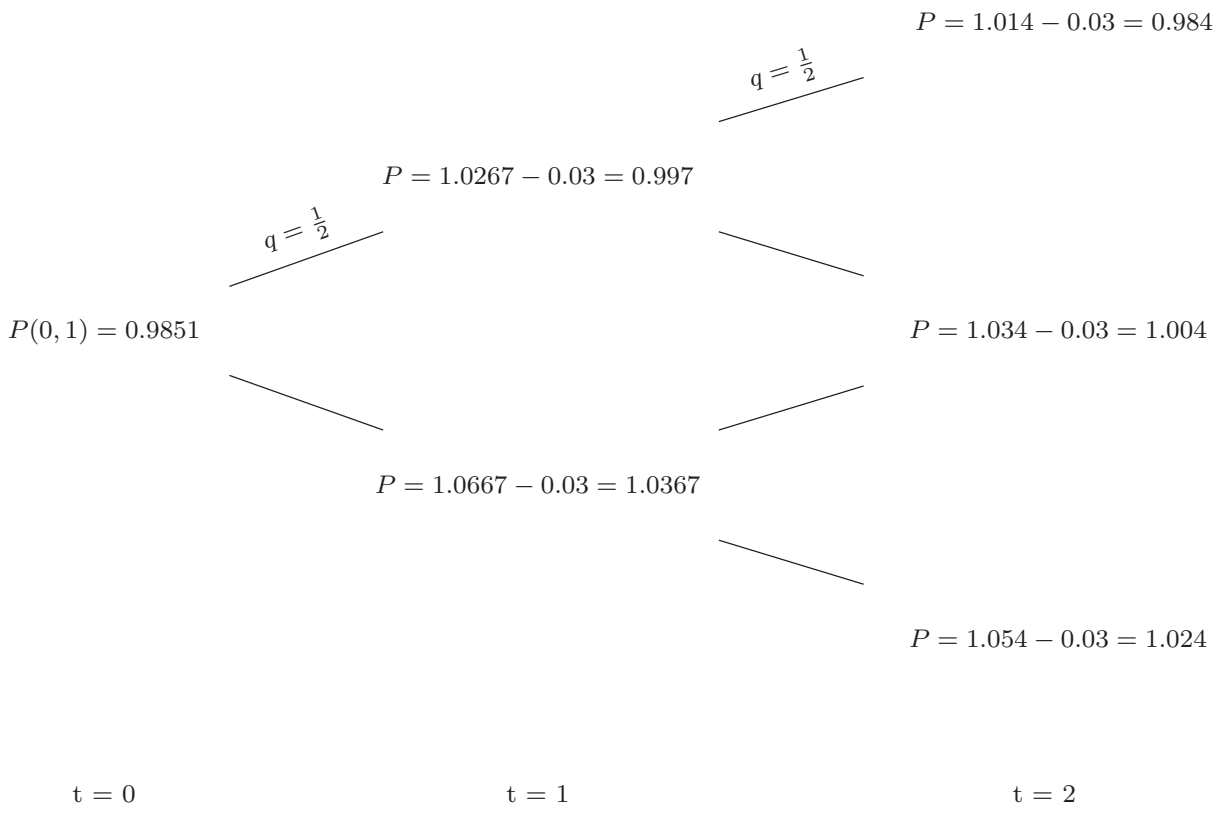
$$P = P(0,1)\left(\frac{1}{2}1.0267 + \frac{1}{2}1.0667\right) = 0.9851\left(\frac{1}{2}1.0267 + \frac{1}{2}1.0667\right) = 1.0311.$$

Naturally, this price could been obtained by simply adding $[P_{\$}(0,1) + P_{\$}(0,2) + P_{\$}(0,3)] \cdot 0.03 + P_{\$}(0,3)$, although the results in the tree above are going to matter while pricing derivatives written on the coupon bearing bond.

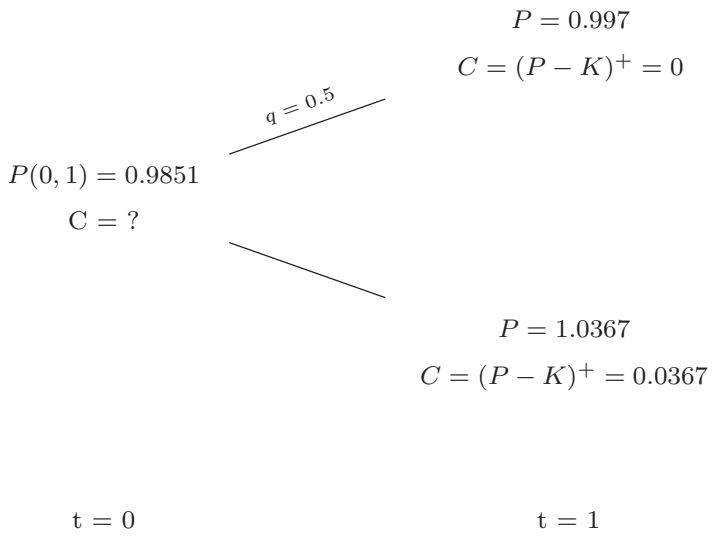
12.5.6.3 Pricing European options

Next, we wish to find the price of options, say the price of two call options on the 1.5Y bond considered in the previous subsection, when the strike price is £1 and the maturities of the options are 6 months and 1 year. Again, we need to figure out the no-arbitrage movements of

the *ex-coupons* bond price. (This is because if we purchase the bond today, we are not entitled to receive any coupon, today. The flow of coupons we are entitled to receive starts from the next period.) We easily obtain the tree below. We must just subtract the coupon, 0.03, from each cum-coupons price in each node of the tree. Then, we obtain:



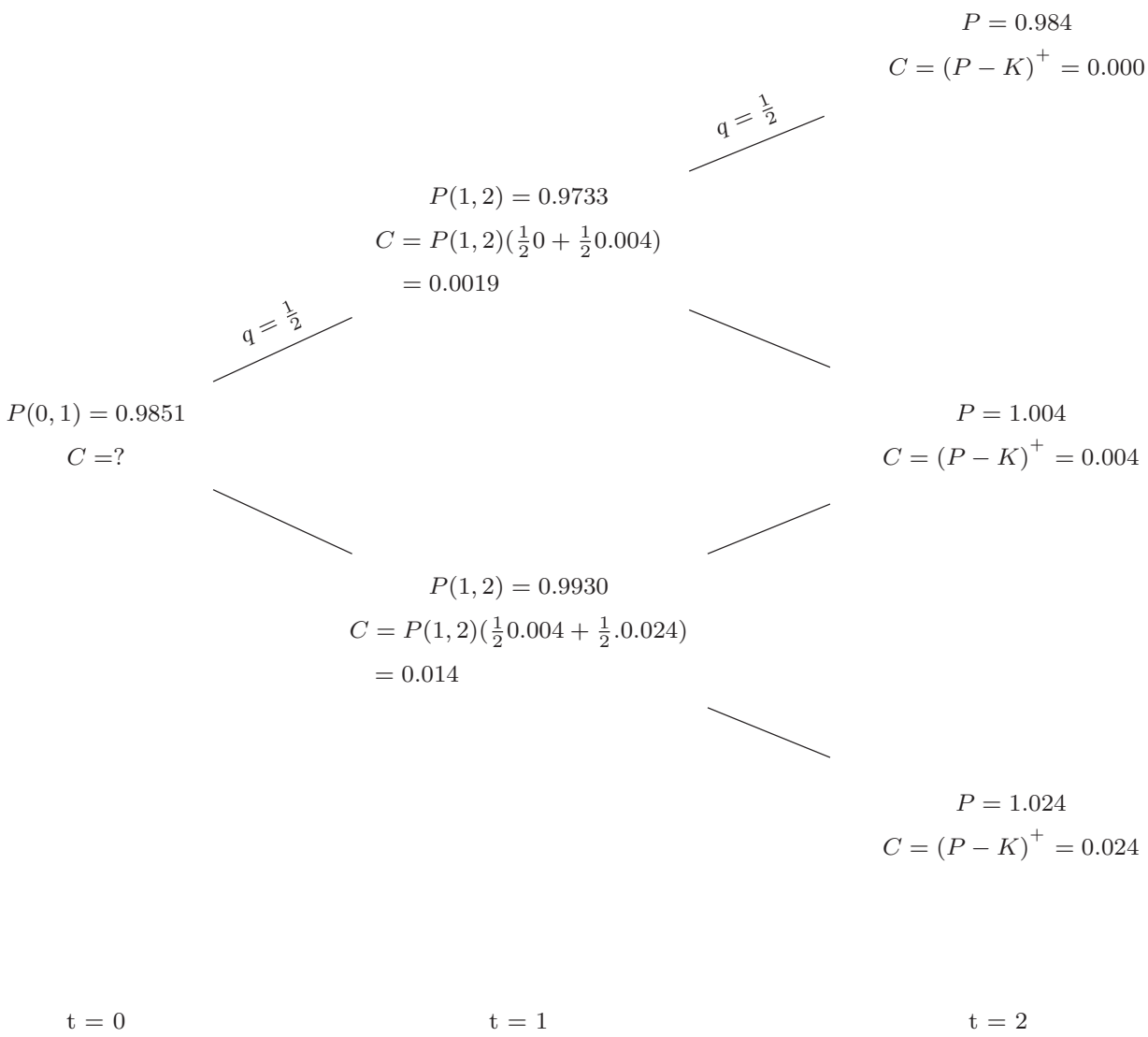
We are ready to price the two options. As for the call option on the 1.5Y bond, with 6 months maturity, and strike price $K = £1$, we have the following tree:



Therefore,

$$C = 0.9851\left(\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0.0367\right) = 1.808\times 10^{-2}.$$

The call option on the 1.5Y bond with 1 year maturity, and strike price $K = £1$, is dealt with similarly. We have the following tree:



Therefore, the price of the option is,

$$C = P(0,1)\left(\frac{1}{2}0.0019 + \frac{1}{2}0.014\right) = 0.9851\left(\frac{1}{2}0.0019 + \frac{1}{2}0.014\right) = 7.831\times 10^{-3}.$$

12.5.7 Continuous-time approximations, with an application to barbell trading

12.5.7.1 The approximation

Consider Eq. (12.39), and define $r = -\Delta t^{-1}\ln P$, and $r_j^{\Delta}(t) \equiv r_j(t)\,\Delta t$, such that:

$$r_j^{\Delta}(t) = \zeta_t - (t(1-q) - j)\ln\delta, \quad \zeta_t \equiv \hat{F}_t(0) + \ln\left(\delta^{-qt}(1-q) + \delta^{(1-q)t}q\right).$$

We have,

$$\mathbb{E}_0\left[r_j^{\Delta}(t)\right]=\zeta_t,\quad \text{and}\quad \sigma^2 t\equiv\mathbb{V}_0\left[r_j^{\Delta}(t)\right]=\left(\ln\delta^{-1}\right)^2 q(1-q)\,t,$$

such that we may define $\ell\equiv\sigma/\sqrt{q(1-q)}$, and, then, $\delta=e^{-\ell}$. Replacing this into the definition of ζ_t, yields, after expanding terms up to the second order,

$$\begin{aligned}\zeta_t&=\hat{F}_t(0)+\ln\left(e^{\ell q t}(1-q)+e^{-\ell(1-q)t}q\right)\\&\approx\hat{F}_t(0)+\ln\left(1+\frac{1}{2}\ell^2 q(1-q)\,t^2\right)\\&\approx\hat{F}_t(0)+\frac{1}{2}\ell^2 q(1-q)\,t^2\\&=\hat{F}_t(0)+\frac{1}{2}\sigma^2 t^2.\end{aligned}$$

Note, this expansion is accurate when ℓt is small, which empirically is indeed, as we have that, typically, $\ell t\approx 10^{-2}t$, which is reasonable small for values of t up to at least 50 years! However, these calculations might also be considered as the starting point for the initial drift of the short-term rate from zero to time t. So, we have, approximately, that,

$$\mathbb{E}_0\left[r_j^{\Delta}(t)\right]=\hat{F}_t(0)+\frac{1}{2}\sigma^2 t^2,\quad \text{and}\quad \mathbb{V}_0\left[r_j^{\Delta}(t)\right]=\sigma^2 t. \tag{12.41}$$

Consistent with the previous calculations, the next chapter shows that in continuous time, the Ho and Lee model predicts the short-term rate to be the solution to:

$$dr_t=\left(\frac{\partial}{\partial t}f_{\$}(0,t)+\sigma^2 t\right)dt+\sigma d\tilde{W}_t, \tag{12.42}$$

where $f_{\$}(0,t)$ is the instantaneous continuously compounded forward rate, and $\tilde{W}_t$ is a Brownian motion defined under the risk-neutral probability (see Section 13.4.2). In fact, the next chapter shows that the instantaneous forward rate predicted by the Ho & Lee model is

$$f(s',t)-f(s,t)=\sigma^2\int_s^{s'}(t-\tau)\,d\tau+\sigma\left(\tilde{W}_{s'}-\tilde{W}_s\right), \tag{12.43}$$

such that, for $r_{s'}=f(s',s')$,

$$r_{s'}-f(s,s')=\sigma^2\int_s^{s'}(s'-\tau)\,d\tau+\sigma\left(\tilde{W}_{s'}-\tilde{W}_s\right), \tag{12.44}$$

the continuous time counterpart to the two conditions in Eqs. (12.41). By combining Eqs. (12.43)-(12.44) leaves

$$f(s',t)-f(s,t)=\sigma^2(s'-s)(t-s')+r_{s'}-f(s,s'). \tag{12.45}$$

As shown in the next chapter (see Section 13.6.1), we have that for any model, including Ho & Lee,[7] the following representation holds true:

$$P(t,T)=\frac{P(0,T)}{P(0,t)}\cdot e^{-\int_t^T[f(t,u)-f(0,u)]du}. \tag{12.46}$$

[7]For example, Eq. (12.36) provides the discrete time counterpart to Eq. (12.46).

Using the expression for $f(s',t) - f(s,t)$ in Eq. (12.45), and integrating,

$$\int_t^T [f(t,u) - f(0,u)]\, du = \frac{1}{2}\sigma^2 (T-t)^2 t + (T-t)(r_t - f(0,t)).$$

Replacing this expression into Eq. (12.46) leaves:

$$P(t,T) = \frac{P(0,T)}{P(0,t)} \cdot \exp\left(-\frac{1}{2}\sigma^2 (T-t)^2 t - (r_t - f(0,t))(T-t)\right). \tag{12.47}$$

It is a neat expression, which we may use, for a variety of purposes, such as option pricing. The next section develops an example relating to barbell trading.

12.5.7.2 Application to barbell trading

We revisit the barbell trading strategy of Section 12.4.3.4, where we argued that this strategy leads to positive profits due to "convexity," as summarized by Figure 12.4. The key point in this argument is that it abstracts from passage of time, and may, in fact, lead us to misinterpret what is a merely static analysis. We may use the Ho and Lee model to analyze the profit and losses of a barbell trade, in a dynamic context free from arbitrage. We consider two situations: one, where the initial yield curve is flat, and a second, where the initial yield curve is upward sloping.

As for the flat yield curve, we use the continuously compounded rate corresponding to the flat 5% of Section 12.4.3.5, delivering $r = \ln 1.05 = 0.04879$. The number of assets to include into the portfolio, θ_1 and θ_2, are as in Eq. (12.16), i.e. $\theta_1 = 0.45706$ and $\theta_3 = 0.56724$. Instantaneous forward rates are $f(0,T) = \lim_{S \downarrow T} F(0,T,S) = -\partial \ln P(0,T)/\partial T = r$. Using Eq. (12.47), with volatility parameter $\sigma = 0.03$, we calculate the value of the strategy a few months later, as follows:

$$\text{Barb}_t = 100 \cdot (\theta_1 P(t,1) + \theta_3 P(t,10) - P(t,5)). \tag{12.48}$$

Figure 12.12 depicts the value of the barbell, Barb_t, for investment horizons equal to 1 month, 3 months, 6 months and one year.

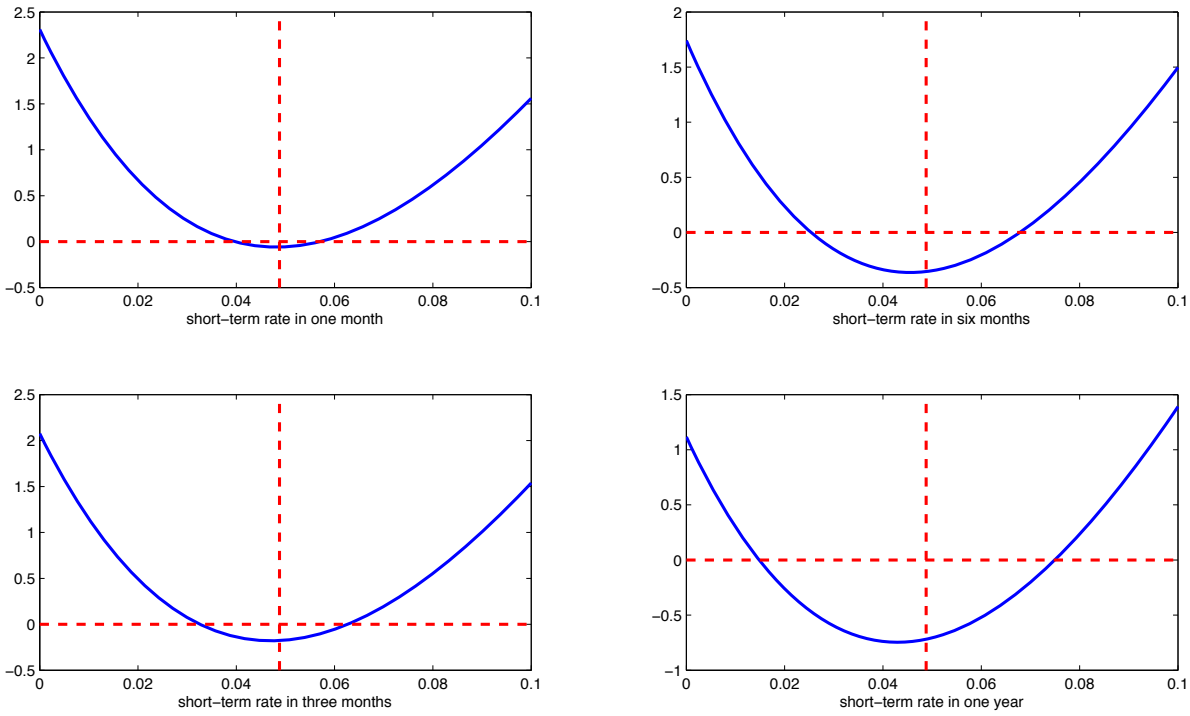


FIGURE 12.12. Profit and losses arising from barbell trading, Barb_t in Eq. (12.48), under the assumption the yield curve is driven by the Ho and Lee model, Eq. (12.47). The initial yield curve is assumed to be flat at $r = 4.8790\%$. Investment horizons are $t = 1/12$ (NW quadrant), $t = 3/12$ (SW quadrant), $t = 6/12$ (NE quadrant) and, $t = 1$ (SE quadrant). The vertical dashed lines pass through $r = 4.8790\%$, and the horizontal dashed lines pass through zero.

This trade is quite risky. For long investment horizons, it pays off when the short-term rate fluctuates significantly away from the initial value, $r = 4.8790\%$. The amount of fluctuations in the short-term rate diminishes as we shrink the investment horizon. Nevertheless, this amount appears to be considerable: for example, at one-month horizon, we should require the short-term rate to move from $r = 4.8790\%$ to either values larger than $r = 6\%$ or lower than $r = 4\%$, in order to claim for positive profits. Actually, these results suggest that a short position in the barbell trade (i.e., sell the barbell portfolio and go long the 5Y bond) should be an interesting strategy to implement in periods where we do not expect high volatility of interest rates. For example, for investment horizons of 6 months, the profits from a short position in the barbell trade are positive within a quite significant range of variation of the short-term rate, [2.5%, 6.8%].

Finally, we consider a scenario where the initial yield curve is upward sloping, and generate prices as $P(0,t) = e^{-tY(t)}$, where $Y(t) \equiv 0.01(1 + \ln t)$. We still determine the value of the portfolio according to Eq. (12.16), i.e., we rely on the self-financing condition in Eq. (12.14) and both (i) the locally riskless condition in Eq. (12.15), $dB_2(\hat{y}_2) = \theta_1 dB_1(\hat{y}_1) + \theta_3 dB_3(\hat{y}_3)$,

and (ii) the (generically incorrect) assumption of parallel shifts in the yield curve, $\frac{\partial B_2(\hat{y}_2)}{\partial \hat{y}_2} = \frac{\partial B_1(\hat{y}_1)}{\partial \hat{y}_1}\theta_1 + \frac{\partial B_3(\hat{y}_3)}{\partial \hat{y}_3}\theta_3$. Figure 12.13 depicts the profit and losses arising from the trade.

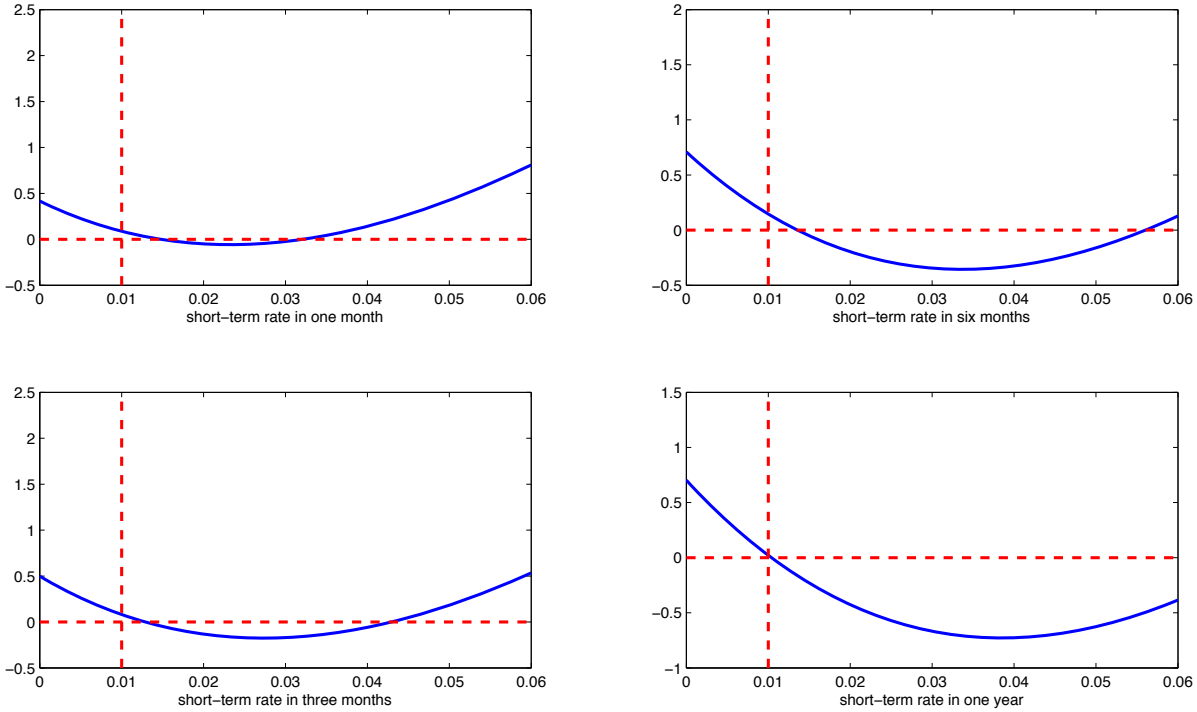


FIGURE 12.13. Profit and losses arising from barbell trading, Barb (t) in Eq. (12.48), under the assumption the yield curve is driven by the Ho and Lee model, Eq. (12.47). The initial yield curve is assumed to be upward sloping, generated by the equation, $Y(t) = 0.01(1+\ln t)$, with prices given by $P(0,t) = e^{-tY(t)}$. Investment horizons are $t = 1/12$ (NW quadrant), $t = 3/12$ (SW quadrant), $t = 6/12$ (NE quadrant) and, $t = 1$ (SE quadrant). The vertical dashed lines pass through the current short-term rate, $r = 1.0\%$, and the horizontal dashed lines pass through zero.

Similarly as for the profit and losses summarized in Figure 12.12, the trade leads to profits only when the short-term rate increases, and significantly, from the initial value $r = 1\%$. In particular, when r moves around 1%, profits increase as r lowers, and decrease, as r goes up. This effect relates to that arising within the static exercise described in Table 12.4: long term bonds benefit from a decrease in r more than short-term, and lose their value more than short-term bonds as r increases. However, as the interest rate increases significantly, the barbell generates profits because the convexity of 10 year bonds dominates overall.

12.6 Beyond Ho and Lee: Calibration through Arrow-Debreu securities

The approach in the previous sections imposes no-arbitrage restrictions on bond prices, which have implications on forward rates, thereby ultimately determining an *implied* stochastic process

of the short-term rate. In this section, we determine the no-arbitrage dynamics of the short-term rate in the first place. The advantage of the approach in this section is that it allows us to implement solutions for models, without requiring to solve them in closed-form. For example, the Ho and Lee (1986) model relies on a number of assumptions, which might be unrealistic, in practice—the short-term rate can take on negative values in this model. It is quite unusual that a model displaying realistic features also has a closed-form solution.

The approach in this section relies on the Arrow-Debreu securities introduced in Chapter 2, and parallels the work of Derman and Kani (1994), Rubinstein (1994) and Dupire (1994) on equity options reviewed in Chapter 11. Arrow-Debreu securities do not exist, in practice, for the reasons put forward in Chapter 2, relating to the simple circumstance we may disagree on models and, hence, on the relevant events and states of nature. However, we can extract the shadow price of these securities from traded securities, and used them to price interest rate derivatives. Naturally, we do so, by relying on a given model. The rationale of all this is that we *do* need models to evaluate interest rate derivatives, and Arrow-Debreu security prices are, in fact, the top of the iceberg of a given model, so to speak. Note that the emphasis in Parts I and II is determining Arrow-Debreu security prices in given economies, by relying on assumptions such as production possibilities, preferences or markets. The approach in this section is to extract the price of Arrow-Debreu securities from that of already traded assets. We illustrate this approach, by elaborating on three points.

First, we show how Arrow-Debreu securities can be used in the specific context of fixed income security evaluation; in particular, we illustrate how to exploit the prices of these assets to turn the martingale restrictions of the previous sections into a set of equivalent conditions that are directly usable for practical purposes.

Second, we just use the previously extracted Arrow-Debreu prices, and develop algorithms to populate the short-term rate tree, while ensuring that the initial yield curve is fitted without errors.

Third, we illustrate this procedure and solve two models: (i) the Ho and Lee model, and (ii) a model developed by Black, Derman and Toy (1990). While we know the solution to the first model, it is useful to review how it could alternatively be solved, as this would naturally give us insights into the mechanisms underlying the calibration algorithms of this section. Note in any event, the retrieving the price of Arrow-Debreu security prices is crucial even in the context of the Ho & Lee model, as they can help determine the price of complex derivatives without a closed-form expression. The second model is one where even the bond price has not a closed-form solution.

12.6.1 Extracting Arrow-Debreu securities from the yield curve

We know, from Chapter 2, that an Arrow-Debreu security is an asset that pays off £1 in some prespecified state of the nature, and zero otherwise. Consider, for example, the diagram in Figure 12.14.

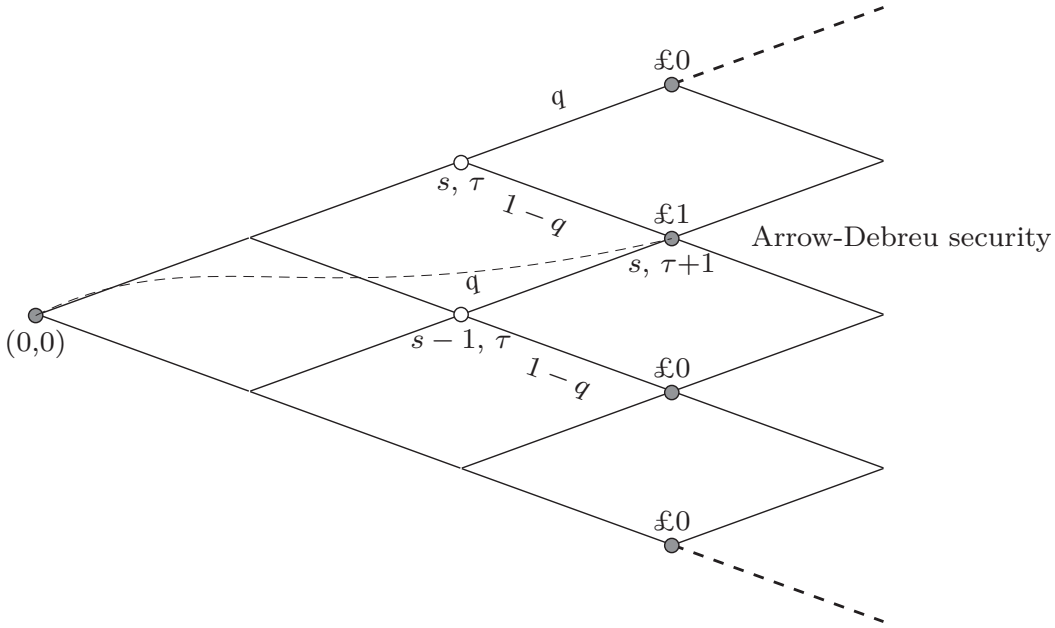


FIGURE 12.14. In the binomial tree of this section, an Arrow-Debreu security for state s at time $\tau+1$ is a security that pays £1 at time $\tau+1$ in state s, and zero otherwise. This section aims to show how to recover Arrow-Debreu prices from the price of fixed income securities.

In this diagram, q is the risk-neutral probability of an upward movement of the short-term rate. A generic pair (s,τ) at each node tracks the number of upward movements of the short-term rate, s, and calendar time, τ, where $s \leq \tau$, as there can only be one possible short-term rate movement in each period. Consider the Arrow-Debreu security for state s at time $\tau+1$. Let $p_s(\tau)$ denote the *current* price of an Arrow-Debreu security, one which pays off £1 at time τ and in state s, and zero otherwise. As we have learnt from many junctures of these lectures, Arrow-Debreu securities can be used to price everything, including bonds. For example, the current market price of a zero coupon bond, maturing at time T is,

$$P_{\$}(0,T) = \sum_{s=0}^{T} p_s(T). \tag{12.49}$$

More generally, consider a derivative that pays off $D_s(\tau)$ in node (s,τ), meaning a dividend equal to $D_1(\tau)$ in state $s=1$, equal to $D_2(\tau)$ in state $s=2,\cdots$, and equal to $D_\tau(\tau)$ in state $s=\tau$. Consistently with explanations given in Chapter 2, the price of this asset, denoted as $C_{\$}(0,T)$, is given by,

$$C_{\$}(0,T) = \sum_{\tau=1}^{T}\sum_{s=0}^{\tau} p_s(\tau) D_s(\tau). \tag{12.50}$$

12.6.1.1 The forward equation for Arrow-Debreu security prices

Our objective is to make use of the initial yield curve, and retrieve the price of all Arrow-Debreu securities, i.e. $p_s(\tau)$ for all s and τ, where $\tau \in \{1,\cdots,T\}$, from the observation of the

initial term-structure of interest rates. Consider the Arrow-Debreu security that pays £1 in node $(s, \tau+1)$ (see Figure 12.14). Denote with $\pi_{j,\tau}[s, \tau+1]$ its value at time τ, and in state j, $j \leq \tau$. What is this value at time τ in all states? A key observation is that in this tree, the node $(s, \tau+1)$ (the filled circle) can only be "accessed to" through the nodes (s, τ) and $(s-1, \tau)$ occurring at time τ (the two empty circles in Figure 12.14). At time τ then, the value $\pi_{j,\tau}[s, \tau+1]$ is zero in all the nodes (j, τ) except the empty circles (s, τ) and $(s-1, \tau)$. Indeed, if we do not happen to be at one of those nodes denoted with empty circles, we could not reach the node $(s, \tau+1)$ (the filled circle), where the Arrow-Debreu security pays off.

So, we are left with finding the values $\pi_{j,\tau}[s, \tau+1]$ in the nodes corresponding to the empty circles (s, τ) and $(s-1, \tau)$, i.e. $\pi_{s,\tau}[s, \tau+1]$ and $\pi_{s-1,\tau}[s, \tau+1]$. Let $r_s(\tau)$ be the continuously compounded short-term rate in node (s, τ). Consider the upper node (s, τ). We have,

$$\pi_{s,\tau}[s, \tau+1] = e^{-r_s(\tau)}[0 \cdot q + 1 \cdot (1-q)] = e^{-r_s(\tau)}(1-q).$$

Similarly, in the lower node, $(s-1, \tau)$,

$$\pi_{s-1,\tau}[s, \tau+1] = e^{-r_{s-1}(\tau)}[1 \cdot q + 0 \cdot (1-q)] = e^{-r_{s-1}(\tau)}q.$$

We can think of our Arrow-Debreu security for $(s, \tau+1)$ as a derivative that at time τ, delivers the following "payoffs"

$$\begin{cases} \pi_{s,\tau}[s, \tau+1] = e^{-r_s(\tau)}(1-q) \\ \pi_{s-1,\tau}[s, \tau+1] = e^{-r_{s-1}(\tau)}q \\ \pi_{j,\tau}[s, \tau+1] = 0, \quad \text{for all } j < s \end{cases} \tag{12.51}$$

These "payoffs" are simply the market value of the Arrow-Debreu security for $(s, \tau+1)$, in the various states occurring at time τ, i.e. the money the holder can make by selling the asset at time τ, in the various states. Therefore, we can apply Eq. (12.50), and obtain,

$$\begin{aligned} p_s(\tau+1) &= \sum_{j=0}^{\tau} p_j(\tau)\pi_{j,\tau}[s, \tau+1] \\ &= p_s(\tau)\pi_{s,\tau}[s, \tau+1] + p_{s-1}(\tau)\pi_{s-1,\tau}[s, \tau+1]. \end{aligned}$$

By replacing the Arrow-Debreu prices in (12.51) into the previous equation, we obtain the so-called *forward equation* for the Arrow-Debreu prices,

$$p_s(\tau+1) = p_s(\tau)e^{-r_s(\tau)}(1-q) + p_{s-1}(\tau)e^{-r_{s-1}(\tau)}q. \tag{12.52}$$

Eq. (12.52) is the counterpart to the forward equation used to fit European equity option prices (see Chapter 11). The approach in this section is distinct from that in Chapter 11: now, we take the risk-neutral probability constant and interest rates time-varying and state-dependent; in contrast, in Chapter 11, interest rates are exogenously given but the risk-neutral probability is time-varying and state-dependent. Naturally, the approach in this section could be generalized to include stochastic risk-neutral probabilities: generalizing this approach may be motivated by the need to fit European options in addition to bond prices. In the next section, we explain how to fit bond prices.

12.6.1.2 The bond price equation

To implement an algorithm, we make a repeated use of the forward equation (12.52) and the following zero pricing equation,

$$P_{\$}(0,\tau+1)=\sum_{s=0}^{\tau}p_{s}(\tau)\,e^{-r_{s}(\tau)}. \tag{12.53}$$

The inputs to the algorithm are a number of zeros equal to the largest maturity date the tree extends to. Note an important feature of the calibration procedure. While we extract Arrow-Debreu security prices, we need to make reference to a given model for the underlying short-term rate movements, $r_s(\tau)$ and, indeed, we shall illustrate the algorithm by hinging upon two examples in the following sections. Instead, in Chapter 11, we have illustrated that in the equity case, cross section of option prices is enough to uniquely pin down the underlying stock price movements.

12.6.2 Two model examples

We begin with Ho and Lee, assuming continuous compounding. By Eq. (12.39), the short-term rate predicted by the Ho and Lee model is:

$$r_{j}(\tau)=\hat{F}_{\tau}(0)+\ln\left((1-q)+q\delta^{\tau}\right)-(\tau-j)\ln\delta, \tag{12.54}$$

where $\hat{F}_{\tau}(0)$ is the continuously compounded forward rate at time zero for maturity $[\tau,\tau+1]$, and j is the number of *upward* movements of the entire set of *bond prices*. Define $s\equiv(\tau-j)$, which is the number of downward movements of the bond prices or, equivalently, the number of *upward* movements of the *short-term rate*. Hence, we can equivalently index the short-term rate by s, instead of j, and rewrite Eq. (12.54) with a slight abuse in notation, as follows:

$$r_{s}(\tau)=\underbrace{\hat{F}_{\tau}(0)+\ln\left((1-q)+q\delta^{\tau}\right)}_{\equiv\, r_{0}(\tau)}+\ln\delta^{-1}\cdot s, \tag{12.55}$$

such that $r_0(\tau)$ is the short-term rate at time τ, in the event of zero upward movements in this rate, and δ is the usual volatility parameter, which can be calibrated through $\ln\delta^{-1}=\frac{\text{Std}(\Delta r)}{\sqrt{q(1-q)}}$, with straightforward notation. Note, incidentally, that the short-term rate movements do depend on the specific value we assign to the risk-neutral probability q.

Naturally, Eq. (12.55) would be sufficient to populate the interest rate tree quite easily. However, even in the context of Ho and Lee, it is still quite important to retrieve Arrow-Debreu security prices, as these would allow us to price all kinds of interest rate derivatives, including those for which a closed-form expression is unavailable.

Let us proceed. At time zero, the price of a zero maturing at time $\tau+1$ is:

$$P_{\$}(0,\tau+1)=\sum_{s=0}^{\tau}p_{s}(\tau)\,e^{-r_{s}(\tau)}=e^{-r_{0}(\tau)}\sum_{s=0}^{\tau}\delta^{s}p_{s}(\tau),$$

where the second equality follows by the assumption that the short-term rate is solution to Eq. (12.55).

By rearranging terms in the previous equation, we obtain a closed-form expression for the future short-term rate at time τ, in the event of zero upward movements,

$$r_0(\tau) = \ln\left(\frac{\sum_{s=0}^{\tau}\delta^s p_s(\tau)}{P_{\$}(0,\tau+1)}\right). \tag{12.56}$$

This is the counterpart to zero pricing equation (12.53).

We use Eq. (12.56) and the forward equation (12.52) to populate the interest rate tree, under the assumption that $q = \frac{1}{2}$. Precisely, the algorithm proceeds as follows:

(i) Given the boundary condition for the Arrow-Debreu price, $p_0(0) = 1$, determine the initial value of the short-term rate, $r_0(0)$, using Eq. (12.56), as $r_0(0) = \ln(1/P_{\$}(0,1))$.

(ii) Suppose we know the *future* value of the short-term rate at time $\tau - 1$, in the event of no upward movements, i.e. $r_0(\tau-1)$. Then, given the value of $r_0(\tau-1)$, and the price of the Arrow-Debreu securities $p_s(\tau-1)$ for $s \leq \tau - 1$, determine $p_s(\tau)$ for $s \leq \tau$, through the forward equation (12.52),

$$p_s(\tau) = p_s(\tau-1)\,\delta^s e^{-r_0(\tau-1)}(1-q) + p_{s-1}(\tau-1)\,\delta^{s-1}e^{-r_0(\tau-1)}q, \quad q = \frac{1}{2},$$

where the last equation follows by plugging Eq. (12.55) into Eq. (12.52).

(iii) Given the Arrow-Debreu prices $p_s(\tau)$ for $s \leq \tau$, use Eq. (12.56) to determine the *future* value of the short-term rate at time τ, in the event of no upward movements, i.e. $r_0(\tau)$.

(iv) If $\tau = T$, stop. Otherwise, go to (ii).

As a second example, consider the Black, Derman and Toy (1990) model. In this model, the short-term rate is solution to,

$$r_s(\tau) = \delta^s r_0(\tau), \tag{12.57}$$

where δ is, once again, a volatility parameter.[8] For computational convenience, this model assumes that the short-term rate in Eq. (12.57) is discretely compounded. Accordingly, we rewrite the forward equation (12.52) in terms of discretely compounded rates,

$$p_s(\tau+1) = p_s(\tau)\frac{1}{1+r_s(\tau)}(1-q) + p_{s-1}(\tau)\frac{1}{1+r_{s-1}(\tau)}q. \tag{12.58}$$

The algorithm proceeds as follows:

(i) Determine the initial value of the short-term rate, $r_0(0)$, as the solution to, $P_{\$}(0,1) = \frac{1}{1+r_0(0)}$.

(ii) Suppose we know the *future* value of the short-term rate at time $\tau - 1$, in the event of no upward movements, i.e. $r_0(\tau-1)$. Then, given the value of $r_0(\tau-1)$, and the price of the Arrow-Debreu securities $p_s(\tau-1)$ for $s \leq \tau - 1$, determine $p_s(\tau)$ for $s \leq \tau$, through the forward equation (12.58),

$$p_s(\tau) = p_s(\tau-1)\frac{1}{1+\delta^s r_0(\tau-1)}(1-q) + p_{s-1}(\tau-1)\frac{1}{1+\delta^{s-1}r_0(\tau-1)}q, \quad q = \frac{1}{2},$$

where the last equation follows by plugging Eq. (12.57) into Eq. (12.58).

[8] In its most general form, this model assumes that $r_s(\tau) = \delta_\tau^s r_0(\tau)$, where δ_τ is a volatility parameter that varies deterministically over time. This more general formulation leads to more flexibility, which is useful to fit the term structure of volatility.

(iii) Given the boundary condition $p_0(0) = 1$, and the Arrow-Debreu prices, $p_s(\tau)$ for $s \leq \tau$, use the pricing equation for the zero,

$$P_{\$}(0, \tau + 1) = \sum_{s=0}^{\tau} p_s(\tau) \frac{1}{1 + \delta^s r_0(\tau)},$$

to solve, numerically, for the *future* value of the short-term rate at time τ, in the event of no upward movements, i.e. $r_0(\tau)$. Note, we did not need this additional step for the solution of the Ho and Lee model, as the short-term rate $r_0(\tau)$ is known in closed form in the Ho and Lee model (see Eq. (12.56)). Note, since $r_0(\tau) > 0$, then, we also have that $r_s(\tau) > 0$ by Eq. (12.57).

(iv) If $\tau = T$, stop. Otherwise, go to (ii).

12.6.2.1 A numerical example

Consider, again the Ho and Lee model example in Section 12.5.5, where three zeros were traded: (i) one zero maturing in 6 months, (ii) one zero maturing in 1 year, and (iii) one zero maturing in 1.5 years, with market prices $P_{\$}(0,1) = 0.9851$, $P_{\$}(0,2) = 0.9685$, $P_{\$}(0,3) = 0.9445$. By Eq. (12.55), the Ho and Lee model assumes that,

$$r_s(\tau) = r_0(\tau) + \left(\ln \delta^{-1}\right) \cdot s. \tag{12.59}$$

We use Eq. (12.59) and find the values of the short-term rate $r_s(\tau)$ in each node, under the assumption that $q = \frac{1}{2}$, and that the standard deviation of the short-term rate is 0.014, annualized. To find δ, we may use the relation, $\ln \delta^{-1} = \frac{\text{Std}(\Delta r)}{\sqrt{q(1-q)}}$, where $q = \frac{1}{2}$ and $\text{Std}(\Delta r)$ is the standard deviation of the short-term rate, which equals $\text{Std}(\Delta r) = 0.014$, annualized. Therefore $\ln \delta^{-1} = \frac{0.014}{\sqrt{2}} / \frac{1}{2} = 0.02$ or $\delta = 0.9802$.

For the Ho & Lee model, we know the closed-form expression for $r_0(\tau)$,

$$r_0(\tau) = \ln\left(\frac{\sum_{s=0}^{\tau} \delta^s p_s(\tau)}{P_{\$}(0, \tau + 1)}\right), \tag{12.60}$$

where $p_s(\tau)$ denotes the price of an Arrow-Debreu security which pays of £1 in state s at time τ, and zero otherwise. Given the term-structure of prices $P_{\$}(0, \tau + 1)$, $\tau = 0, 1, 2$, we "populate" the tree using Eq. (12.60) and the forward Arrow-Debreu prices equation (12.52),

$$p_s(\tau) = \frac{1}{2} e^{-r_0(\tau - 1)} \left[\delta^s p_s(\tau - 1) + \delta^{s-1} p_{s-1}(\tau - 1)\right], \tag{12.61}$$

with the appropriate boundary conditions.

So we have to calculate interest rates and Arrow-Debreu prices for $\tau = 0, 1, 2$.

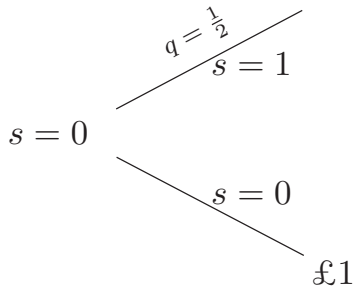
- $\tau = 0$. Eq. (12.60) is trivial. It leads to, $r_0(0) = \ln\left(\frac{1}{P_{\$}(0,1)}\right) = 0.015$.The forward equation for the Arrow-Debreu prices, Eq. (12.61), is also trivial, $p_0(0) = 1$.

- $\tau = 1$. Let us use Eq. (12.61), the forward equation for the Arrow-Debreu prices, to find $p_0(1)$ and $p_1(1)$. We have two cases:

– $s = 0$. We have:

$$p_0(1) = \frac{1}{2} e^{-r_0(0)} [p_0(0) + 0] = \frac{1}{2} e^{-r_0(0)} = 0.4925.$$

The previous relation holds because $p_0(1)$ is the current price of the Arrow-Debreu security which pays off £1 in state 0 at time 1, as illustrated by the tree in the figure below:



– $s = 1$. By a similar reasoning,

$$p_1(1) = \frac{1}{2} e^{-r_0(0)} [0 + p_0(0)] = \frac{1}{2} e^{-r_0(0)} = 0.4925.$$

This Arrow-Debreu security is valued the same as that for state $s = 0$ because the risk-neutral probability is 50%.

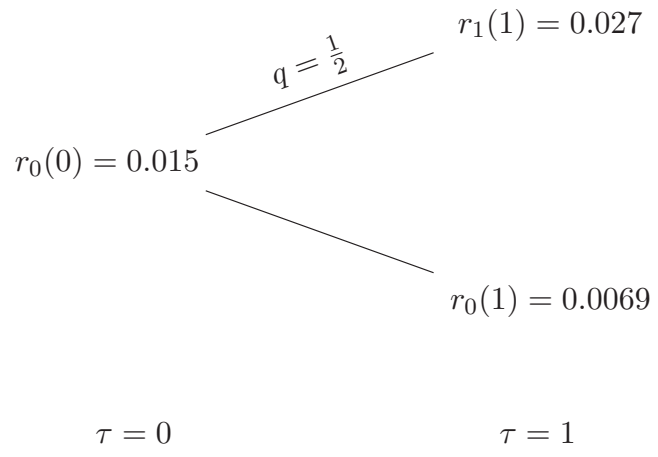
Eq. (12.60) is, now,

$$r_0(1) = \ln\left(\frac{p_0(1) + \delta p_1(1)}{P_{\$}(0,2)}\right) = \ln\left(\frac{0.4925 \cdot (1 + 0.9802)}{0.9685}\right) = 0.0069.$$

Hence, by Eq. (12.59),

$$r_1(1) = r_0(1) + \left(\ln \delta^{-1}\right) = 0.0069 + 0.02 = 0.0270.$$

To sum up, we have the tree below,

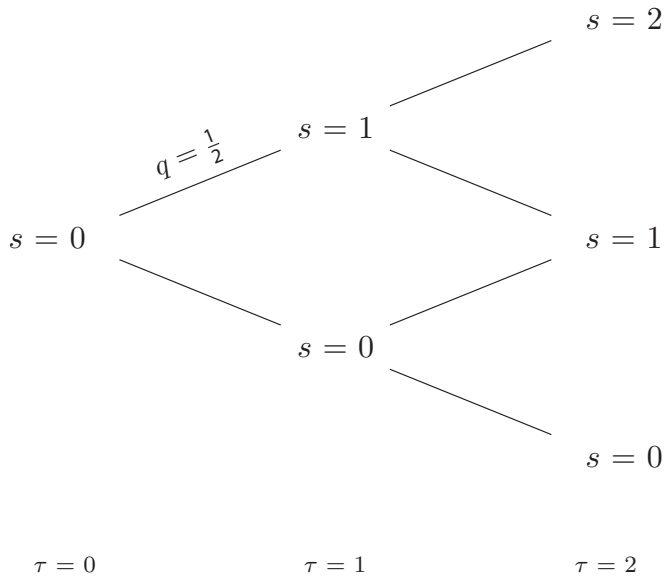


We can now calculate the values of the short-term rate for one further period.

- $\tau = 2$. By Eq. (12.61), the forward equation for the Arrow-Debreu prices, we have the following three cases:

$$\begin{array}{ll} (s=0) & p_0(2) = \frac{1}{2}e^{-r_0(1)}\left[p_0(1) + 0\right] = 0.2446 \\ (s=1) & p_1(2) = \frac{1}{2}e^{-r_0(1)}\left[\delta p_1(1) + p_0(1)\right] = 0.4843 \\ (s=2) & p_2(2) = \frac{1}{2}e^{-r_0(1)}\left[0 + \delta p_1(1)\right] = 0.2397 \end{array}$$

  The tree below further illustrates how to obtain these prices.



  Consider, for example, $p_0(2)$. It is the price of the Arrow-Debreu security for time 2, under two consecutive downward movements of the short-term rate. This state can only be accessed to through the state $s = 0$ at time $\tau = 1$. But at state $s = 0$ at time $\tau = 1$, the value of the Arrow-Debreu asset is $\frac{1}{2}e^{-r_0(1)}$. Hence, $p_0(2) = p_0(1) \cdot \frac{1}{2}e^{-r_0(1)}$. By a similar reasoning, we have that $p_2(2) = p_1(1) \cdot \frac{1}{2}e^{-r_1(1)} = p_1(1) \cdot \frac{1}{2}e^{-r_0(1)}\delta$. Note, there is some symmetry in the distribution of the Arrow-Debreu prices, with $p_1(2)$ being the largest, being the price of the security that pays off with the highest likelihood. However,

$p_0(2) > p_2(2)$, even if q is constant and equal to 50%, because discounting is more severe whilst crossing the nodes leading to $s = 2$, compared to the nodes leading to $s = 0$.

We can now calculate the values of the short-term rate for each node. Eq. (12.60) is, now,

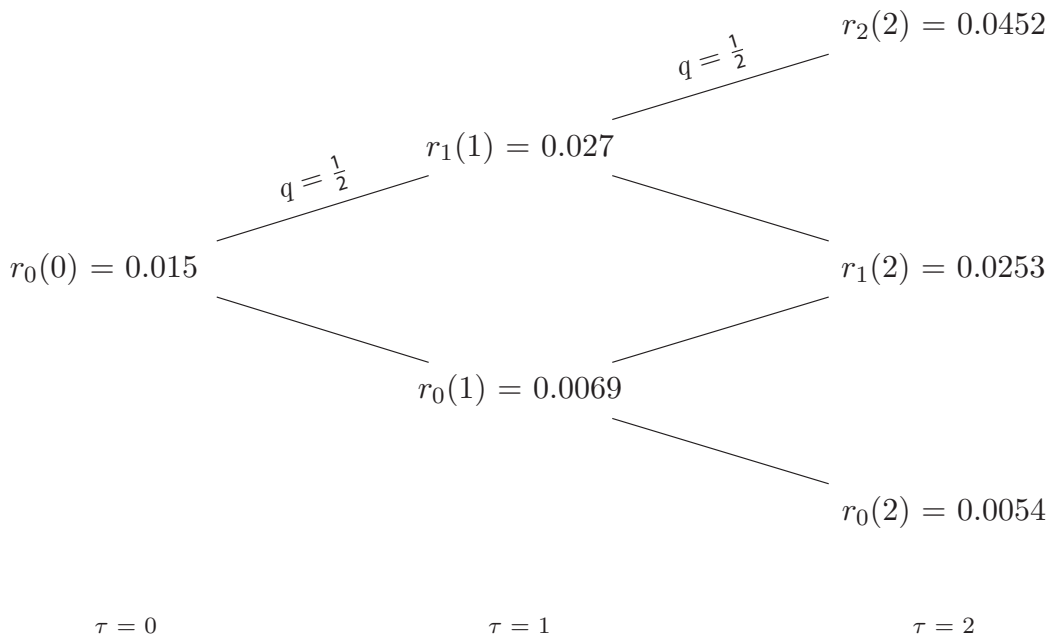
$$\begin{aligned} r_0(2) &= \ln\left(\frac{p_0(2) + \delta p_1(2) + \delta^2 p_2(2)}{P_{\$}(0,3)}\right) \\ &= \ln\left(\frac{0.2446 + 0.9802 \cdot 0.4843 + (0.9802)^2 \cdot 0.2397}{0.9445}\right) = 0.0054. \end{aligned}$$

Hence, by Eq. (12.59),

$$r_s(2) = r_0(2) + \left(\ln \delta^{-1}\right) \cdot s = 0.0054 + 0.02 \cdot s, \qquad s = 0, 1, 2.$$

This yields the following values for the short-term rate: $r_0(2) = 0.0054$, $r_1(2) = 0.0253$, and $r_2(2) = 0.0452$.

The diagram below summarizes the implied tree for the short-term rate in this model.



Naturally, the prices $P = e^{-r}$ in the nodes of the previous tree match those calculated in Section 12.6.5, apart from discrepancies arising due to rounding errors.

12.6.2.2 A second example: time-varying probabilities and interest rate volatility swaps

Assume that the spot yield curve is 2.5% for $t = 1$ year, 4.5% for $t = 2$ years, and 6% for $t = 3$ years, continuously compounded and annualized. Consider the following model:

$$r_s(t) = r_0(t) + a \cdot s, \tag{12.62}$$

where a is a constant and equal to 0.01, $r_s(t)$ is the continuously compounded short-term rate as of time t, after s upward movements and, finally, the unit period of time is taken to be one

year. As we know, the Ho & Lee model predicts that the price as of time zero of an Arrow-Debreu security paying off in state s at time t, denoted as $p_s(t)$, satisfies the following forward equation, for $t > 1$, $s \geq 0$ and $s \leq t$:

$$p_s(t) = e^{-r_0(t-1)} \left[(1-q) \left(e^{-a}\right)^s p_s(t-1) + q \left(e^{-a}\right)^{s-1} p_{s-1}(t-1) \right], \tag{12.63}$$

where q is the risk-neutral probability of an upward movement in the short-term rate. Furthermore, according to this model, the price of a zero coupon bond, paying £1 at time t, $P_{\$}(0,t)$, equals,

$$P_{\$}(0,t) = e^{-r_0(t-1)} \sum_{s=0}^{t-1} \left(e^{-a}\right)^s p_s(t-1). \tag{12.64}$$

Suppose, next, that the risk-neutral probability of an upward movement at any time t is not a constant q, but a function of calendar time, say q_t: q_t is, then, the probability of an upward movement in the short-term rate from time t to time $t+1$. Naturally, the assumption that q_t is time-varying, makes this model markedly distinct from Ho & Lee model. To calibrate this model, we consider the recursive equation for the Arrow-Debreu security prices:

$$p_s(t) = e^{-r_0(t-1)} \left[(1-q_{t-1}) \left(e^{-a}\right)^s p_s(t-1) + q_{t-1} \left(e^{-a}\right)^{s-1} p_{s-1}(t-1) \right], \tag{12.65}$$

where q_{t-1} denotes the risk-neutral probability of an upward movement in the short-term rate from time $t-1$ to time t. The boundary conditions are the usual ones: $p_0(0) = 1$, $p_s(t) = 0$, for $s \geq t$ and $s < 0$. Eq. (12.65) can be derived through the same arguments in Section 12.7.1.

Next, suppose the risk-neutral probability of an upward movement in the short-term rate in the first period equals $\frac{1}{2}$. Suppose, further, that available for trading is a derivative, which pays off an amount of £1 in state $s = 2$ and an amount of £1 in state $s = 0$, both at time $t = 2$. The current price of this derivative equals 0.45514. The interpretation of the derivative is that of a contract that pays off when the interest rate experiences extreme movements (up-up or down-down)—a very basic *interest rate volatility contract*. Its price can be expressed as the sum of the two Arrow-Debreu securities for these extreme interest rate movements. Let us set the nominal values of the zero coupon bonds to £1. To populate the interest rate tree, we need to determine the three zero prices, which are:

$$P_{\$}(0,1) = e^{-0.025} = 0.97531, \; P_{\$}(0,2) = e^{-0.045\times 2} = 0.91393, \; P_{\$}(0,3) = e^{-0.06\times 3} = 0.83527.$$

We can start populate the tree. Eq. (12.64) can be rewritten as:

$$r_0(t) = \ln\left(\frac{\sum_{s=0}^{t} \left(e^{-a}\right)^s p_s(t)}{P_{\$}(0,t+1)} \right). \tag{12.66}$$

We have,

- $t = 0$. In this case, Eq. (12.66) is:

$$r_0(0) = \ln\left(\frac{1}{0.97531}\right) = 0.025.$$

- $t = 1$. We have two nodes to fill: $s = 0$ & $s = 1$. We use Eq. (12.65), as follows:

– $s = 0$: We have,

$$p_0(1) = e^{-r_0(0)}(1 - q_0)\, p_0(0) = 0.97531 \times 0.5 \times 1 = 0.48766.$$

– $s = 1$: We have,

$$p_1(1) = e^{-r_0(0)} q_0 p_0(0) = 0.97531 \times 0.5 \times 1 = 0.48766.$$

Then, Eq. (12.66) is,

$$\begin{aligned} r_0(1) &= \ln\left(\frac{p_0(1) + (e^{-a})\, p_1(1)}{P_{\$}(0,2)}\right) = \ln\left(\frac{0.48766 \times (1 + e^{-0.01})}{0.91393}\right) = 0.06 \\ r_1(1) &= r_0(1) + 0.01 = 0.07. \end{aligned}$$

- $t = 2$. There are now three nodes to fill, corresponding to $s = 0$, $s = 1$ and $s = 2$. We use Eq. (12.65), as follows:

  – $s = 0$: We have,

  $$p_0(2) = e^{-r_0(1)}(1 - q_1)\, p_0(1) = e^{-0.06}(1 - q_1)\, 0.48766.$$

  – $s = 1$: We have,

  $$p_1(2) = e^{-r_0(1)}\left[(1 - q_1)\, e^{-a} p_1(1) + q_1 p_0(1)\right] = e^{-0.06}\left[(1 - q_1)\, e^{-0.01} + q_1\right] 0.48766.$$

  – $s = 2$: We have,

  $$p_2(2) = e^{-r_0(1)} q_1 e^{-a} p_1(1) = e^{-0.06} q_1 e^{-0.01} 0.48766.$$

  We do not know yet q_1. Yet the "rate volatility" asset, which quotes for 0.45514, can be used to extract q_1. At time $t = 1$, its price is either $Z_u = e^{-0.07} q_1$ (in the up state of the world), or $Z_d = e^{-0.06}(1 - q_1)$ (in the down state of the world). So by no-arbitrage, its current price, satisfies

  $$0.45514 = \frac{1}{2} e^{-0.025}(Z_u + Z_d) = \frac{1}{2} e^{-0.025}\left[e^{-0.07} q_1 + e^{-0.06}(1 - q_1)\right].$$

  Solving for q_1 yields, $q_1 = 0.90$. Naturally, the same result is obtained by calibrating q_1 so as to make the price of the derivative, 0.45514, match the sum of the prices of the Arrow-Debreu securities paying off in states 0 and 2 at $t = 2$, viz $q_1 : 0.45514 = p_0(2) + p_2(2) = e^{-0.06}(1 - q_1)\, 0.48766 + e^{-0.06} q_1 e^{-0.01} 0.48766$. So now, we can use $q_1 = 90\%$ and calculate the Arrow-Debreu prices, obtaining:

  $$\begin{aligned} p_0(2) &= e^{-0.06}(1 - .9)\, 0.48766 = 0.04592 \\ p_1(2) &= e^{-0.06}\left[(1 - .9)\, e^{-0.01} + .9\right] 0.48766 = 0.4588 \\ p_2(2) &= e^{-0.06}.9 e^{-0.01} 0.48766 = 0.40922 \end{aligned}$$

  Note, there is no symmetry at all in the distribution of these Arrow-Debreu security prices. The price $p_0(2)$ is very low, due to the fact that q_1 is very high, such that the probability of reaching the lowest node of the tree at time $t = 2$ is quite low.

Next, by Eq. (12.66),

$$\begin{aligned} r_0(2) &= \ln\left(\frac{p_0(2)+e^{-a}p_1(2)+e^{-2a}p_2(2)}{P_\$(0,3)}\right) \\ &= \ln\left(\frac{0.04592+e^{-0.01}0.4588+e^{-2\times 0.01}0.40922}{0.83527}\right)=0.07605, \end{aligned}$$

and,

$$\begin{aligned} r_1(2) &= r_0(2)+0.01=0.07605+0.01=0.08605 \\ r_2(2) &= r_0(2)+2\times 0.01=0.07605+2\times 0.01=0.09605 \end{aligned}$$

Finally, we wish to evaluate a European call option at time zero, written on the three year zero coupon bond with nominal value equal to £1. This option expires at $t=2$ and has a strike price equal to £0.91000. At expiry, the option pays off:

$$\begin{aligned} C_2(2) &\equiv \left(e^{-r_2(2)}-0.91\right)^+ = \left(e^{-0.09605}-0.91\right)^+ = 0 \\ C_1(2) &\equiv \left(e^{-r_1(2)}-0.91\right)^+ = \left(e^{-0.08605}-0.91\right)^+ = 0.00755 \\ C_0(2) &\equiv \left(e^{-r_0(2)}-0.91\right)^+ = \left(e^{-0.07605}-0.91\right)^+ = 0.01677 \end{aligned}$$

Then,

$$\begin{aligned} C_u &= e^{-r_1(1)}\left(q_1C_2(2)+(1-q_1)C_1(2)\right) \\ &= e^{-0.07}\left(0.9\times 0+0.1\times 0.00755\right)=7.0396\times 10^{-4} \\ C_d &= e^{-r_0(1)}\left(q_1C_1(2)+(1-q_1)C_0(2)\right) \\ &= e^{-0.06}\left(0.9\times 0.00755+0.1\times 0.01677\right)=7.9786\times 10^{-3} \end{aligned}$$

which leads to $C=e^{-0.025}\frac{1}{2}\left(C_u+C_d\right)=4.2341\times 10^{-3}$.

Finally, using all the market data so far, we wish to evaluate a second European call option written on the three year bond, expiring in one year, and struck at £0.85. Its no-arbitrage price, denoted with C_T, is:

$$C_T = e^{-0.025}\frac{1}{2}\left(\left(P_u(1,3)-0.85\right)^+ + \left(P_d(1,3)-0.85\right)^+\right),$$

where

$$\begin{aligned} P_u(1,3) &= e^{-r_1(1)}\left(q_1e^{-r_2(2)}+(1-q_1)e^{-r_1(2)}\right) \\ &= e^{-0.07}\left(0.90\times e^{-0.09605}+0.10\times e^{-0.08605}\right)=0.84786 \\ P_d(1,3) &= e^{-r_0(1)}\left(q_1e^{-r_1(2)}+(1-q_1)e^{-r_0(2)}\right) \\ &= e^{-0.06}\left(0.90\times e^{-0.08605}+0.10\times e^{-0.07605}\right)=0.86498 \end{aligned}$$

That is, $C_T=e^{-0.025}\frac{1}{2}\left(0.01498\right)=0.00730$. Suppose, now, that the market value of this option diverges from C_T, i.e. $C_T\neq C_\$$, where $C_\$$ is the market value of the option. For example, $C_T<C_\$$. To implement this arbitrage opportunity, we can sell the option, and use the proceeds to build up a portfolio comprising the bond expiring in three years and a money market account, with initial value:

$$V_0=\Delta P_\$(0,3)+M,$$

where Δ and M are chosen to match the payoffs promised by the option at time 1:

$$(\Delta, M): \begin{cases} \Delta P_u(1,3) + Me^{r_0} = \pi_u \\ \Delta P_d(1,3) + Me^{r_0} = \pi_d \end{cases}$$

where π_u and π_d are the payoffs of the one year option. The solution is,

$$\hat{\Delta} = \frac{\pi_u - \pi_d}{P_u(1,3) - P_d(1,3)}, \quad \hat{M} = e^{-r_0} \frac{\pi_d P_u(1,3) - \pi_u P_d(1,3)}{P_u(1,3) - P_d(1,3)}.$$

Using the numerical values obtained so far, $\pi_u = 0$, $\pi_d = 0.01498$, $P_u(1,3) = 0.84786$, $P_d(1,3) = 0.86498$, we have:

$$\hat{\Delta} = \frac{-0.01498}{0.84786 - 0.86498} = 0.875, \quad \hat{M} = e^{-0.025} \frac{0.01498 \times 0.84786}{0.84786 - 0.86498} = -0.72356.$$

The current value of this portfolio is,

$$\hat{V}_0 = \hat{\Delta} P_{\$}(0,3) + \hat{M} = 0.875 \times 0.83527 - 0.72356 = 0.00730 = C_T.$$

Finally, if $C_T > C_{\$}$, we can buy the option, claim for $[\pi_u, \pi_d]$ from the option seller, and sale short the previous $(\hat{\Delta}, \hat{M})$ portfolio. Net profits from the trade are $C_T - C_{\$} > 0$ at time $t = 0$; at time $t = 1$, we use the claims from the option we purchased to honour the sale of the $(\hat{\Delta}, \hat{M})$-portfolio.

12.7 Callables, puttable and convertibles with trees

This section provides an introductory discussion of the pricing of callable, puttable and convertible bonds, with and without credit risk, and develops basic pricing examples for callable and convertibles, relying on binomial trees. Chapter 13 develops a continuous time evaluation framework for callable and puttable bonds, while Chapter 14 contains a continuous time model to evaluate convertible bonds.

12.7.1 Definitions and rationale

Callable bonds are assets that can be called back by the issuer at a pre-specified strike price, either at a fixed maturity date or at any fixed date before the expiration. The rationale behind this optionality is that at the date of issuance, the market might not share the same optimism as the issuer as regards the issuer's future creditworthiness. By adding a call provision, the bond issuer gives itself the option to refinance at some future date, at hopefully better market conditions. Although this specific example might link to agency problems or difference in beliefs between the bond issuer and market participants, the option to call the bond may simply arise due to pure hedging motives—a concern that future interest rates might lower.

Naturally, the right to call the bonds rises the cost of capital, to the extent of the value of this (call) option to redeem the bonds. Let us examine the details of these issues. Let $V_t(T)$ denote the value of a callable zero coupon bond expiring at T, which we assume to embed an American style option to call, i.e. one that could be exercized at any point in time before the expiration. For alleviate notation, set $V_t \equiv V_t(T)$. For each t, the firm decides to reedem the bonds whenever the exercise price K is less than the (discounted value of the) callable bond

expected to prevail in the next period. The firm, then, would exercise at t, should it expect its cost of capital will decrease at $t+1$, which would boost the market evaluation of its debt. In this case, the price of a callable bond is clearly just K. Otherwise, the value of the callable bond is its discounted expected value over the next period. To summarize,

$$V_t = \min\{K, e^{-r_t}\mathbb{E}_t\left(V_{t+1}\right)\} = e^{-r_t}\mathbb{E}_t\left(V_{t+1}\right) - \max\{e^{-r_t}\mathbb{E}_t\left(V_{t+1}\right) - K, 0\}. \tag{12.67}$$

We can also view this problem under a slightly different angle, one where the firm may decide to issue non-callable zero coupon bonds right after having reedemed its (callable) bonds, such that the price of any callable bond could be neatly decomposed as the price of a straight, non-callable bond, minus the option to call the bond. To achieve this decomposition, let $P_\tau\left(T\right)$ denote the price of a non-callable zero coupon bond as usual and suppose that at some point in time τ^*, interest rates have decreased to an extent to have made $P_{\tau^*}\left(T\right)$ sufficiently large, in a sense to be explained in a moment. The problem we want to study is actually one where the issuer is seeking for an optimal stopping time τ^* at which it can redeem the bonds for K, and issue new *non-callable* debt at $\tau = \tau^*$, priced at $P_{\tau^*}\left(T\right)$. This would allow the issuer to cash in a difference equal to $P_{\tau^*}\left(T\right) - K$. Note that by doing so, the bond-issuer is left with the same optionalities it would have by not exercising the option to call, but with the additional "money-shower," $P_{\tau^*}\left(T\right) - K$. It is, therefore, in the interest of the bond-issuer to exercise at τ^*, whenever the difference $P_{\tau^*}\left(T\right) - K$ is positive and sufficiently large, and it is obviously not otherwise. Naturally, we consider re-issuance of non-callable debt because we wish to achieve a neat decomposition of the value of callable debt in terms of *non*-callable debt and an option to call.

How large the difference $P_{\tau^*}\left(T\right) - K$ has to be? It is a real option problem of the kind studied in Chapter 5 or in Part II (see Chapters 9 and 11). The objective of the firm is to maximize the present value of the "money shower" at some optimal stopping time τ^*, viz

$$\mathcal{C}_t \equiv \sup_\tau \mathbb{E}_t\left(e^{-\sum_{u=t}^{\tau} r_u}\left(P_\tau\left(T\right) - K\right)^+\right), \tag{12.68}$$

which gives rise to a free-exercise boundary, $\tau^* = \inf_{\tau\in[t,T]}\{\tau : P_\tau\left(T\right) \geq P_\tau^*\}$. Note that due to finite maturity, the free boundary P_τ^* may well be time-dependent, as in the dice examples of Chapter 5 (see Section 5.7).

We conjecture that the value of callable debt can be decomposed at any time as the value of a non-callable debt minus the American option price $\mathcal{C}_t$ in Eq. (12.68),

$$V_t = P_t\left(T\right) - \mathcal{C}_t. \tag{12.69}$$

Replacing Eq. (12.69) into Eq. (12.67) leaves

$$\begin{aligned} P_t\left(T\right) - \mathcal{C}_t &= e^{-r_t}\mathbb{E}_t\left(P_{t+1}\left(T\right) - \mathcal{C}_{t+1}\right) - \max\{e^{-r_t}\mathbb{E}_t\left(P_{t+1}\left(T\right) - \mathcal{C}_{t+1}\right) - K, 0\} \\ &= P_t\left(T\right) - \max\{\left(P_t\left(T\right) - K\right)^+, e^{-r_t}\mathbb{E}_t\left(\mathcal{C}_{t+1}\right)\}, \end{aligned}$$

where the second equality follows by the martingale property of the bond price rescaled by the money market account, and by rearranging terms. That is,

$$\mathcal{C}_t = \max\{\left(P_t\left(T\right) - K\right)^+, e^{-r_t}\mathbb{E}_t\left(\mathcal{C}_{t+1}\right)\}, \tag{12.70}$$

confirming Eq. (12.68). In other words, the optimal stopping time for the problem in Eq. (12.67) collapses to that in Eq. (12.68).

Puttable bonds, instead, are assets that give the holder the right to sell the bonds back to the issuer at some exercise price, either at a fixed maturity date or any fixed date before the expiration. The bondholders would exercise their option to tender the bonds to the issuer when market conditions improve from their perspective, i.e. when interest rates are high enough, so as to make bond prices lower than the exercise price. Therefore, issuing puttable bonds leads to a lower cost of capital, to the extent of the value of the American *put* option given to the bondholders to tender the bonds at the strike K.

Suppose, for example, that the price of a *non*-puttable bond, $P_\tau(T)$ as usual, lowers to a level sufficiently lower than the strike price K, in a sense to be determined in a moment. The bondholders, then, will find convenient to tender the bonds at K, buying conventional bonds at $P_\tau(T)$, thereby cashing in $K - P_\tau(T)$, and then wait until maturity. This trade would provide bondholders with a "money-shower" equal to $K - P_\tau(T)$, at the exercise date. Alternatively, the bondholders would not exercise, and wait until maturity, in which case they would not receive the profit $K - P$, at the exercise date. Therefore, it is optimal to exercise when $K > P_\tau(T)$, with the price being sufficiently low for some τ^*, and it is obviously not otherwise. Thus, the value of a puttable zero coupon bond, $V_t^{\mathrm{p}} \equiv V_t^{\mathrm{p}}(T)$ say, satisfies

$$V_t^{\mathrm{p}} = \max\{K, e^{-r_t}\mathbb{E}_t\left(V_{t+1}^{\mathrm{p}}\right)\} = e^{-r_t}\mathbb{E}_t\left(V_{t+1}\right) + \max\{K - e^{-r_t}\mathbb{E}_t\left(V_{t+1}^{\mathrm{p}}\right), 0\}. \tag{12.71}$$

Conjecture that

$$V_t^{\mathrm{p}} = P_t(T) + \mathcal{P}_t, \tag{12.72}$$

where $\mathcal{P}_t$ is the value of an American put on the zero-coupon bond, solution to,

$$\mathcal{P}_t = \max\{(K - P_t(T))^+, e^{-r_t}\mathbb{E}_t(\mathcal{P}_{t+1})\}. \tag{12.73}$$

Substituting Eq. (12.72) into Eq. (12.71) leaves Eq. (12.73) by arguments similar to those leading to Eq. (12.70).

Convertible bonds are assets that give the holder the right to convert them into a prespecified number of shares of the firm. Their value at each date when the conversion takes place is $\max\{\mathrm{CV}, P\} = P + \max\{\mathrm{CV} - P, 0\}$, where CV denotes the conversion value of the bonds, expressed in terms of the value of the firm's shares: issuing convertible bonds now lowers the cost of capital to the extent of the option given to the bondholders to convert the bonds into shares. Convertible bonds can be made callable by the bond-issuers, at a strike K. Usually, if the bonds are called, the convertible bondholders have the option to either tender the bonds to the firm, or to convert them. On the other hand, the only reason the bond-issuers might call is that the price of the convertibles is up, compared to the strike price. Therefore, the option to make convertible bonds also callable puts a ceiling to the price of the convertibles bonds, given by the exercise price, K. Precisely, the value of a *convertible callable* bond at each potential conversion date is $\max\{\mathrm{CV}, \min\{P, K\}\}$: the option to call back takes away some of the optionality from the bondholders, who are, in effect, forced to convert, as soon as the price P increases to a level beyond K.

For the previous mechanism to work, the conversion value and the bond price cannot be both continuous processes. Alternatively, we need to think in terms of a discrete time context. Consider the following events triggering the convertible bond-holder to convert after the issuer decides to call. First, the issuer calls as soon as $K < P_t$, such that $\max\{\mathrm{CV}_t, \min\{P_t, K\}\} = \max\{\mathrm{CV}_t, K\}$. Then, the convertible bond-holder decides to convert when $\mathrm{CV}_t > K$, regardless of whether CV_t is larger than P_t. For example, it may be that $\mathrm{CV}_t < P_t$, in which case conversion

would not have taken place without the issuer option to call at K. Note that with this option to call, we could well have that[9]

$$K < \mathrm{CV}_t < P_t. \tag{12.74}$$

That is, the option to call provides investors with incentives to exercise the option to convert earlier: now, bondholders convert even if CV_t is less than the bond price.

12.7.2 Callable bonds

12.7.2.1 Copying with credit risk

We can price callable bonds through trees in a way that the initial yield curve is fitted without errors, relying on the methodology in this chapter. One issue to take into account is the presence of credit risk. We may proceed as follows.

(i) First, we "populate" a short-term rate tree through one of the models described in this chapter—say, for example, through the Black, Derman and Toy (1990) model.

(ii) Second, we rely on the implied short-term rate process of the previous step and price a callable coupon bearing bond without default risk. In each node, we compare the strike price K with the rolled-back (ex-coupon) bond value, take the minimum of the two, add the coupon to this minimum and find, then, the market value of the callable bond at the relevant node,

$$V_t(s) = \min\{K, e^{-r_t(s)}\mathbb{E}_t(V_{t+1}|\, s)\} + \text{coupon},$$

where $V_t(s)$ denote the value of the callable coupon bearing bond at time t and state s, $r_t(s)$ is the short-term process at t and state s, and $e^{-r_t(s)}\mathbb{E}_t(V_{t+1}|\, s)$ is the rolled-back bond value at time t and state s. As usual, this rolled-back value is found recursively, i.e., by discounting the risk-neutral expectation of the future values of the coupon-bearing callable bond value.

(iii) Third, we correct for credit risk. The price of the callable bond in the previous step is likely to exceed the market price, due to credit risk. One then proceeds with adding a constant *spread* to the short-term rate process in step one. The resulting credit risk-adjusted interest rate tree is used to implement step (ii). If the model-based callable bond is valued less than the market, one re-calibrates the interest rate tree with a lower credit spread, until convergence is achieved by which market and model prices of the callable bonds are the same. The resulting spread is usually referred to as the *option-adjusted spread* to emphasize it is determined while taking into account the optionalities regarding the callable bonds.

At this point, we may price derivatives written on callable defaultable coupon bearing bonds—for example, options.[10] We now illustrate a simple case, relating to the the pricing of a callable bond without credit risk.

[9] If $\mathrm{CV}_0 < P_0$, and both CV_t and P_t are continuous, the inequalities in (12.74) could not hold. If at some point $s < t$, it so happens that $K < P_s$, the issuer would call the bond while providing the bondholder with the option to convert. But the bondholder would not convert as due to continuity, CV_t would still be below K.

[10] Ho and Lee (2004) (Chapter 8, Section 8.3 p. 274-278) contain exercises on the pricing of options on callable bonds relying on trees such as those described in this section.

12.7.2.2 A numerical example: without credit

Assume that the discretely compounded six-month rate, or the "short-term rate," evolves over time according to the tree described in Figure 12.15.

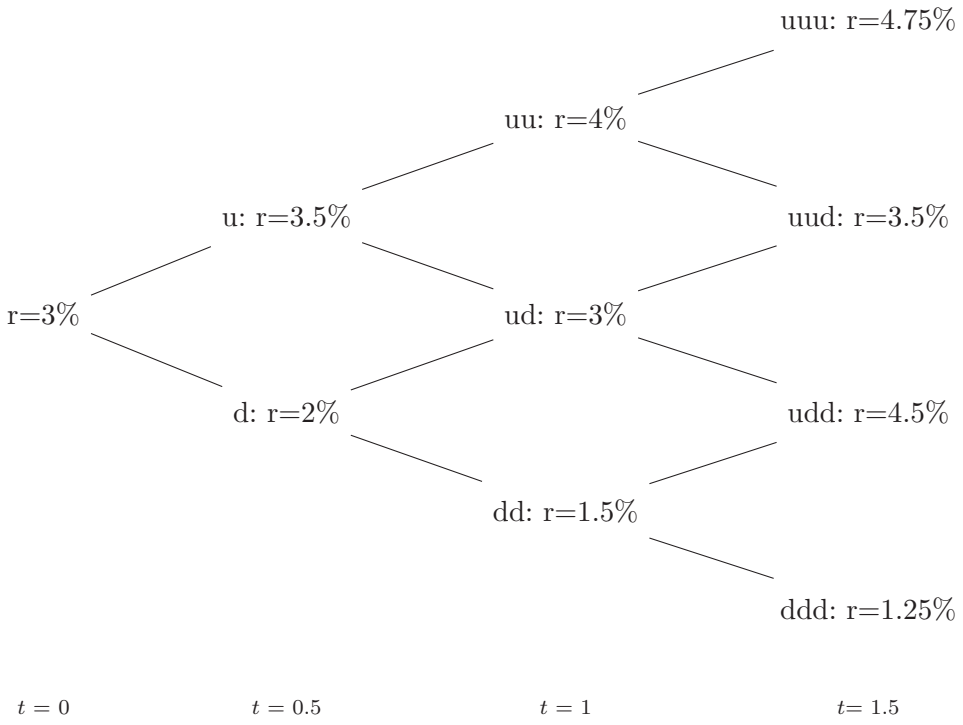


FIGURE 12.15.

Next, consider a bond expiring in two years, paying off coupon rates of 3% of the principal of £1 every six months, and callable at any time by the issuer, at par value. Let this bond be labeled "BCX." Suppose that the prices of three zero coupon bonds expiring in one year, eighteen months and two years are, respectively, 0.94632, 0.91876 and 0.89166. We can use these market data to calibrate the risk-neutral probabilities of upward movements in the short-term rate implied by the binomial tree in Figure 12.15, provided these risk-neutral probabilities depend only on calendar time t, not on the specific state of nature at time t.

We assume that available for trading is also a conventional, (i.e. non-callable) bond maturing in two years and paying coupons semiannually, at 3% of the principal of £1. We wish to calculate the price movements of the non-callable coupon-bearing two year bond. We have, $P_{\$}(0,0.5) = \frac{1}{1.03} = 0.97087$. Furthermore, as regards the zero expiring in one year:

$$P_{\$}(0,1) = 0.94632 = P_{\$}(0,0.5)\left(q_0\frac{1}{1.035} + (1-q_0)\frac{1}{1.02}\right),$$

which solved for q_0 delivers $q_0 = 0.40$. As for the zero expiring in 1.5 years,

$$\begin{aligned} P_{\$}(0,1.5) &= 0.91876 = P_{\$}(0,0.5)\left(q_0 P_u(0.5,1.5) + (1-q_0)P_d(0.5,1.5)\right) \\ &= 0.97087 \cdot \left(0.40 \cdot P_u(0.5,1.5) + 0.60 \cdot P_d(0.5,1.5)\right), \end{aligned}$$

where

$$P_u(0.5,1.5)=\frac{1}{1.035}\left(q_1\frac{1}{1.04}+(1-q_1)\frac{1}{1.03}\right),\ P_d(0.5,1.5)=\frac{1}{1.02}\left(q_1\frac{1}{1.03}+(1-q_1)\frac{1}{1.015}\right)$$

Solving for q_1 leaves $q_1=0.70$. As for the zero expiring in 2 years:

$$\begin{aligned}P_{\$}(0,2)&=0.89166=P_{\$}(0,0.5)\cdot(q_0P_u(0.5,2)+(1-q_0)P_d(0.5,2))\\&=0.97087\cdot(0.40\cdot P_u(0.5,2)+0.60\cdot P_d(0.5,2)),\end{aligned}$$

where

$$\begin{aligned}P_u(0.5,2)&=\frac{1}{1.035}(q_1P_{uu}(1,2)+(1-q_1)P_{ud}(1,2))\\&=\frac{1}{1.035}(0.70\cdot P_{uu}(1,2)+0.30\cdot P_{ud}(1,2))\\P_d(0.5,2)&=\frac{1}{1.02}(q_1P_{ud}(1,2)+(1-q_1)P_{dd}(1,2))\\&=\frac{1}{1.02}(0.70\cdot P_{ud}(1,2)+0.30\cdot P_{dd}(1,2))\end{aligned}$$

and:

$$\begin{aligned}P_{uu}(1,2)&=\frac{1}{1.04}\left(q_2\frac{1}{1.0475}+(1-q_2)\frac{1}{1.035}\right)\\P_{ud}(1,2)&=\frac{1}{1.03}\left(q_2\frac{1}{1.035}+(1-q_2)\frac{1}{1.02}\right)\\P_{dd}(1,2)&=\frac{1}{1.015}\left(q_2\frac{1}{1.02}+(1-q_2)\frac{1}{1.0125}\right)\end{aligned}$$

Solving for q_2 leaves $q_2=0.60$.

The price of a coupon bearing bond yielding 3% of the principal every six months is easy to calculate,

$$B(0,2)=0.03\times(0.97087+0.94632+0.91876+0.89166)+0.89166=1.0035. \tag{12.75}$$

Given the market data and the previously calibrated risk-neutral probabilities, we now proceed with the calculation of the price of the callable coupon bearing bond. We discount the expected cash flows, through the evaluation formula, $\min\{D,1\}+0.03$, where D is the present value of the future expected discounted cash flows promised at each node by a callable bond with the same strike price K. We have:

(i) At $t=1.5$ years,

- uuu: $\frac{1.03}{1.0475}=0.98329$ vs $1\Rightarrow$ wait, and the value of the callable bond is 0.98329.
- uud: $\frac{1.03}{1.035}=0.99517$ vs $1\Rightarrow$ wait, and the value of the callable bond is 0.99517.
- udd: $\frac{1.03}{1.02}$ vs $1\Rightarrow$ exercise, and the value of the callable bond is 1.
- ddd: $\frac{1.03}{1.0125}$ vs $1\Rightarrow$ exercise, and the value of the callable bond is 1.

(ii) At $t = 1$ year, we have that $q = 60\%$, and, then:

- uu: $\frac{1}{1.04}\left[0.6\frac{1.03}{1.0475} + 0.4\frac{1.03}{1.035} + 0.03\right] = 0.97889$ vs $1 \Rightarrow$ wait, and the value of the callable bond is 0.97889.
- ud: $\frac{1}{1.03}\left[0.6\frac{1.03}{1.035} + 0.4 \cdot 1 + 0.03\right] = 0.99719$ vs $1 \Rightarrow$ wait, and the value of the callable bond is 0.99719.
- dd: $\frac{1}{1.015}\left[0.6 \cdot 1 + 0.4 \cdot 1 + 0.03\right] = 1.0148$ vs $1 \Rightarrow$ exercise, and the value of the callable bond is 1.

(iii) At $t = 0.5$ years, we have that $q = 70\%$, and, then:

- u: $\frac{1}{1.035}\left[(0.7 \times 0.97889 + 0.3 \times 0.99719) + 0.03\right] = 0.98008$ vs $1 \Rightarrow$ wait, and the value of the callable bond is 0.98008.
- d: $\frac{1}{1.02}\left[(0.7 \times 0.99719 + 0.3 \times 1) + 0.03\right] = 1.0079$ vs $1 \Rightarrow$ exercise, and the value of the callable bond is 1.

Finally, at the time of evaluation, we have that $q = 40\%$, and, then, the price of the callable bond is:

$$P^c = \frac{1}{1.03}\left(0.40 \times 0.98008 + 0.60 \times 1 + 0.03\right) = 0.99226.$$

Naturally, the callable bond is valued less than the conventional bond $B(0,2)$ in Eq. (12.75): the difference is the value of the option given to the issuer to redeem these bonds, and arises when the interest rates go sufficiently down—negative convexity.

How would we proceed to price the BCX bond if the previous market data were unavailable? In particular, suppose that (i) the risk-neutral probabilities of upward movements in the short-term rate are: (i.a) unknown from time zero to 0.5 years; (i.b) 70%, from 0.5 to one year; and (i.c) 60%, from one to 1.5 years; (ii) available for trading is a European call option written on the BCX bond; (iii) this option, which quotes for £1.7226×10^{-3}, expires in 1.5 years, is struck at £0.99000, and becomes worthless as soon as the underlying callable bond is called back by the issuer. First, note that at the expiration, $t = 1.5$ years, the payoffs of the option are:

$$C_{uuu} = 0, \quad C_{uud} = 0.00517,$$

and, because of the sudden death assumption,

$$C_{udd} = C_{ddd} = 0.00000.$$

At $t = 1$ year, we have that $q = 60\%$, and, then

$$\begin{aligned} C_{uu} &= \frac{1}{1.04}\left(0.6 \times 0 + 0.4 \times 0.00517\right) = 1.9885 \times 10^{-3} \\ C_{ud} &= \frac{1}{1.03}\left(0.6 \times 0.00517 + 0.4 \times 0\right) = 3.0117 \times 10^{-3} \\ C_{dd} &= 0 \end{aligned}$$

At $t = 0.5$ years, we have that $q = 70\%$, and, then

$$\begin{aligned} C_u &= \frac{1}{1.035}\left(0.70 \cdot C_{uu} + 0.30 \cdot C_{ud}\right) \\ &= \frac{1}{1.035}\left(0.70 \cdot 1.9885 \times 10^{-3} + 0.30 \cdot 3.0117 \times 10^{-3}\right) = 2.2178 \times 10^{-3}. \end{aligned}$$

$C_d = 0$, by the sudden death assumption.

At the time of evaluation, the price of the call is

$$C = 1.7226 \times 10^{-3} = \frac{1}{1.03}\left(qC_u + (1-q)\,C_d\right) = \frac{1}{1.03} \cdot q \cdot 2.2178 \times 10^{-3},$$

where q is the risk-neutral probability of an upward movement in the short-term rate during the first six months. We can solve for this q, obtaining $q = 80\%$. Finally, given this probability, we can calculate the price of the callable bond. We have

$$P^{\text{ca}} = \frac{1}{1.03}\left(0.80 \times 0.98008 + 0.20 \times 1 + 0.03\right) = 0.98453.$$

It is lower than the price calculated earlier, because the price of the option is giving more weight (80%) than before (40%) to the occurrence of the state of the world where the interest rate goes up.

12.7.3 Convertible bonds

12.7.3.1 Evaluation issues

Consider the following convertible and callable bond. Let K be the strike at which the bond can be called by the bond-issuer, and let the parity, or *conversion value*, be $\text{CV} = \text{CR} \times S$, where S is the price of the common share. To evaluate this bond through a binomial tree, we may proceed through the following three steps:

(i) We set the life of the tree equal to the life of the callable convertible bond.

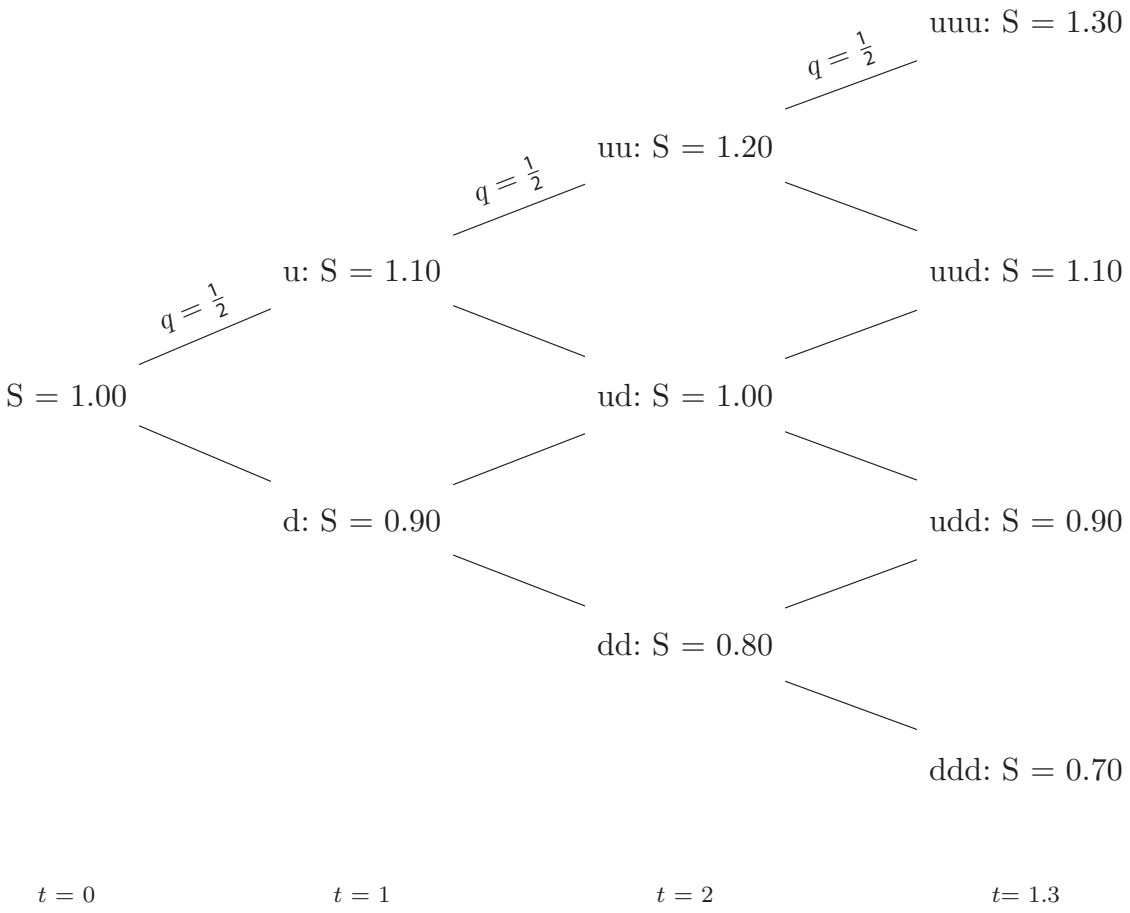
(ii) We assess the evolution of the stock price along the tree, under the risk-neutral probability. This is done following the standard Cox, Ross and Rubinstein (1979) approach.

(iii) We determine the value of the bond as $\max\{\text{CV}, \min\{B, K\}\}$, where B is the value of the bond, "rolled-back" from the values of the bond in the next nodes through the usual recursive, backward method—relying on calculating the present value of the risk-neutral expectation of the future payoffs. That is, assuming the bondholder does not convert, the value is $B^* = \min\{B, K\}$, where B is the "rolled-back" value of the bond. Then, the value is $\max\{\text{CV}, B^*\}$.

Note, this procedure leads to fill in the nodes, once we know the appropriate interest rate. If the firm was not subject to default risk, we would simply use the riskless interest rate. However, the firm is obviously subject to default risk. In practice, we proceed as follows. In each node, the value of the bond is decomposed into two parts. One part, related to the "pure debt component," discounted at the defaultable interest rate; and one part related to the "pure equity component," discounted at the default-free interest rate. Exercise 25.7 in Hull (2003) (p. 653-654) illustrates a specific example.

12.7.3.2 A numerical example: without credit

Consider a three year convertible bond, which can be converted at any time into one share of the underlying firm's stock. The bond has a face value equal to 1, it is default-free, and pays off a coupon of 3% of the face value every year, except the time at which it is issued. Moreover, in each period, it pays off the coupon, regardless of whether it will be converted or not.



The price of the share is assumed to be unaffected by any decision relating to the conversion of the bond, and evolves over time as described by the following tree:

In the previous diagram, each period corresponds to one year, S denotes the price of the share, and $q = \frac{1}{2}$ is the constant risk-neutral probability of price movements. Assume, finally, that the yield curve is flat at 3%, discretely compounded, and that it will remain such over the next three periods, and in each state of the world.

We proceed to calculate the conversion value of the convertible bond at each node of the tree. We shall identify, then, the nodes where it is optimal for the bond-holder to convert. Finally, we shall determine the value of the convertible bond at time $t = 0$, as well as the value of the option to convert. As for the conversion value, we know this is simply the product of the conversion ratio times the current value of the outstanding stock, and equals $\text{CV}_t = \text{CR} \cdot S_t = S_t$, as the conversion ratio is one. To find the current value of the convertible bond, we proceed recursively, as explained earlier, and calculate, for each date t and each node, $\max\{\text{CV}_t, B_t\}$, where B_t denotes the present value of the future cash flows of the convertible, in case of no conversion at time t. The payoffs at time $t = 3$ are:

uuu: $\text{CV} = S = 1.30$, $B_{uuu} \equiv \max\{\text{CV}, 1\} + 0.03 = 1.33$, convert

uud: $\text{CV} = S = 1.10$, $B_{uud} \equiv \max\{\text{CV}, 1\} + 0.03 = 1.13$, convert

udd: $\text{CV} = S = 0.90$, $B_{udd} \equiv \max\{\text{CV}, 1\} + 0.03 = 1.03$

ddd: $\text{CV} = S = 0.70$, $B_{ddd} \equiv \max\{\text{CV}, 1\} + 0.03 = 1.03$

We have:

$t=2$ uu: $\text{CV}=S=1.20$, $B_{uu}\equiv\max\{\text{CV},A_1\}+0.03$, $A_1\equiv\frac{1}{1.03}\frac{1}{2}(B_{uuu}+B_{uud})=\frac{1}{1.03}\frac{1}{2}(1.33+1.13)=1.19420$. Hence $B_{uu}=1.23000$, convert

ud: $\text{CV}=S=1.00$, $B_{ud}\equiv\max\{\text{CV},A_2\}+0.03$, $A_2\equiv\frac{1}{1.03}\frac{1}{2}(B_{uud}+B_{udd})=\frac{1}{1.03}\frac{1}{2}(1.13+1.03)=1.04850$. Hence $B_{ud}=1.07850$

dd: $\text{CV}=S=0.80$, $B_{dd}\equiv\max\{\text{CV},A_3\}+0.03$, $A_3\equiv\frac{1}{1.03}\frac{1}{2}(B_{udd}+B_{ddd})=1$. Hence $B_{dd}=1.03000$

$t=1$ u: $\text{CV}=S=1.10$, $B_u\equiv\max\{\text{CV},L_1\}+0.03$, $L_1\equiv\frac{1}{1.03}\frac{1}{2}(B_{uu}+B_{ud})=\frac{1}{1.03}\frac{1}{2}(1.23+1.07850)=1.1206$. Hence $B_u=1.15060$

d: $\text{CV}=S=0.90$, $B_d\equiv\max\{\text{CV},L_2\}+0.03$, $L_2\equiv\frac{1}{1.03}\frac{1}{2}(B_{ud}+B_{dd})=\frac{1}{1.03}\frac{1}{2}(1.07850+1.03000)=1.0235$. Hence $B_d=1.0535$

Finally, we have that:

$$B^{\text{co}}=\frac{1}{0.03}\frac{1}{2}(B_u+B_d)=\frac{1}{1.03}\frac{1}{2}(1.15060+1.0535)=1.0700.$$

Instead, and because the YTM is the same as the coupon rate, the value of a non-convertible three year coupon bearing bond is 1 (see Eq. (12.6)). Therefore, the option to convert is worth 0.07000.

Next, assume that the convertible bond is also callable by the issuer, at any time, and at a strike value of 1.02000, and if it is called, the bond-holder has the option to tender the bond or to convert it into one share. This convertible, and callable, bond can be evaluated as in the previous calculations, although the formula to use in each node now is $\max\{\text{CV}_t,\min\{B_t,K\}\}$, with $K=1.02000$. The payoffs at time $t=3$ are:

uuu: $\text{CV}=S=1.30$, $B_{uuu}\equiv\max\{\text{CV},\min\{B=1,1.02\}\}+0.03=1.33$, convert

uud: $\text{CV}=S=1.10$, $B_{uud}\equiv\max\{\text{CV},\min\{B=1,1.02\}\}+0.03=1.13$, convert

udd: $\text{CV}=S=0.90$, $B_{udd}\equiv\max\{\text{CV},\min\{B=1,1.02\}\}+0.03=1.03$

ddd: $\text{CV}=S=0.70$, $B_{ddd}\equiv\max\{\text{CV},\min\{B=1,1.02\}\}+0.03=1.03$

We have:

$t=2$ uu: $\text{CV}=S=1.20$, $B_{uu}\equiv\max\{\text{CV},\min\{A_1,1.02\}\}+0.03$, $A_1\equiv\frac{1}{1.03}\frac{1}{2}(B_{uuu}+B_{uud})=\frac{1}{1.03}\frac{1}{2}(1.33+1.13)=1.19420$. Hence $B_{uu}=\max\{\text{CV},1.02\}+0.03=1.23$. The bond is called, and then converted

ud: $\text{CV}=S=1.00$, $B_{ud}\equiv\max\{\text{CV},\min\{A_2,1.02\}\}+0.03$, $A_2\equiv\frac{1}{1.03}\frac{1}{2}(B_{uud}+B_{udd})=\frac{1}{1.03}\frac{1}{2}(1.13+1.03)=1.04850$. Hence $B_{ud}=\max\{\text{CV},1.02\}+0.03=1.05$. The bond is called, but not converted

dd: $\text{CV}=S=0.80$, $B_{dd}\equiv\max\{\text{CV},\min\{A_3,1.02\}\}+0.03$, $A_3\equiv\frac{1}{1.03}\frac{1}{2}(B_{udd}+B_{ddd})=1$. Hence $B_{dd}=\max\{\text{CV},1\}+0.03=1.03000$

$t = 1$ u: $\text{CV} = S = 1.10$, $B_u \equiv \max\{\text{CV}, \min\{L_1, 1.02\}\} + 0.03$, $L_1 \equiv \frac{1}{1.03}\frac{1}{2}(B_{uu} + B_{ud}) = \frac{1}{1.03}\frac{1}{2}(1.23 + 1.05) = 1.1068$. Hence $B_u = \max\{\text{CV}, 1.02\} + 0.03 = 1.13000$. The bond is called, and then converted. The event at this node illustrates how the presence of the option to call the bond incentivizes investors to convert "earlier." Indeed, without this optionality opened to the firm at $t = 1$, the bond value would be L_1, which is larger than the conversion value, $\text{CV} < L_1$, in which case the bondholder would not convert

d: $\text{CV} = S = 0.90$, $B_d \equiv \max\{\text{CV}, \min\{L_2, 1.02\}\} + 0.03$, $L_2 \equiv \frac{1}{1.03}\frac{1}{2}(B_{ud} + B_{dd}) = \frac{1}{1.03}\frac{1}{2}(1.050 + 1.03000) = 1.0097$. Hence, $B_d = \max\{\text{CV}, 1.0097\} + 0.03 = 1.0397$. The bond is not called and is not converted

Therefore, we have that

$$B^{\text{co-ca}} = \frac{1}{1.03}\frac{1}{2}(B_u + B_d) = \frac{1}{1.03}\frac{1}{2}(1.1300 + 1.0397) = 1.0533.$$

As expected, the value of a convertible callable is less than that of the convertible, due to the option given to the bond-issuers to call the bond.

12.8 Appendix 1: Botstrapping and no-arbitrage restrictions

A. Extracting zeros from bond prices

Standard techniques are available to "extract" the price of zeros from that of coupon-bearing bonds, provided at least a sufficient spread of bonds exist across maturities. Consider the following example, in which available for trading are three coupon-bearing bonds: the first pays off at T_1, the second bond pays off at T_1, T_2, the third bond pays off at T_1, T_2, T_3. By no-arbitrage,

$$B \equiv \begin{bmatrix} B(T_1) \\ B(T_2) \\ B(T_3) \end{bmatrix} = \begin{bmatrix} C_{11}+1 & 0 & 0 \\ C_{21} & C_{22}+1 & 0 \\ C_{31} & C_{32} & C_{33}+1 \end{bmatrix} \begin{bmatrix} P(T_1) \\ P(T_2) \\ P(T_3) \end{bmatrix} \equiv C \cdot P, \tag{12A.1}$$

for some coupons C_{ij}. We can use Eq. (12A.1) to invert for the prices of the zeros, $P = C^{-1}B$. There is a mathematically equivalent inversion algorithm, a procedure known as "bootstrapping," based on the observation that the prices of the zeros can be solved for recursively. Let B_i be the price of a bond that pays off coupons at on dates $t_1, t_2, \cdots, t_i$, and the principal of £1 at t_i. Let P_i be the price of the zero maturing at t_i. Then, P_n can be estimated as follows:

$$P_n = \frac{B_n - \sum_{i=1}^{n-1} C_{ni} P_i}{C_{nn}+1}, \quad n = 1, \cdots, N, \tag{12A.2}$$

where N is the largest available maturity. It is straightforward to verify this formula using the example $N = 3$ in (12A.1).

To illustrate, suppose the bonds maturing at n have fixed coupons, $C_n \equiv C_{ni}$ for all i. We define the par yield as in Eq. (12.3), as the fixed sequence C_n such that the price B_n is forced to equal 100%. The following example shows how to use Eq. (12A.2) and extract zeros and, then, reconstruct a discretely compounded yield curve.

Coupon	Maturity, n	Zero price	$\sum_{i=1}^{n} P_i$	Yield curve*
6.00%	1	0.9434	0.9434	6.00%
7.00%	2	0.8728	1.8162	7.04%
8.00%	3	0.7914	2.6076	8.11%
9.50%	4	0.6870	3.2946	9.84%
9.00%	5	0.6454	3.9400	9.15%
10.50%	6	0.5306	4.4706	11.14%
11.00%	7	0.4579	4.9285	11.81%
11.25%	8	0.4005	5.3290	12.12%
11.50%	9	0.3472	5.6762	12.47%
11.75%	10	0.2980	—	12.87%

Note that Eq. (12A.2) relies on the assumption that no maturities are missing. When some of the maturity dates are not available, the required coupon rate C_{ni} can be replaced with a linear interpolation $\hat{C}_{ni}$ between $C_{n,i-1}$ and $C_{n,i+1}$, as follows,

$$\hat{C}_{ni} = \frac{t_{i+1} - t_i}{t_{i+1} - t_{i-1}} C_{n,i-1} + \frac{t_i - t_{i-1}}{t_{i+1} - t_{i-1}} C_{n,i+1}.$$

The effects of the interpolation should be "visible" near the missing maturitites.

B. Splines

Alternative to bootstrap are techniques that aim to cope with situations where the number of bonds is less than the maturity dates we want to fit. Suppose we observe N bonds, where the i-th bond entitles

to receive the coupons C_{ij}, for $j = 1, \cdots, M_i$, and that these prices are observed with errors, or

$$B(M_i) = \sum_{j=1}^{M_i} C_{ij} P(t_j) + P(t_{M_i}) + \epsilon_i, \quad i = 1, \cdots, N,$$

where ϵ_i is the measurement error for the i-th bond.

We aim to find the curve $T \mapsto P(T)$ that minimizes the errors, in some statistical sense. The natural device is to "parametrize" the function $P(T)$, with a number of k parameters, where $k < N$. To parametrize the function $P(t_j)$ for a generic t_j, we can use *polynomials*, as originally suggested by McCulloch (1971, 1975),

$$P(t_j) = 1 + a_1 t_j + a_2 t_j^2 + \cdots + a_k t_j^k,$$

where a_i are the parameters. *Cubic splines* are polynomials up to the third order, and are very popular. The parameters a_i can be estimated by minimizing the sum of the squared errors, $\sum_{i=1}^{N} \epsilon_i^2$. A well-known pitfall of polynomials is that a high k might imply that while the polynomial approximation works reasonably well near the observed maturities, it may exhibit an erratic behavior in between. To avoid this problem, we can use *local polynomials*, which are low-order polynomials (typically splines) fitted to non-overlapping subintervals.

Naturally, we may also want to parametrize the spot rates, $R(T)$, as polynomials. The next chapter (Section 13.3) explains that an alternative to polynomials are parametrizations introduced by Nelson and Siegel (1987), where

$$R(T) = \beta_1 + \beta_2 \left(\frac{1 - e^{-\lambda T}}{\lambda T} \right) + \beta_3 \left(\frac{1 - e^{-\lambda T}}{\lambda T} - e^{-\lambda T} \right), \tag{12A.3}$$

and β_i and λ are parameters. We provide interpretations of this parametrization in the next chapter, where we also explain how this has been used in practice to forecast the yield curve.

C. No arb restrictions

Bond prices need to satisfy restrictions that prevent arbitrage. We use figures taken from Tuckman (2002) (p. 8-12), and illustrate how an arbitrage opportunity can arise and be exploited in this context.

Data

Suppose that on some hypothetical date, say Februrary 15, 09, we observe Set I of bond prices in the left panel of the following table, and that bootstrap leads to to the implicit zeros in the corresponding right panel of the table. Also assumed is that we observe additional bond prices, those in Set II in the lower part of the table. Are these additional prices, those in Set II, compatible with those in Set I, in terms of arbitrage opportunities?

Set I: Treasury Bond prices

Coupon	Maturity	Market price
7.875%	8/15/09	101.400
14.250%	2/15/10	108.980
6.375%	8/15/10	102.160
6.250%	2/15/11	102.570
5.250%	8/15/11	100.840

Bootstrapped zeros

Time to maturity	Implicit zero
0.5	$p(0, 0.5) = 0.97557$
1.0	$p(0, 1.0) = 0.95247$
1.5	$p(0, 1.5) = 0.93045$
2.0	$p(0, 2.0) = 0.90796$
2.5	$p(0, 2.5) = 0.88630$

Set II: Treasury Bond prices

Coupon	Maturity	Market price
13.375%	8/15/09	104.080
10.750%	2/15/11	110.938
5.750%	8/15/11	102.020
11.125%	8/15/11	114.375

A basic no-arb condition

We cast the problem in a more general format. Suppose we observe a vector of N bond prices, with a $N \times N$ matrix C of coupons, where each row of C gives the stream of the coupons promised by a given asset. We know that the $N \times 1$ vector of zeros P, satisfies, $B = CP$. That is, assuming that the matrix C is invertible,

$$P = C^{-1}B. \tag{12A.4}$$

Next, suppose there exists some asset that: (i) promises to pay:

$$c^* = \left[\begin{array}{cccc} c_1^* & c_2^* & \cdots & c_N^* + 100 \end{array} \right],$$

and (ii) has a price, b^*, such that:

$$b^* < c^*P. \tag{12A.5}$$

The right hand side of this inequality, c^*P, is the "no-arbitrage price" of the asset, which in this example is greater than the market price. The inequality gives rise to an arbitrage opportunity, which can be exploited by going long the asset, and shorting a portfolio "synthesizing" it. To synthesize the asset to go long for, we solve the following system of N equations with N unknowns,

$$\pi C = c^*, \tag{12A.6}$$

where the vector of unknowns, π, contains the number of assets in the synthesizing portfolio: by purchasing the portfolio π, one is entitled to receive πC in the future, which we want to equal c^*. The solution to Eq. (12A.6) is:

$$\hat{\pi} = c^*C^{-1}. \tag{12A.7}$$

Accordingly, the value of this portfolio, V say, is given by, $V = \hat{\pi}B = c^*C^{-1}B = c^*P > b^*$, where the last equality follows by the "zero pricing equation" (12A.4), and the inequality holds by the inequality (12A.5).

To summarize, we now have the following situation: (i) the asset we hold produces the cash flows that are needed to pay out the coupons of the "synthesizing" portfolio we sold, and (ii) the price of the asset we go long is less than the value of the portfolio we short. This situation is an arbitrage opportunity, as initially claimed. We now use these insights to check whether arbitrage opportunities exist and exploited, using the data in Tables 12.1 through 12.3.

First step: detecting arbitrage opportunities

First, we determine *no-arbitrage* prices of the bonds in Set II, using the implicit zeros extracted from Set I. Denote these prices with B_1 (for the six month 13.375%), B_2 (for the two year 10.750%), B_3 (for the 2.5 year 5.750%), and B_4 (for the 2.5 year 11.125%). They are given by,

$$\begin{aligned}
B_1 &= \left(\tfrac{13.375}{2} + 100\right) p\,(0, 0.5) \\
B_2 &= \tfrac{10.750}{2} \left[p\,(0, 0.5) + p\,(0, 1) + p\,(0, 1.5)\right] + \left(\tfrac{10.750}{2} + 100\right) p\,(0, 2.0) \\
B_3 &= \tfrac{5.75}{2} \left[p\,(0, 0.5) + p\,(0, 1.0) + p\,(0, 1.5) + p\,(0, 2.0)\right] + \left(\tfrac{5.75}{2} + 100\right) p\,(0, 2.5) \\
B_4 &= \tfrac{11.125}{2} \left[p\,(0, 0.5) + p\,(0, 1.0) + p\,(0, 1.5) + p\,(0, 2.0)\right] + \left(\tfrac{11.25}{2} + 100\right) p\,(0, 2.5)
\end{aligned}$$

The next table provides the numerical values of these theoretical prices, comparing to their market counterparts:

Set II: Treasury Bond prices

Coupon	Maturity	Market price	No-arb price
13.375%	8/15/09	104.080	104.080
10.750%	2/15/11	110.938	111.041
5.750%	8/15/11	102.020	102.007
11.125%	8/15/11	114.375	114.511

While there are no arbitrage opportunities for the 13.375% bond expiring in six months, the price of the 10.750% bond expiring in 2 years is less than its no-arbitrage price: this bond "trades cheap." In contrast, the 2.5 year 5.750% bond "trades rich," although the resulting arbitrage does not seem to be quite sensible.

Second step: implementing the arbitrage

We proceed to exploit the mispricing related to the 10.750% bond expiring in 2 years, using the previous insights. We have, $N = 4$, and $c^* = \left[\begin{array}{cccc} \frac{1}{2}10.750 & \frac{1}{2}10.750 & \frac{1}{2}10.750 & \frac{1}{2}10.750 + 100 \end{array}\right]$. We use the first four bonds in Set I to construct an arbitrage portfolio. In terms of the coupon matrix C, we have,

$$C = \left[\begin{array}{cccc} \frac{1}{2}7.875 + 100 & 0 & 0 & 0 \\ \frac{1}{2}14.250 & \frac{1}{2}14.250 + 100 & 0 & 0 \\ \frac{1}{2}6.375 & \frac{1}{2}6.375 & \frac{1}{2}6.375 + 100 & 0 \\ \frac{1}{2}6.250 & \frac{1}{2}6.250 & \frac{1}{2}6.250 & \frac{1}{2}6.250 + 100 \end{array}\right].$$

We implement the following trade: (i) buy x 10.750% bonds expiring in 2 years, which cost $110.938 \cdot x$; (ii) create x portfolios satisfying Eq. (12A.7),

$$\hat{\pi} = c^* C^{-1} = \left[\begin{array}{cccc} 0.0189 & 0.0197 & 0.0212 & 1.0218 \end{array}\right].$$

If we short x of these portfolios, then, by construction, the coupons we need to pay are exactly matched by the coupons we receive from the x 10.750% bonds expiring in 2 years. However, the market value of the x portfolios we short equals,

$$x \cdot V = x \cdot \hat{\pi} B = x \cdot \left[\begin{array}{cccc} 0.0189 & 0.0197 & 0.0212 & 1.0218 \end{array}\right] \left[\begin{array}{c} 101.40 \\ 108.98 \\ 102.16 \\ 102.57 \end{array}\right] = x \cdot 111.041,$$

where the vector of the market prices, B, is taken from Set I. Therefore, the gains from this trade are, $x \cdot (111.041 - 110.938) = 0.103 \cdot x$. For example, by trading £1,000,000 at face value, i.e. $x = 10000$, then, arbitrage profits equal £1030.

12.9 Appendix 2: Proof of Eq. (12.17)

Let $P(r,T_i)$ denote the price of a zero with maturity $T_i,\ i=1,2$, when the interest rate is equal to r. We wish to replicate a zero with maturity T_1 by means of a portfolio that comprises (i) Δ zeros with maturity T_2 and (ii) $\pounds M$ in the MMA. Let V_0 be the current value of this portfolio; it is clearly a function of the current short-term rate r, and equals

$$V_0(r) = P(r,T_2)\,\Delta + M.$$

In the second period, the value of the portfolio is random, as it depends on the development of the short-term rate $\tilde{r}$. Precisely, the value of the portfolio in the second period, is

$$V(\tilde{r}) = \begin{cases} V(r^+) = P(r^+,T_2)\Delta + M(1+r), \text{ with probability } p \\ V(r^-) = P(r^-,T_2)\Delta + M(1+r), \text{ with probability } 1-p \end{cases}$$

We also know that in the second period, the value of the second zero is,

$$P(\tilde{r},T_1) = \begin{cases} P(r^+,T_1), & \text{with probability } p \\ P(r^-,T_1), & \text{with probability } 1-p \end{cases}$$

Next, we select Δ and M to make the portfolio value match that of the zero maturing at T_1, in each state of nature, viz

$$V(\tilde{r}) = P(\tilde{r},T_1), \quad \text{in each state.}$$

Mathematically, this is tantamount to solving the following system of two equations with two unknowns (Δ and M),

$$\begin{cases} V(r^+) = P(r^+,T_2)\,\Delta + M(1+r) = P(r^+,T_1) \\ V(r^-) = P(r^-,T_2)\,\Delta + M(1+r) = P(r^-,T_1) \end{cases} \tag{12A.8}$$

The solution is,

$$\hat{\Delta} = \frac{P(r^+,T_1) - P(r^-,T_1)}{P(r^+,T_2) - P(r^-,T_2)}, \quad \hat{M} = \frac{P(r^-,T_1)P(r^+,T_2) - P(r^+,T_1)P(r^-,T_2)}{[P(r^+,T_2) - P(r^-,T_2)](1+r)}.$$

By construction, the previous portfolio, $(\hat{\Delta},\hat{M})$, replicates the value of the second zero in the second period. But if two assets (the portfolio, and the second zero) yield the same payoffs in each state of the nature, they must be worth the same, in the absence of arbitrage. Therefore, we must have,

$$V_0(r)|_{\Delta=\hat{\Delta},M=\hat{M}} = P(r,T_2)\,\hat{\Delta} + \hat{M} = P(r,T_1),$$

or,

$$(1+r)\,\hat{M} = (1+r)\,P(r,T_1) - (1+r)\,P(r,T_2)\,\hat{\Delta}. \tag{12A.9}$$

Next, let us figure out the prediction of the model in terms of the expected return it generates for the price of the bond maturing at T_2, when $(\Delta,M) = (\hat{\Delta},\hat{M})$. To do this, multiply the first equation in (12A.8) by p, and multiply the second equation in (12A.8) by $1-p$. Add the result for $\Delta = \hat{\Delta}, M = \hat{M}$ to obtain,

$$\left[pP(r^+,T_2) + (1-p)\,P(r^-,T_2)\right]\hat{\Delta} + \hat{M}(1+r) = pP(r^+,T_1) + (1-p)P(r^-,T_1).$$

Replacing Eq. (12A.9) into the previous equation yields,

$$\begin{aligned} &\left[\left(pP(r^+,T_2) + (1-p)\,P(r^-,T_2)\right) - (1+r)P(r,T_2)\right]\hat{\Delta} \\ &= \left[pP(r^+,T_1) + (1-p)P(r^-,T_1)\right] - (1+r)P(r,T_1). \end{aligned}$$

Finally, replacing the solution for $\hat{\Delta}$ into the previous equation leaves,

$$\frac{[pP(r^+,T_2)+(1-p)P(r^-,T_2)]-(1+r)P(r,T_2)}{P(r^+,T_2)-P(r^-,T_2)}=\frac{[pP(r^+,T_1)+(1-p)P(r^-,T_1)]-(1+r)P(r,T_1)}{P(r^+,T_1)-P(r^-,T_1)}.$$

The previous equation is easy to interpret. The numerators are the expected *excess* returns from holding the assets. They equal $E_p[P(\tilde{r},T_i)]-(1+r)P(r,T_i)$, where $E_p[P(\tilde{r},T_i)]$ is what the investors expect to receive, the next period, by investing $£P(r,T_i)$ today, in the bond; and $(1+r)P(r,T_i)$ is what the investors expect to receive, the next period, by investing $£P(r,T_i)$ today, in the MMA. The denominators constitute a measure of volatility related to holding the assets. The previous equation then tells us that the Sharpe ratios (or the unit risk premiums) on the two zeros agree.

Let the Sharpe ratio on any zero be equal to some function λ of the short-term rate r only (and possibly of calendar time). This function, λ, does not clearly depend on the maturity of the zeros. Then, we have,

$$\begin{aligned}\left[pP(r^+,T_2)+(1-p)P(r^-,T_2)\right]-(1+r)P(r,T_2)&=\left[P(r^+,T_2)-P(r^-,T_2)\right]\lambda\\&=\frac{P(r^+,T_2)-P(r^-,T_2)}{r^+-r^-}\cdot[(r^+-r^-)\lambda]. \quad \text{(12A.10)}\end{aligned}$$

We can interpret (r^+-r^-) as a measure of interest rate volatility, and define $\text{Vol}(\tilde{r}-r)\equiv(r^+-r^-)$. Eq. (12.17) follows by rewriting Eq. (12A.10) for a generic maturity date $T>2$.

12.10 Appendix 2: The Ho and Lee price representation

Define the discretely compounded forward rate, as the number $F_T(\tau) \equiv F(\tau, T, T+1)$, satisfying: $\frac{P(\tau,T+1)}{P(\tau,T)} = \frac{1}{1+F_T(\tau)}$, as in Eq. (12.4) of the main text. Iterating this equation leaves:

$$P(\tau,T) = \prod_{S=\tau}^{T-1} \frac{1}{1+F_S(\tau)} = \frac{P(t,T)}{P(t,\tau)} \frac{P(t,\tau)}{P(t,T)} \prod_{S=\tau}^{T-1} \frac{1}{1+F_S(\tau)}.$$

Therefore, for any $t : t < \tau < T$, we have that,

$$P(\tau,T) = \frac{P(t,T)}{P(t,\tau)} \prod_{S=\tau}^{T-1} \frac{1+F_S(t)}{1+F_S(\tau)}. \tag{12A.11}$$

Eq. (12A.11) is a convenient representation of the bond price at a future date τ: it is the ratio of the two *current* prices $P(t,T)$ and $P(t,\tau)$, and a factor relating to the development of *forward rates* from the current time t to time τ, i.e. $\frac{1+F_S(t)}{1+F_S(\tau)}$, for $S = \tau, \cdots, T-1$. Hence, once we model forward rates, we have implications for bond price movements.

We normalize the time-line and set $t = 0$. Redefining $\tau = t$, Eq. (12A.11) reduces to,

$$P(t,T) = \frac{P(0,T)}{P(0,t)} \prod_{S=\tau}^{T-1} \frac{1+F_S(0)}{1+F_S(t)}. \tag{12A.12}$$

Eq. (12.36) in the main text follows by Eq. (12A.12).

Next, we search for the model's predictions on forward rates, i.e. we prove Eq. (12.37). The proof is by induction. Eq. (12.37) holds true for $t = 0$. Next, suppose that it holds at time t. We wish to show that in this case, Eq. (12.37) would also hold at time $t+1$. At time $t+1$, we have two cases.

- *Case 1*: A *positive* price jump occurs between time t and time $t+1$. In this case,

$$\begin{aligned}
\hat{F}_S^{j+1}(t+1) &= \ln \frac{P_{j+1}(t+1,S)}{P_{j+1}(t+1,S+1)} \\
&= \ln \left[u(S-t) \frac{P_j(t,S)}{P_j(t,t+1)} \right] - \ln \left[u(S+1-t) \frac{P_j(t,S+1)}{P_j(t,t+1)} \right] \\
&= \ln \frac{u(S-t)}{u(S+1-t)} + \hat{F}_S^j(t) \\
&= \ln \frac{u(S+1-(t+1))}{u(S+1)} + \hat{F}_S(0) - [(t+1)-(j+1)] \ln \delta,
\end{aligned}$$

  where the first equality and the third follow by the definition of $\hat{F}_S^{j+1}(t)$, the second equality holds by the definition of the jump in Eq. (12.27), the fourth equality follows by Eq. (12.37). Hence, Eq. (12.37) holds at time $t+1$ in the occurrence of a *positive* price jump between time t and time $t+1$.

- *Case 2*: A *negative* price jump occurs between time t and time $t+1$. In this case,

$$\begin{aligned}
\hat{F}_S^j(t+1) &= \ln \frac{P_j(t+1,S)}{P_j(t+1,S+1)} \\
&= \ln \left[d(S-t) \frac{P_j(t,S)}{P_j(t,t+1)} \right] - \ln \left[d(S+1-t) \frac{P_j(t,S+1)}{P_j(t,t+1)} \right] \\
&= \ln \frac{d(S-t)}{d(S+1-t)} + \hat{F}_S^j(t) \\
&= \ln \frac{d(S-t)\,\delta^{-(S-t)+1}}{d(S+1-t)\,\delta^{-(S+1-t)+1}} \delta^{-1} + \hat{F}_S(0) + \ln \frac{u(S+1-t)}{u(S+1)} - (t-j)\ln\delta \\
&= \ln \frac{u(S-t)}{u(S+1)} + \hat{F}_S(0) - [(t+1)-j]\ln\delta,
\end{aligned}$$

where the first four equalities follow by the same arguments produced in Case 1, the fifth equality holds by the relation $u(T) = d(T)\,\delta^{-(T-1)}$ in Eq. (12.33) and the last equality follows by rearranging terms. Hence, Eq. (12.37) holds at time $t+1$ in the occurrence of a *negative* price jump between time t and time $t+1$.

References

Bernanke, B. S. and A. Blinder (1992): "The Federal Funds Rate and the Channels of Monetary Transmission." *American Economic Review* 82, 901-921.

Black, F. and M. Scholes (1973): "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81, 637-659.

Black, F., E. Derman and W. Toy (1990): "A One Factor Model of Interest Rates and its Application to Treasury Bond Options." *Financial Analysts Journal* (January-February), 33-39.

Cox, J. C., S. A. Ross and M. Rubinstein (1979): "Option Pricing: A Simplified Approach." *Journal of Financial Economics* 7, 229-263.

Derman, E. and J. Kani (1994): "Riding on a Smile." *Risk* 7, 32-39.

Dupire, B. (1994): "Pricing with a Smile." *Risk* 7, 18-20.

Heath, D., R. Jarrow and A. Morton (1992): "Bond Pricing and the Term-Structure of Interest Rates: a New Methodology for Contingent Claim Valuation." *Econometrica* 60, 77-105.

Ho, T. S. Y. and S.-B. Lee (1986): "Term Structure Movements and the Pricing of Interest Rate Contingent Claims." *Journal of Finance* 41, 1011-1029.

Ho, T. S. Y. and S.-B. Lee (2004): *The Oxford Guide to Financial Modeling*. Oxford University Press.

Hull, J. C. (2003): *Options, Futures, and Other Derivatives*. Prentice Hall. 5th edition (International Edition).

Hull, J. C. and A. White (1990): "Pricing Interest Rate Derivative Securities." *Review of Financial Studies* 3, 573-592.

McCulloch, J. (1971): "Measuring the Term Structure of Interest Rates." *Journal of Business* 44, 19-31.

McCulloch, J. (1975): "The Tax-Adjusted Yield Curve." *Journal of Finance* 30, 811-830.

Nelson, C.R. and A.F. Siegel (1987): "Parsimonious Modeling of Yield Curves." *Journal of Business* 60, 473-489.

Rubinstein, M. (1994): "Implied Binomial Trees." *Journal of Finance* 49, 771-818.

Tuckman, B. (2002): *Fixed Income Securities*. Wiley Finance.

Vasicek, O. (1977): "An Equilibrium Characterization of the Term Structure." *Journal of Financial Economics* 5, 177-188.

13
Interest rates

13.1 Introduction

This chapter surveys empirical facts and models regarding the term-structure of interest rates and derivatives based thereon. It differs from the previous introductory chapter, as we now largely rely on continuous-time methods while providing a systematic approach to a variety of important topics. These topics range from the stylized facts such as the factors driving the yield curve, their business cycle components, or bond returns predictability and volatility, to more conceptual aspects regarding how we would need to think about duration in a random environment, or the pricing details of interest rate derivatives such as bond options, puttable and callable bonds, swaps, caps, floors, or swaptions, to mention a few.

We know from previous chapters that an important objective arising while pricing derivatives is that we make sure that the price of the underlying assets is pinned down without errors. When it comes to interest rate derivatives, this task is more challenging, because the yield curve relies on risks that are typically not traded. Consider, for example, a model where the price of a zero coupon bond is only driven by random movements of the short-term rate—a one-factor model. Let $P(r, \tau, T)$ be the price at time τ of a zero coupon bond expiring at time T, when the short-term rate is r. The exact functional form of the pricing function $P(r, \tau, T)$ depends on the assumed dynamics of the short-term rate and the market risk-appetite. Models of this kind, and generalizations to multi-state variables, are known as "models of the short-term rate," and are discussed in Section 13.4.

Models of the short-term rate are very important because once they are made complex enough to address the main facts we see in the data, they might perform a series of tasks. For example, they can be used to forecast developments in fixed income markets. They could also be used for trading purposes should they reasonably point to market inefficiencies. Note that these models lead to pricing errors, and it is actually the presence of these errors to justify their potential use for trading purposes.

A second class of models that does not lead to pricing errors is that developed by Heath, Jarrow and Morton (1992), which generalizes the Ho and Lee (1986) model described in the previous chapter. In the previous chapter, we have seen tree-based instances of additional "no-arbitrage models." This chapter provides a systematic continuous time treatment (in Sections

13.6 and 13.6). A principle underlying these models is that current bond prices need not be modeled in the first place. Rather, current bond prices are taken as primitives, with the modeling focus being shifted to the ongoing development of forward rates, i.e. interest rates prevailing today for borrowing in the future. There is a relation linking bond prices to forward rates. No arbitrage then restricts the joint behavior of future bond prices and forward rates. We shall emphasize the use of these models to price derivatives.

In more detail, the plan of this chapter is as follows. The next section provides definitions of interest rates and markets, and foundational issues regarding two basic representations of bond prices—one in terms of the short-term rate and another in terms of forward rates. Section 13.3 contains an introduction to a number of very important empirical topics, such as bond return predictability, or the relation between the yield curve and the business cycle. Sections 13.4 and 13.5 deal with models of the short-term rate, and Section 13.6 with their perfectly fitting extensions, i.e. the extensions that make these models fit the initial yield curve without errors. Section 13.7 contains a treatment of models that fit the yield curve, based on the Heath, Jarrow and Morton (1992) framework. Section 13.8 is an introduction to the main interest rate derivatives, and describes how these assets can be priced relying on the models of the short-term rate (and their perfectly fitting extensions). Section 13.9 provides an alternative pricing framework, known as "market model," whereby derivatives are evaluated through Black's (1976) "pricers." A number of appendixes provides technical details omitted from the main text. This chapter does not assume default risk except in parts of Section 13.4. Default risk is, instead, systematically dealt with in the next chapter.

13.2 Bond prices and interest rates

13.2.1 A first representation of bond prices

13.2.1.1 Zeros

Let $P(\tau, T)$ the price at τ of a zero coupon bond expiring at T, and consider the discretely compounded interest rate L for the time interval $[\tau, T]$ introduced in the previous chapter (see Section 12.2), and defined as

$$P(\tau, T) = \frac{1}{1 + (T - \tau) L(\tau, T)}. \tag{13.1}$$

Given $L(\tau, T)$, the *short-term rate* process r is obtained as:

$$r_\tau \equiv \lim_{T \downarrow \tau} L(\tau, T).$$

Next, let Q be a risk-neutral probability, and $\mathbb{E}_\tau(\cdot)$ denote the time τ conditional expectation under Q. By the FTAP, there is no arbitrage if and only if $P(\tau, T)$ satisfies, for all $\tau \in [t, T]$,

$$P(\tau, T) = \mathbb{E}_\tau \left(e^{-\int_\tau^T r_s ds} \right). \tag{13.2}$$

The proof of Eq. (13.2) relies on arguments that are now quite standard in these lectures, but its "if part" is provided again in Appendix 1, because it highlights a few key hedging arguments that underlie it.

13.2.1.2 Fixed coupon and floating rate bonds

Given a set of dates $\{T_i\}_{i=0}^n$, a fixed coupon bond pays off a known *coupon* stream, i.e. c_i at T_i, $i = 1, \cdots, n$, and £1 at T_n; typically, the coupon paid at at T_i compensates for the time-interval $T_i - T_{i-1}$. By the FTAP, the value of a fixed coupon bond is

$$P_{\text{fcb}}(t, T_n) = P(t, T_n) + \sum_{i=1}^{n} c_i P(t, T_i).$$

A floating rate bond pays off coupons indexed to short-term rates,

$$c_i = \delta_{i-1} L(T_{i-1}, T_i) = \frac{1}{P(T_{i-1}, T_i)} - 1, \tag{13.3}$$

where $\delta_i \equiv T_{i+1} - T_i$, and where the second equality follows by Eq. (13.1). By the FTAP, the price p_{frb} as of time t of a floating rate bond is:

$$\begin{aligned} p_{\text{frb}}(t) &= P(t, T_n) + \sum_{i=1}^{n} \mathbb{E}_t \left(e^{-\int_t^{T_i} r_\tau d\tau} \delta_{i-1} L(T_{i-1}, T_i) \right) \\ &= P(t, T_n) + \sum_{i=1}^{n} \mathbb{E}_t \left(\frac{e^{-\int_t^{T_i} r_\tau d\tau}}{P(T_{i-1}, T_i)} \right) - \sum_{i=1}^{n} P(t, T_i) \\ &= P(t, T_n) + \sum_{i=1}^{n} P(t, T_{i-1}) - \sum_{i=1}^{n} P(t, T_i) \\ &= P(t, T_0). \end{aligned} \tag{13.4}$$

where the second line follows by Eq. (13.3) and the third line will be justified in a moment (see Eq. (13.5) below). That is, a floating rate bond would quote at par at its first reset date, $p_{\text{frb}}(T_0) = P(T_0, T_0) = 100\%$.[1]

Regarding the third line of Eq. (13.4),

$$P(t, T) = \mathbb{E}_t \left(\frac{e^{-\int_t^S r_\tau d\tau}}{P(T, S)} \right), \tag{13.5}$$

consider the following economic interpretation. Suppose that at time t, $£P(t, T)$ are invested in a bond maturing at time T. At time T, this investment will obviously pay off $£1$. And at time T, $£1$ can be further rolled over another bond maturing at time S, thus yielding $£\ 1/\ P(T, S)$ at time S. In other words, an investment at t equal to $£P(t, T)$, leads to a "payoff" at S equal to $£\ 1/\ P(T, S)$, whence Eq. (13.5).[2]

[1] This property also holds in a market where the floating rates continuously pay off the instantaneous short-term rate r. Indeed, let $T_0 = t$, and let p_{frb} is solution to the partial differential equation (13.94) in Section 13.6, with $\pi(r) = r$, and boundary condition $p_{\text{frb}}(T) = 1$. Then, it can be verified that $p_{\text{frb}} = 1$ is solution to Eq. (13.94).

[2] Mathematically, we have, by the Law of Iterated Expectations, that

$$\mathbb{E}_t \left[\frac{e^{-\int_t^S r(\tau) d\tau}}{P(T, S)} \right] = \mathbb{E}_t \left[\mathbb{E} \left(\frac{e^{-\int_t^T r(\tau) d\tau} e^{-\int_T^S r(\tau) d\tau}}{P(T, S)} \middle| \mathcal{F}(T) \right) \right] = P(t, T).$$

13.2.2 Forward rates

Forward rates are interest rates that make the value of a *forward rate agreement* (FRA, henceforth) equal to zero at origination. The previous chapter provides the definition of a forward rate agreement (see Section 12.2), which we re-state below for reasons clarified in a moment. Forward rates as of time t, regarding a FRA for the future time-interval $[T, S]$, are denoted with $F(t, T, S)$, and link to bond prices through the following relation, derived in Chapter 12:

$$\frac{P(t,T)}{P(t,S)} = 1 + (S-T)\,F(t,T,S)\,. \tag{13.6}$$

Clearly, the forward rate agreed at T for the time interval $[T, S]$ is the short-term rate applying to the same period:

$$F(T,T,S) = L(T,S)\,. \tag{13.7}$$

Next, consider a more general FRA, whereby a party agrees: (i) to pay a counterparty an interest rate on a given principal at time T, fixed at some $K \neq F(t,T,S)$, and (ii) to receive, in exchange, the future interest rate prevailing at time T for the time interval $[T, S]$, $L(T,S)$. The time S payoff originated by this forward starting "interest rate swap" is:

$$\pi_{\text{fra}}(T,S) \equiv (S-T)\,[L(T,S) - K]\,. \tag{13.8}$$

This payoff is the same as the P&L to a party who enters a FRA at time t (with no costs) for the time-interval $[T, S]$, as a future borrower. Come time T, the party shall honour the FRA by borrowing £1 for the time-interval $[T, S]$ at a cost of K. At the same time, the party can lend this very same £1 at the random interest rate $L(T,S)$. The time S payoff deriving from this trade is, of course, the same as $\pi_{\text{fra}}(T,S)$ in Eq. (13.8). In Section 13.8, we shall deal with the evaluation of this and related payoffs.

13.2.3 A second representation of bond prices

13.2.3.1 Prices as forward looking indicators

Bond prices can be expressed in terms of these forward interest rates, namely in terms of the "instantaneous" forward rates. First, rearrange terms in Eq. (13.6) so as to obtain:

$$F(t,T,S) = -\frac{P(t,S) - P(t,T)}{(S-T)P(t,S)}.$$

The *instantaneous forward rate* $f(t,T)$ is defined as

$$f(t,T) \equiv \lim_{S \downarrow T} F(t,T,S) = -\frac{\partial \ln P(t,T)}{\partial T}. \tag{13.9}$$

It can be interpreted as the marginal rate of return from committing a bond investment for an additional instant. To express bond prices in terms of f, integrate Eq. (13.9), $f(t,\ell) = -\frac{\partial \ln P(t,\ell)}{\partial \ell}$, with respect to the maturity date ℓ, use the condition that $P(t,t) = 1$, and obtain:

$$P(t,T) = e^{-\int_t^T f(t,\ell)\,d\ell}. \tag{13.10}$$

Eq. (13.10) suggests a natural modeling approach of the yield curve that emphasizes the dynamics of the forward rates, and dealt in detail in Sections 13.5 and 13.6.

13.2.3.2 The marginal nature of forward rates

Consider the *yield-to-maturity* introduced in Chapter 12 (see Section 12.2), defined to be the function $R(t,T)$ such that

$$P(t,T) \equiv e^{-(T-t)\cdot R(t,T)}. \tag{13.11}$$

Comparing Eq. (13.11) with Eq. (13.10) yields:

$$R(t,T) = \frac{1}{T-t}\int_t^T f(t,\tau)d\tau. \tag{13.12}$$

By differentiating Eq. (13.12) with respect to T yields:

$$\frac{\partial R(t,T)}{\partial T} = \frac{1}{T-t}\left[f(t,T) - R(t,T)\right]. \tag{13.13}$$

Eq. (13.13) underscores the "marginal nature" of forward rates: the yield-curve, $R(t,T)$, is increasing in, decreasing in, or stationary at T, according to whether $f(t,T)$ exceeds, is lower, or equal the spot rate for maturity T. It is a simple but powerful result: it is intuitive that forward rates should summarize information about the market expectations regarding future interest rates, as explained in Section 13.3. For example, an increasing yield curve would signal high rates to be expected in the future, under conditions explained in Section 13.3.

13.3 Stylized facts

13.3.1 The expectation hypothesis

The *expectation hypothesis* holds that the *forward rates equal expected future short-term rates*, viz

$$f(t,T) = E_t(r_T), \tag{13.14}$$

where $E_t(\cdot)$ denotes expectation under the physical probability. Eq. (13.14) has two implications. First, through Eq. (13.12),

$$R(t,T) = \frac{1}{T-t}\int_t^T E_t(r_\tau)\,d\tau, \tag{13.15}$$

and second, by Eq. (13.13),

$$\frac{\partial R(t,T)}{\partial T} = \frac{1}{T-t}\left[E_t(r_T) - R(t,T)\right]. \tag{13.16}$$

Note a very straight conclusion we can make from Eq. (13.16): short-term rates are expected to be high at some future date T (compared to the corresponding yield) if, and only if, the yield curve is increasing in T. One implication is that an increasing yield curve (i.e. one where r_t is always less than $R(t,T)$) signals the expectation that future short-term rates will increase and, similarly, an inverted yield curve signals the expectation that future short-term rates will fall.

For example, in December 2012, the yield curve in the United Kingdom was substantially flat at less than 0.5% for maturities of less than three years, and then steadily increased to 2% for maturities up to ten years. Taken at face value, the expectation hypothesis would simply imply

that in December 2012, the market did not expect UK short-term to move over the following three years. Naturally, the market reasoning could have changed over new information.

A natural question arises as to whether in the data the forward rate for maturity T is higher than the short-term rate expected to prevail at time T, viz

$$f(t,T) \geq E_t\left(r_T\right). \tag{13.17}$$

It is an old issue. One possibility is that (13.17) holds due to risk-aversion. Consider the Hicks-Keynesian *normal backwardation* hypothesis reviewed in Chapter 11.[3] In the context of this chapter, forward rates can be higher than expected spot rates as firms demand long-term funds with fund suppliers preferring to lend at shorter maturity dates. In this case, the market is cleared by intermediaries requiring a *liquidity premium* to be compensated for their risky activity of borrowing at short and lending at long maturities—intermediaries worry about their profits at time T, thereby lending at higher rates than the expected spots.[4]

Finally, consider the following definition of the *term-premium*, the difference between the spot rate and the future expected average short-term rate, for the same horizon, as follows:

$$\mathrm{TP}\left(t,T\right) \equiv R\left(t,T\right) - \frac{1}{T-t}\int_t^T E_t\left(r_\tau\right) d\tau = \frac{1}{T-t}\int_t^T \left[f\left(t,\tau\right) - E_t\left(r_\tau\right)\right] d\tau,$$

where the second equality follows by Eq. (13.12). We see that the expectation hypothesis, and possible violations of it, bring content to the term-premium. The next section aims to provide evidence on the expectation hypothesis, as linked to the predictability of bond returns.

13.3.2 Bond returns predictability

What does the empirical evidence suggest about the expectation hypothesis? How does the expectation hypothesis link to bond returns? First, we explain under which probability should the expectation hypothesis hold. Second, we explain the main issues regarding the empirical evidence.

13.3.2.1 Forward-adjusted expectation hypothesis

In more advanced sections of this chapter (see Section 13.9), we shall explain that forward rates cannot be unbiased expectations of the future spot rates, not even under the risk-neutral probability. They only are such, under a certain probability, known as forward probability, already introduced in Chapter 5. Section 13.8, for example, shows that

$$f\left(t,T\right) = \mathbb{E}_{Q_F^T}\left(r_T\right), \tag{13.18}$$

where $\mathbb{E}_{Q_F^T}$ denotes the expectation under the forward probability, Q_T^F (see Eq. (13.97)). This result points to a basic difficulty with the expectation hypothesis: Eq. (13.18) just tells us that the expectation hypothesis can only hold true under the forward probability. There are two

[3]Note that the *normal backwardation* (*contango*) hypothesis states that forward prices are lower (higher) than future expected spot prices, which implies the inequality in (13.17), stated in terms of rates.

[4]The inequality in (13.17) cannot hold in a purely risk-neutral market. Indeed, we have that, $e^{-\int_t^T f(t,\ell)d\ell} \equiv P\left(t,T\right) \geq e^{-\int_t^T \mathbb{E}_t(r_\tau)d\tau}$, which implies that $\int_t^T \left(f(t,\tau) - \mathbb{E}_t\left(r_\tau\right)\right) d\tau \leq 0$ for eact T and therefore $f(t,\tau) \leq \mathbb{E}_t\left(r_\tau\right)$ for each τ, contradicting (13.17).

sources of risk, which might actually prevent the expectation theory from holding, in practice. To illustrate, let us elaborate on Eq. (13.18),

$$\begin{aligned} f(t,T) &= \mathbb{E}_{Q_F^T}(r_T) \\ &= \mathbb{E}_t(\eta_T(T)\, r_T) \\ &= \mathbb{E}_t(r_T) + cov_{Q,t}(\eta_T(T), r_T) \\ &= E_t(r_T) + cov_{P,t}\left(\eta_{P,T}(T), r_T\right) + cov_{Q,t}(\eta_T(T), r_T), \end{aligned} \tag{13.19}$$

where $\eta_{P,T}(T)$ denotes the Radon-Nikodym derivative of the physical probability against the risk-neutral; $\mathbb{E}_t$ and $cov_{Q,t}$ denote time t conditional expectation and covariance under the risk-neutral probability, with E_t and $cov_{P,t}$ denoting the counterparts under the physical probability. That is, forward rates might deviate from future expected short-term rates, because of risk-aversion corrections (the second term in the last equality) and randomness in interest rates (the third term in the last equality).

Note that we can determine these covariance terms through the Radon-Nikodym theorem. Show that the second covariance $cov_{Q,t}(\eta_T(T), r_T)$ is zero, when T goes to t. Then, as T goes to t, we are only left with risk-aversion corrections, the first covariance $cov_{P,t}\left(\eta_{P,T}(T), r_T\right)$. The *local expectation hypothesis* postulates that this second correction is zero.]

[In progress]

13.3.2.2 Organizing the empirical evidence

Denote the continuously compounded returns on a zero expiring at some date T as $r_{t+1}^T = \ln \frac{P(t+1,T)}{P(t,T)}$. Using the definition of spot rates, $R(t,T)$, the excess returns, $\hat{r}_{t+1}^T$ say, can be expressed as:

$$\begin{aligned} \hat{r}_{t+1}^T &\equiv \ln \frac{P(t+1,T)}{P(t,T)} - \ln \frac{1}{P(t,t+1)} \\ &= \ln \frac{P(t+1,T)}{P(t,T)} - R(t,t+1) \\ &= -(T-t-1)\, R(t+1,T) + (T-t)\, R(t,T) - R(t,t+1), \end{aligned}$$

such that the expected change in the yield curve is negatively related to the expected excess returns and positively related to the slope of the yield curve:

$$E_t[R(t+1,T) - R(t,T)] = -\frac{1}{T-t-1} E_t\left(\hat{r}_{t+1}^T\right) + \frac{1}{T-t-1}[R(t,T) - R(t,t+1)].$$

The expectation hypothesis implies that the risk-premium, $E_t\left(\hat{r}_{t+1}^T\right)$, is zero. Indeed, assuming that the expectation hypothesis holds,

$$\begin{aligned} E_t(r(t+1)) &= f(t,t+1) \\ &= E_t(r(t+1)) + cov_{P,t}\left(\eta_{P,t+1}(t+1), r(t+1)\right) + cov_{Q,t}\left(\eta_{t+1}(t+1), r(t+1)\right), \end{aligned}$$

where the first equality holds by the expectation hypothesis, and the second is Eq. (13.19). Therefore, the sum of the last two terms in the last equality is zero, implying that $E_t\left(\hat{r}_{t+1}^T\right)$ is zero. Veronesi (2010, Chapter 7) builds up a parametric example to illustrate these relations, within an affine model. In the Appendix, we illustrate how these relations work, analytically, by hinging upon a simple but famous model—that of Vasicek (1977). [In progress]

Empirically, we can test for the expectation theory, by running the following regression:

$$R(t+1,T) - R(t,T) = \alpha_T + \beta_T \frac{1}{T-t-1}\left[R(t,T) - R(t,t+1)\right] + \text{Residual}_t,$$

and test for the null: $\alpha_T = 0$ and $\beta_T = 1$. A widely known empirical feature of US data is that the estimates of β_T are typically negative for all maturities T, and somewhat increasing with T in absolute value. In fact, Fama and Bliss (1987) show that bond returns are predictable, in that the risk-premium $E_t\left(\hat{r}_{t+1}^T\right)$ relates to the forward spreads, defined as $f_t^T - R(t,t+1)$, and that regressing,

$$\hat{r}_{t+1}^T = \alpha_T + \beta_T\left(f_t^T - R(t,t+1)\right) + \text{Residual}_t,$$

delivers statistically significant and positive values of β_T for many maturities T.

Cochrane and Piazzesi (2005) go one step further and consider the following regressions:

$$\hat{r}_{t+1}^T = \alpha_T + \beta_{1T} R(t,t+1) + \sum_{j=2}^{5} \beta_{j,T} f_t^j + \text{Residual}_t,$$

where f_t^T is the forward rate for maturity $T-1$, $f_t^T = -\ln \frac{P(t,T)}{P(t,T-1)}$. They document a "tent shape" for the estimates of the coefficients $\left(\beta_{j,T}\right)_{j=1}^5$, for bond maturities $T \in \{1,\cdots,5\}$, and where t is in years so as to make returns calculated on a yearly basis. They document that this tent shape is robust to estimating a factor model, in that the shape persists in the estimates of the coefficients $(b_j)_{j=1}^5$ in:

$$\hat{r}_{t+1}^T = \alpha_T + \beta_{1T} Z_t + \text{Residual}_t, \quad Z_t = b_1 R(t,t+1) + \sum_{j=2}^{5} b_j f_t^j,$$

where Z_t is the common factor among the bond maturities $T \in \{1,\cdots,5\}$. Moreover, they argue that the predicting power of their factors is not destroyed, in sample, while conditioning on the standard known to explain movements in the yield curve (see Section 12.3).

13.3.3 The yield curve and the business cycle

There is a simple prediction about the shape the yield-curve that we can make. By Jensen's inequality, $e^{-(T-t)R(t,T)} \equiv P(t,T) = \mathbb{E}_t[e^{-\int_t^T r_\tau d\tau}] \geq e^{-\int_t^T \mathbb{E}_t(r_\tau)d\tau}$. Therefore, the yield curve satisfies: $R(t,T) \leq \frac{1}{T-t}\int_t^T \mathbb{E}_t\left(r_\tau\right) d\tau$. For example, suppose that the short-term rate is a martingale under the risk-neutral probability, viz $\mathbb{E}_t\left(r_\tau\right) = r_t$. Then, the yield curve is bound to be: $R(t,T) \leq r_t$. That is, the yield curve is not increasing in time-to-maturity, T, at least for small maturities. Positively sloped yield curve, then, likely arise because the short-term rate is *not* a martingale under the risk-neutral probability, which happens because of two fundamental, and not necessarily mutually exclusive, reasons: (i) interest rates are expected to increase, (ii) investors are risk-averse. On average, the US yield curve is upward sloping at maturity from one up to ten years.

There is strong empirical evidence since at least Kessel (1965) or, later, Laurent (1988, 1989), Stock and Watson (1989), Estrella and Hardouvelis (1991) and Harvey (1991, 1993), that inverted yield curves predict recessions with a lead time of about one to two years. Figure 13.1 illustrates these empirical facts through a plot of the difference between long-term and short-term yields on Treasuries—in short, the "term spread."

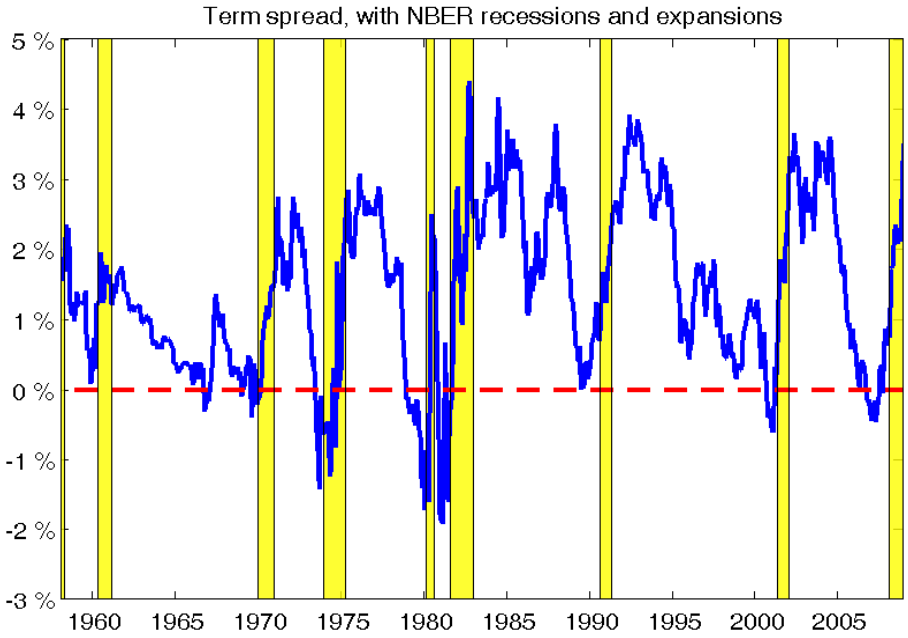


FIGURE 13.1. This picture depicts the time series of the term spread, defined as the difference between the 10 year yield minus the 3 month yield on US Treasuries. Sample data cover the period from January 1957 to December 2008. The shaded areas mark recession periods, as defined by the National Bureau of Economic Research. The end of the last recession was announced to have occurred in June 2009.

Naturally, there are recession episodes preceded by mild yield curve inversions. But the really striking empirical regularity is the sharp movements of the term spread towards a negative territory, occurring prior to any recession episode. Note, it is not really important whether it is the short-term rate that goes up or the long-term rate that goes down. The empirical regularity is that the *term spread* goes down and becomes negative prior to a recession. The explanations of these statistical facts are challenging, and might hinge upon both (i) the conduct of monetary policy and the expectations about it, and (ii) the risk-premiums agents require to invest in long-term bonds. We discuss these two points below.

(i) *The monetary channel*:

(i.1) During expansions, monetary policy tends to be restrictive, to prevent the economy from heating up. At the height of an expansion, then, short-term yields go up.

(i.2) Moreover, during recessions, monetary policy tends to keep interest rates low. At the height of an expansion, agents may be anticipating an incoming recession and expecting central banks to lower future interest rates. Therefore, at the height of an expansion, future interest rates may be expected to lower. The expectation hypothesis in Eq. (13.15) would then predict that the slope of the yield curve would decrease. Granted, in the previous subsection, we have just learnt that the expectation hypothesis does not hold empirically. Bond markets command risk-premiums. However, a risk-premium channel would reinforce the conclusion that the slope of the yield curve decreases during expansions, as argued in the next point.

(ii) *The risk-premium channel*: From Chapter 8, we know that risk-premiums are countercyclical, being high during recessions and low during expansions. The conditional equity premium is countercyclical, and so is the long-bond premium.[5] In fact, long-term yields and equity expected returns are likely to be driven by the same state variables affecting the pricing kernel of the economy.[6]

Let's summarize. On the one hand, countercyclical monetary policy may be responsible of the negative price changes affecting short-term bonds. On the other, expectations about countercyclical monetary policy as well as procyclical risk-appetite may be responsible for positive price changes affecting long-term bonds. These price movements, we have argued, should occur at the height of an expansion. But the sample data we have are those where expansions are followed by recessions. Whence, the statistical facts regarding the predictive content of the yield curve. In Section 13.4, we rely on a simple model to formalize this reasoning.[7]

The empirical evidence that links yield curve inversions to recessions is quite important, but there are a few caveats. First, there may be other explanations for this evidence, relying on the presence of financial intermediaries. [Cite the appropriate reference]. Indeed, it is interesting to note that these inversions did also use to occur prior to the creation of the Federal Reserve system. The creation of the US Central Bank can be a "Natural Experiment" to rely on and perform statistical inference about the importance of the gaming between central banks and market expectations regarding the future conduct of monetary policy.

Second, inverted yield curve may occurr for resons unrelated to the business cycle—in this case, some of the empirical evidence gathered so far might just have arisen by chance. For example, the inversion of the yield curve in 2006 is widely considered to have occurred due to a strong demand for long-term bonds, as warned by some policy-makers at the time.[8] This demand pressure was exerted by institutional investors such as pension funds performing asset-liability management duties. It is very difficult to quantify this demand pressure. On the other hand, since 2006, the Federal Reserve had started to target higher interest rates to deal with inflationary pressures, inflationary pressures that at least partially originated from a quite expansionary policy following the 2001 recession, the Twin Towers attacks and, perhaps, the 2003 corporate scandals too. It is an open question as to whether the markets thought that this increased tightening was marking the end of an expansion, thereby feeding an expectation future interest rates would drop again in the near future. Undeniably, the sharp tightening of the FED policy at the time would carry implications on financial and economic developments such as the 2007 subprime crisis and the crisis arising therefrom as explained in the next chapter.

13.3.4 Additional stylized facts about the US yield curve

There are three additional features of data, which need to be noted.

[5] An objection to this line of reasoning is that countercyclical risk-premiums might lead to expect future bond prices to decrease over a future recession, thereby destroying the effects of a procyclical short-term rate. In Section 13.3.3.2, we develop a model where these effects do not arise as soon as the effects of countercyclical risk-premiums are assumed to be bounded.

[6] That long term bonds and stock market are informally acknowledged to be tightly related is witnessed by a quite raw rule of thumb, whereby a stock market correction, such as a crash say, is deemed to be imminent when the spread 30 year bond yield minus the earning-price ratio is larger than 3%. This spread, which is usually around 1% or 2%—and on average, zero, once corrected for inflation—was indeed larger than 3% in 1987 and in 1997.

[7] There may be other channels. Inverted yield curves lead to negative margins for banks, which might then contribute to a credit crunch, determined by a less aggressive attitude to lend due to the negative margins. This might depress demand, leading to the expectation of an imminent recession. This expectation leads to a negative term spread due to the mechanism analyzed in the main text, over a vicious cycle.

[8] See, e.g., the European Central Bank Monthly Bulletin, February 2006, p. 27.

(i) Yields are highly correlated (say three year yields with four year yields, with five year yields, etc.), and suggest the existence of common factors driving all of them, discussed in Section 13.3.5 below.

(ii) Yields are also highly persistent, and this persistence bears important consequences on derivative pricing, as explained in Section 13.8.1.

(iii) The term-structure of unconditional volatility is downward sloping, a feature also rationalized in Section 13.8.1.

13.3.5 Common factors affecting the yield curve

Which systematic risks affect the whole term-structure of interest rates? How many factors are needed to explain the variation of the yield curve? Standard "duration hedging" practice relies on the idea that most of the variation in the yield curve is successfully captured by a single factor that produces parallel shifts in the yield curve (see Chapter 11). How reliable is this idea in practice? This section reviews famous evidence that most of the variation of the US yield curve is explained by just a "few factors," interpreted as (i) a level factor, (iii) a steepness factor, and (iii) a curvature factor. It is natural to expect that only a few factors explain the yield curve: intuitively, there is no so much difference between a 10Y zero and a 10Y+1day zero, and then, 10Y+2days, etc. Repeating this reasoning ad infinitum suggests that most of the yield curve should be likely driven by common sources of variation.

In a very famous study, Litterman and Scheinkman (1991) demonstrate that most of the variation (more than 95%) of the yield curve can be attributed to the variation of three unobservable factors, which they label (i) a "level" factor, (ii) a "steepness" (or "slope") factor, and (iii) a "curvature" factor. Their analysis relies on a *fixed-factor* model. Succinctly, this methodology can be described as follows.

Suppose that p returns computed from bond prices at p different maturities are generated by a linear factor model, with a fixed number k of factors,

$$\underset{p\times 1}{R_t} = \underset{p\times 1}{\bar{R}} + \underset{p\times k}{B}\,\underset{k\times 1}{F_t} + \underset{p\times 1}{\epsilon_t}\,, \tag{13.20}$$

where R_t is the vector of returns, F_t is the zero-mean vector of common factors affecting the returns, assumed to be zero mean, $\bar{R}$ is the vector of unconditional expected returns, ϵ_t is a vector of idiosyncratic components of the return generating process, and B is a matrix containing the factor loadings. Each row of B contains the factor loadings for all the common factors affecting a given return, i.e. the sensitivities of a given return with respect to a change of the factors. Each comumn of B contains the *term-structure of factor loadings*, i.e. how a change of a given factor affects the term-structure of excess returns.

13.3.5.1 Methodological details

Estimating the model in Eq. (13.20) leads to econometric challenges, mainly because the vector of factors F_t is unobservable.[9] However, there exists a simple method, known as *principal*

[9]Suppose that in Eq. (13.20), $F \sim N(0, I)$, and that $\epsilon \sim N(0, \Psi)$, where Ψ is diagonal. Then, $R \sim N\left(\bar{R}, \Sigma\right)$, where $\Sigma = BB^\top + \Psi$. The assumptions that $F \sim N(0, I)$ and that Ψ is diagonal are necessary to identify the model, but not sufficient. Indeed, any orthogonal rotation of the factors yields a new set of factors which also satisfies Eq. (13.20). Precisely, let T be an orthonormal matrix. Then, $(BT)(BT)^\top = BTT^\top B^\top = BB^\top$. Hence, the factor loadings B and BT have the same ability to generate the matrix Σ. To obtain a unique solution, one needs to impose extra constraints on B. For example, Jöreskog (1967) develop a maximum

components analysis (PCA, henceforth), which leads to empirical results qualitatively similar to those holding for the general model in Eq. (13.20). We discuss these empirical results in the next subsection. We now describe the main methodological issues arising within PCA.

The main idea underlying PCA is to transform the original p correlated variables R into a set of new uncorrelated variables, the *principal components*. These principal components are linear combinations of the original variables, and are arranged in order of decreased importance: the first principal component accounts for as much as possible of the variation in the original data, etc. Mathematically, we are looking for p linear combinations of the demeaned excess returns,

$$Y_i = C_i^\top \left(R - \bar{R}\right), \qquad i = 1, \cdots, p, \tag{13.21}$$

such that, for p vectors $C_i^\top$ of dimension $1 \times p$, (i) the new variables Y_i are uncorrelated, and (ii) their variances are arranged in decreasing order. The logic behind PCA is to ascertain whether a few components of $Y = [Y_1 \cdots Y_p]^\top$ account for the bulk of variability of the original data. Let $C^\top = [C_1^\top \cdots C_p^\top]$ be a $p \times p$ matrix such that we can write Eq. (13.21) in matrix format, $Y_t = C^\top \left(R_t - \bar{R}\right)$ or, by inverting,

$$R_t - \bar{R} = C^{\top -1} Y_t. \tag{13.22}$$

Next, suppose that the vector $Y^{(k)} = [Y_1 \cdots Y_k]^\top$ accounts for most of the variability in the original data,[10] and let $C^{\top(k)}$ denote a $p \times k$ matrix extracted from the matrix $C^{\top -1}$ through the first k rows of $C^{\top -1}$. Since the components of $Y^{(k)}$ are uncorrelated and they are deemed largely responsible for the variability of the original data, it is natural to "disregard" the last $p - k$ components of Y in Eq. (13.22),

$$\underset{p\times 1}{R_t - \bar{R}} \;\approx\; \underset{p\times k}{C^{\top(k)}}\; \underset{k\times 1}{Y_t^{(k)}}.$$

If the vector $Y_t^{(k)}$ really accounts for most of the movements of R_t, the previous approximation to Eq. (13.22) should be fairly good.

Let us make more precise what the concept of variability is in the context of PCA. Suppose that the variance-covariance matrix of the returns, Σ, has p distinct eigenvalues, ordered from the highest to the lowest, as follows: $\lambda_1 > \cdots > \lambda_p$. Then, the vector C_i in Eq. (13.21) is the eigenvector corresponding to the i-th eigenvalue. Moreover,

$$var\left(Y_i\right) = \lambda_i, \qquad i = 1, \cdots, p.$$

Finally, we have that

$$\mathrm{R_{PCA}} = \frac{\sum_{i=1}^{k} var\left(Y_i\right)}{\sum_{i=1}^{p} var\left(R_i\right)} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}. \tag{13.23}$$

(Appendix 4 provides technical details and proofs regarding the previous formulae.) It is in the sense of Eq. (13.23) that we say that the first k principal components account for $\mathrm{R_{PCA}}\%$ of the total variation of the data.

---

likelihood approach in which the log-likelihood function is, $-\frac{1}{2}N\left[\ln|\Sigma| + \mathrm{Tr}\left(S\Sigma^{-1}\right)\right]$, where S is the sample covariance matrix of R, and the constraint is that $B^\top \Psi B$ be diagonal *with elements arranged in descending order*. The algorithm is: (i) for a given Ψ, maximize the log-likelihood with respect to B, under the constraint that $B^\top \Psi B$ be diagonal with elements arranged in descending order, thereby obtaining $\hat{B}$; (ii) given $\hat{B}$, maximize the log-likelihood with respect to Ψ, thereby obtaining $\hat{\Psi}$, which is fed back into step (i), etc. Knez, Litterman and Scheinkman (1994) describe this approach in their paper. Note that the identification device they describe at p. 1869 (Step 3) roughly corresponds to the requirement that $B^\top \Psi B$ be diagonal *with elements arranged in descending order*. Such a constraint is clearly related to principal component analysis.

[10] There are no rigorous criteria to say what "most of the variability" means in this context. Instead, a likelihood-ratio test is most informative in the context of the estimation of Eq. (13.20) by means of the methods explained in the previous footnote.

13.3.5.2 The empirical evidence

The striking feature of the empirical results uncovered by Litterman and Scheinkman (1991) is that they have been confirmed to hold across a number of countries and sample periods. Moreover, the economic nature of these results is the same, independently of whether the statistical analysis relies on a rigorous factor analysis of the model in Eq. (13.20), or a more back-of-envelope computation based on PCA. Finally, the empirical results that hold for bond returns are qualitatively similar to those that hold for bond yields.

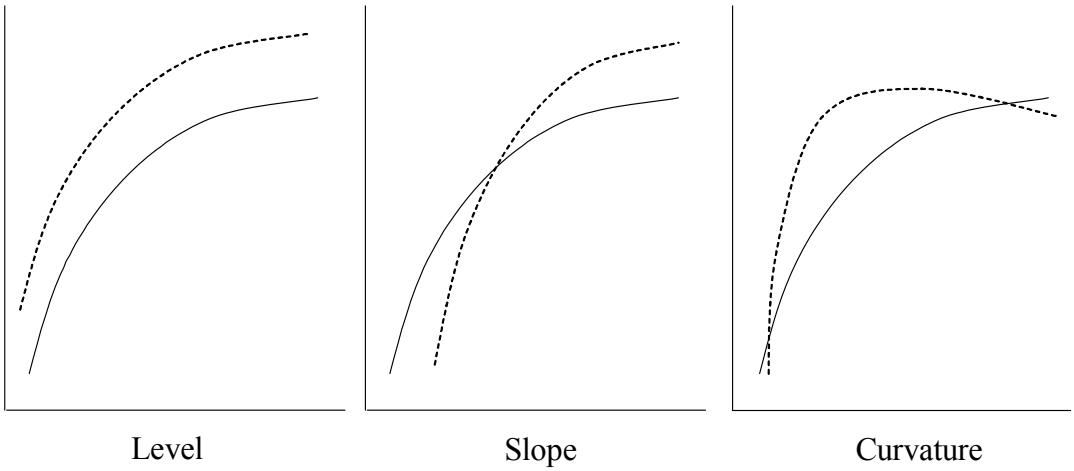


FIGURE 13.2. Changes in the term-structure of interest rates generated by changes in the "level," "slope" and "curvature" factors.

Figure 13.2 visualizes the effects that the three factors have on the movements of the term-structure of interest rates.

- The first factor is called a "level" factor as its changes lead to parallel shifts in the term-structure of interest rates. Thus, this "level" factor produces essentially the same effects on the term-structure as those underlying the "duration hedging" portfolio practice. This factor explains approximately 80% of the total variation of the yield curve.

- The second factor is called a "steepness" factor as its variations induce changes in the slope of the term-structure of interest rates. After a shock in this steepness factor, the short-end and the long-end of the yield curve move in opposite directions. The movements of this factor explain approximately 15% of the total variation of the yield curve.

- The third factor is called a "curvature" factor as its changes lead to changes in the curvature of the yield curve. That is, following a shock in the curvature factor, the middle of the yield curve and both the short-end and the long-end of the yield curve move in opposite directions. This curvature factor accounts for approximately 5% of the total variation of the yield curve.

Understanding the origins of these three factors is still challenging to financial economists and macroeconomists. For example, macroeconomists explain that central banks affect the short-end of the yield curve, e.g. by inducing variations in Federal Funds rate in the US. However,

decisions taken by the Federal Reserve rely on current macroeconomic conditions. Therefore, the short-end of the yield-curve likely links to macroeconomic developments. Instead, movements in the long-end of the yield curve should primarily depend on market expectations and risk-aversion surrounding future interest rates and economic conditions. Financial economists, then, should expect to see the long-end of the yield curve as being driven by expectations of future economic activity, and by risk-aversion.

Indeed, Ang and Piazzesi (2003) provide some empirical evidence that (i) macroeconomic factors such as inflation and real economic activity can explain movements at the short-end and the middle of the yield curve and, interestingly, (ii) the long-end of the yield curve is driven by unobservable factors. However, it is not clear whether such unobservable factors are driven by time-varying risk-aversion or changing expectations. A compelling lesson is that models of the yield curve driven by only one factor are likely to be misspecified, due to the complexity of roles and objectives of the so many diverse participants in the fixed income markets, and the links to the macroeconomy that decisions taken by these instititions have.

13.3.5.3 Forecasting the yield curve

Chapter 12 explains that simple techniques are available to fit the yield curve for the purpose of mere statistical descriptions of the data (see Appendix 1 in Chapter 12). One, proposed by Nelson and Siegel (1987), and reproduced here for convenience, postulates that the yield curve can be modeled as:

$$R(T) = \beta_1 + \beta_2 \left(\frac{1 - e^{-\lambda T}}{\lambda T} \right) + \beta_3 \left(\frac{1 - e^{-\lambda T}}{\lambda T} - e^{-\lambda T} \right).$$

The three coefficients, β_1, β_2 and β_3, can be interpreted in terms of the three factors reviewed in this section. The coefficient β_1 governs the level of the yield curve. The coefficient β_2 relates to the slope, as an increase in this coefficient increases short yields more than long yields.The coefficient β_3 shapes the curvature, as an increase in this coefficient has little effect on very short and very long yields, but increases the middle of the yield curve. Moreover, the coefficient λ controls the exponential decay of the yield curve: small values of λ translate to slow decay and can better fit the curve at long maturities; large values of λ, instead, lead to a fast decay, which helps fit the short-end of the yield curve. Finally, λ determines where the loading on β_3 achieves its maximum. Diebold and Li (2006) rely on this model, and estimate β_i for each date, and then use these estimated time series of β_i to forecast future values of β_i through vector autoregressions and, then, the future yield curve.

13.4 Models of the short-term rate: Introduction

The short-term rate is simply the growth rate, or velocity, at which "locally" riskless investments appreciate over the next infinitesimal amount of time. Naturally, this velocity is not a traded asset. Models where bond prices are tied up to interest rates are likely to be incomplete, in that to hedge against a bond, we cannot rely on anything underlying the bond price movements—what is traded is a money market account as well as the bond itself, not the interest rate. The evaluation framework in this context is one where bond prices can be replicated through other bond prices, as explained in a discrete time setting in Chapter 12. This issues is similar to that encountered in Chapter 11, where to price options in environments with random volatility, we needed to replicate options through other options.

13.4.1 Models versus representations

The pricing formula in Eq. (13.2),

$$P(t,T) = \mathbb{E}_t \left(e^{-\int_t^T r_u du} \right), \tag{13.24}$$

suggests a natural approach to model the arbitrage-free price P, based on the assumption that $(r_\tau)_{\tau \geq 0}$ is exogenously given. For example, we can assume that r_τ is the solution to

$$dr_\tau = b\left(r_\tau, \tau\right) d\tau + a\left(r_\tau, \tau\right) dW_\tau, \quad \tau \in (t, T], \tag{13.25}$$

where W_τ is a standard Brownian motion, and b, a are two functions that satisfy standard regularity conditions.

How should we choose these two functions? One possibility is to select functions that lead to a solution fast to compute in practice. Another possibility is that the chosen functions lead to empirically accurate model predictions. Well-known models solved in closed form are the celebrated "affine" models (see Section 13.5.3). The point any model should be accurate is more subtle. We may never achieve "perfect" accuracy in a model such as that in Eqs. (13.24)-(13.25). This model is one that *determines* the yield curve and it cannot then *exactly fit* it. Instead, a perfect fit of the yield curve is crucial when the aim is to price interest rate derivatives, as discussed in the chapter. However, we shall explain that the models in this section may well fit the yield curve, once equipped with an "infinite-dimensional" parameter calibrated to the observed yield curve.

Models with such a perfect accuracy are known as "no-arbitrage" models. They are the continuous-time counterparts to the implied trees dealt with in Chapter 12. In a sense, no-arbitrage models are *representations*, not *models* really. They force the short-term rate *process* to exactly pin down the yield curve at any instant and for this reason, they likely lead to intertemporal inconsistencies—the parameters that pin down the yield curve today, likely differ from those as of tomorrow. The rationale underlying their use goes to the opposite extreme of what we are describing in this section. In this section, the short-term rate is the *input* of all subsequent movements of the term-structure of interest rates. Instead, no-arbitrage models are a means through which we model future movements of the yield curve, by feeding information relating to the observed yield curve, not only the *current* short-term rate.

Models of the short-term rate are, then, and obviously, very useful, when the objective is to *explain* market behavior through a few inputs—a data reduction scientific principle. In these models, prices are economically admissible, i.e. no-arbitrage, and move while being driven by random changes in the state variables. The pricing errors they lead to may be seen as a virtue, not a defect. A permanent deviation of market data from the model's predictions might be a market anomaly rather than a model's imperfection, which the model could help exploit. Instead, and by construction, no-arbitrage models are needed to deal with pricing problems where bond prices have to match market data, and being able to change over time as a result of a probability distribution determined by the current market data, even when these market data are mispriced.

13.4.2 The bond pricing equation

13.4.2.1 A first derivation

Suppose bond prices are solutions to the following stochastic differential equation:

$$\frac{dP_i}{P_i} = \mu_{bi} d\tau + \sigma_{bi} dW, \tag{13.26}$$

where W is a standard Brownian motion in $\mathbb{R}^d$, μ_{bi} and σ_{bi} are some progressively measurable functions (σ_{bi} is vector-valued), and $P_i \equiv P(\tau, T_i)$. The exact functional form of μ_{b_i} and σ_{b_i} is *not* given, as in the BS case. Rather, it is endogenous and must be found as a part of the equilibrium.

Based on general results given in Chapter 5, we can show that the price system in (13.26) is arbitrage-free if and only if

$$\mu_{bi} = r + \sigma_{bi}\lambda, \tag{13.27}$$

for some $\mathbb{R}^d$-dimensional process λ satisfying some basic regularity conditions. Even if Eq. (13.27) follows by standard no-arbitrage arguments, Appendix 1 derives it again to emphasize how no arbitrage does specifically leads when it comes to bond markets.

The meaning of Eq. (13.27) is better understood by replacing this equation into Eq. (13.26), obtaining

$$\frac{dP_i}{P_i} = (r + \sigma_{bi}\lambda)\, d\tau + \sigma_{bi} dW.$$

The previous equation tells us that the growth rate of P_i is the short-term rate plus a *term-premium* equal to $\sigma_{bi}\lambda$. In the bond market, there are no obvious economic arguments enabling us to sign term-premia. Empirical evidence suggests that term-premia may take both signs. But term-premia would be zero in a risk-neutral world, where bond prices would, then, satisfy:

$$\frac{dP_i}{P_i} = r d\tau + \sigma_{bi} d\tilde{W},$$

where $\tilde{W} = W + \int \lambda d\tau$ is a Brownian motion under the risk-neutral probability, Q.

To illustrate the derivation of Eq. (13.27) when $d = 1$, consider the dynamics of the value V of a self-financed portfolio in two bonds and a money market account

$$dV = (\pi_1(\mu_{b1} - r) + \pi_2(\mu_{b2} - r) + rV)\, d\tau + (\pi_1\sigma_{1b} + \pi_2\sigma_{b2})\, dW,$$

where π_i is wealth invested in bond maturing at T_i: $\pi_i = \theta_i P_i$. We can zero uncertainty by setting

$$\pi_1 = -\frac{\sigma_{b2}}{\sigma_{b1}}\pi_2.$$

By replacing this into the dynamics of V,

$$dV = \underbrace{\left(-\frac{\mu_{b1} - r}{\sigma_{b1}}\sigma_{b2} + (\mu_{b2} - r)\right)}_{\equiv \ell}\pi_2 d\tau + rV d\tau.$$

Note that π_2 can always be chosen such that the value of this portfolio appreciates at strictly more than r: just set $\text{sign}\,(\pi_2) = \text{sign}\,(\ell)$. Therefore, to rule out arbitrage, $\ell = 0$, or

$$\frac{\mu_{b1} - r}{\sigma_{b1}} = \frac{\mu_{b2} - r}{\sigma_{b2}}.$$

That is, the Sharpe ratio for any two bonds has to equal a process λ, say, and Eq. (13.27) follows. Clearly, λ is independent of the two maturity dates, T_1 or T_2, which were arbitrary indeed. It is natural that λ is independent of any bond maturity, as this is unit price of risk required to compensate for randomness in the short-term rate.

The two functions μ_{bi} and σ_{bi} in Eq. (13.26) can be determined through Itô's lemma. Let $P(r,\tau,T)$ be the price at τ of a bond maturing at T when the state at τ is r—the "rational" bond pricing function. Because r is solution to Eq. (13.25), Itô's lemma implies that

$$dP = \left(\frac{\partial P}{\partial \tau} + bP_r + \frac{1}{2}a^2 P_{rr}\right) d\tau + aP_r dW,$$

where subscripts denote partial derivatives.

Comparing this equation with Eq. (13.26), and identifying drifts and diffusion terms, one finds that it must hold that

$$\mu_b P = \frac{\partial P}{\partial \tau} + bP_r + \frac{1}{2}a^2 P_{rr} \quad \text{and} \quad \sigma_b P = aP_r.$$

Finally, replace these functions into Eq. (13.27), and obtain that P satisfies the following partial differential equation (PDE, henceforth), for all r and $\tau \in [t,T)$:

$$\frac{\partial P}{\partial \tau} + bP_r + \frac{1}{2}a^2 P_{rr} = rP + \lambda a P_r, \tag{13.28}$$

with the boundary condition $P(r,T,T) = 1$, holding for all r. We will return to the the interpretation of this bond pricing equation in Section 13.4.2.

13.4.2.2 Derivation based on duration

An alternative idea is to replicate the price of a bond expiring at some time T_1, say $P^1 \equiv P(r,\tau,T_1)$, with a self-financed portfolio comprising a money market account and a second bond expiring at time $T_2 > T_1$. This approach is the continuous-time equivalent to the pricing approach in Chapter 12. It is, also, the interest rate counterpart to option evaluation with stochastic volatility in Chapter 11. Let, then, V be the value of our self-financed portfolio, $V = \Delta \cdot P^2 + M$, where Δ is the number of bonds maturing at T_2 to include in the portfolio, $P^2 = P(r,\tau,T_2)$, and M is the dollar amount in the money market account. Since the portfolio is self-financed, by the usual arguments we have that

$$dV = \Delta \cdot dP^2 + dM = \left(\Delta \cdot LP^2 + rM\right) d\tau + \Delta \cdot aP_r^2 dW, \tag{13.29}$$

where $LP^2 = \frac{\partial P^2}{\partial \tau} + bP_r^2 + \frac{1}{2}a^2 P_{rr}^2$. And, obviously,

$$dP^1 = LP^1 d\tau + aP_r^1 dW. \tag{13.30}$$

Let the initial value of the portfolio match the bond price. Then, comparing the diffusive terms in Eq. (13.29) and Eq. (13.30), we find the delta to be:

$$\hat{\Delta} = \frac{\partial P(r,\tau,T_1)/\partial r}{\partial P(r,\tau,T_2)/\partial r}.$$

Comparing the drift terms in Eq. (13.29) and Eq. (13.30),

$$LP^1 = \Delta \cdot LP^2 + rM = \Delta \cdot LP^2 + r\left(V - \Delta P^2\right) = \Delta \cdot LP^2 + r\left(P^1 - \Delta P^2\right),$$

where the last line follows because we are using the values (Δ, M), such that the portfolio value matches that of the first bond. Rearranging terms yields, $LP^1 - rP^1 = \Delta \cdot (LP^2 - rP^2)$, and evaluating this for $\Delta = \hat{\Delta}$,

$$\frac{LP^1 - rP^1}{P_r^1} = \frac{LP^2 - rP^2}{P_r^2} \equiv \Lambda \equiv \lambda a,$$

for some Λ and λ independent of calendar time.

The delta, $\hat{\Delta}$, can be interpreted as the ratio of the durations of the two bonds, as explained in Chapter 12. Such a duration-based derivation suggests a trading strategy. Suppose that the first bond price is mispriced against a given model, in that we have some confidence that our model does better than the market is currently doing while pricing this bond. For example, we might observe that the model quite often predicts that the price of this first bond is lower than the market, $P^1 < P^1_{\$}$, but that at the same time, the very same model prices well bonds on other points of the curve. We can implement the following relative value trade. We sale short the first bond for $P^1_{\$}$, realize an initial gain equal to $P^1_{\$} - P^1$, and invest P^1 into the previously described portfolio, comprising a money market account and the second bond, aiming to replicate the first bond at maturity $T_1 < T_2$. Naturally, the assumption is that the model is better than the market, which means that such a replication is indeed possible. Veronesi (2010, Chapter 16) develops examples that show the extent of success of this strategy based on the models of the short-term rate dealt with in this section.

13.4.2.3 Interpreting the bond pricing equation: market incompleteness and preference-free evaluation

Eq. (13.28) shows that the bond price, P, depends on both the drift of the short-term rate, b, and the risk-aversion correction, λ. Let us elucidate the circumstances leading to this fact. They link to market incompleteness.

Consider a benchmark complete market model, Black-Scholes, in which the option is redundant, given the initial market structure. As pointed out in Chapter 12, in markets such as Black & Scholes, the stock price *is* the source of randomness, and markets are trivially completed by the very same asset.

By contrast, the bond price is *not* the risk. Rather, the bond draws value from the interest rate risk, *as if* it was a "derivative" on r_t, where r_t forms, alone, "an incomplete market structure." For this reason, the bond price cannot be determined in a preference-free format. Similarly, in equity markets with stochastic volatility (see Chapter 11), two sources of randomness arise, the stock price and its stochastic volatility, such that a portfolio comprising a stock *and* one additional option is needed to hedge against one given option—and no preference-free formulae are possible. Let us mention another example. We are facing a situation similar to those encountered in Part II, where given a certain dividend risk, the stock price could not, then, expressed in a preference-free fashion: the dividend in that context was the risk, and the stock price was like a derivative written on that risk.

In Chapter 2, we learnt indeed that given an incomplete market, there might be securities that could be introduced to make markets complete. For example, a market with two securities and three states is incomplete. Yet it may be completed once a third security is introduced, which together with the first two, can generate any consumption bundle in all the possible states of the world—we need that the payoff matrix of the three securities be no-singular. Does this mean that the third security can be priced in a preference-free format? Of course not. By construction, the payoffs promised by the third security cannot be replicated by trading the first two. In other words, the third security is not redundant, for otherwise it would not complete the initial three-state/two security market. And if it is not redundant, its price reflects the services it provides the society with and thus, it cannot be preference-free. An hypothetical fourth security could, however, be replicated by trading the three securities, and would thus be preference-free.

The market in this section shares similarities with this simple example, but with qualifications. As mentioned, the bond completes the market, but its price cannot be replicated through

the short-term rate, because the short-term rate is not traded. However, we can replicate a second bond price by trading the first bond (and a money market account). But the second bond price cannot be expressed in a preference-free format: there are actually no bond prices that could be expressed in a preference-free format, for the very reason that the first bond cannot! The situation in the context of this section is actually involved. Issues arise not only because we try to replicate a bond price with another bond price. Suppose, indeed, that we try to replicate the price of an option written on this bond throughout the very same bond. We could not achieve a preference-free solution either. By the reasoning in the previous sections, we need to make sure that the volatility of the bond price is matched to that of the option price. Granted, we can achieve this. However, the volatility of the bond price is not exogeneous, as it depends on the price sensitivity with respect to r_t, which is obviously not preference-free.

Finally, the reason the bond price depends on b and a is intimately related to the previous explanations. In this market, uncertainty is generated by an untraded risk, r_t, such that the initial market structure has one untraded risk, r_t, and zero assets, as explained. But because the short-term rate r_t is not a discounted martingale under Q, its drift does not equal to $r_t \cdot r_t = r_t^2$ under Q but, rather $b - \lambda a$, thereby leading to Eq. (13.28). All in all, the bond price depends on the specific functional forms of b, a and λ.

This dependence on risk-appetite might appear as a kind of hindrance to practitioners, but it may actully be a source of value, for its potential to inform on agents' risk-appetite λ: once two functions (b, a) are estimated, the yield curve could be "inverted" to infer the implied λ, which could help policy makers to take more informed decisions about how to affect the yield curve—admittedly though, the task of estimating (b, a) is far from trivial.

By specifying (b, a) and identifying the risk-premium λ, the PDE in Eq. (13.28) can then be solved, either analytically or numerically. Choices concerning the exact functional form of b, a and λ are often made on the basis of analytical or empirical reasons. In Section 13.4.4, we introduce the first, famous models in which b, a and λ have a particularly simple form. Section 13.4.5 discusses analytical merits but major empirical drawbacks of these models. In Section 13.4.6 we provide a very succinct description of models exhibiting jump (and default) phenomena. We, first, highlight how models of the short-term rate could be used to set up a metric of duration when this concept is ore difficult to quantify than in the contexts outlined in Chapter 12.

13.4.3 Stochastic duration

Duration is a measure of risk tailored to capture the notion of fixed income volatility. Cox, Ingersoll and Ross (1979) hinge upon models of the short-term rate and introduce the notion of *stochastic duration*, generalizing the notion of modified duration discussed in Chapter 12. Suppose the price of a zero-coupon bond is a function of the short-term rate only, $P(r, T-t)$. Define the *basis risk* as the semi-elasticity of the bond price with respect to the short-term rate,

$$\Psi(r, T-t) \equiv -\frac{P_r(r, T-t)}{P(r, T-t)}.$$

Naturally, we want to make sure that the measure of duration for a zero.coupon bond equals time-to-maturity. Therefore, Ψ cannot be a measure of duration: except in the trivial case in which r is constant, Ψ does not equal $T-t$ for a zero. The idea underlying "stochastic duration" of a given bond is to search for the time-to-maturity $T^* - t$ of an hypothetical zero-coupon bond such that its basis risk is the same as that of the given bond (e.g., a coupon bearing bond or a

callable bond), viz

$$\Psi\left(r, T^{*}-t\right)=-\frac{B_{r}\left(r, S-t\right)}{B\left(r, S-t\right)},$$

where $B\left(r, S-t\right)$ is the price of the given bond that delivers its face value at time S, if no events preventing this occur prior to time S (say, default or the exercise of the optionalities embedded into the institutional details of the bond). Stochastic duration of this bond is defined as the time-to-maturity $T^{*}-t$ of the hypothetical zero-coupon bond:

$$T^{*}-t\equiv D\left(r, S-t\right)=\Psi^{-1}\left(-\frac{B_{r}\left(r, S-t\right)}{B\left(r, S-t\right)}\right), \tag{13.31}$$

where Ψ^{-1} is the inverse function of $\Psi\left(r, \tau\right)$ with respect to time to maturity τ. It is immediate to verify that for a zero, $D\left(r, S-t\right)=S-t$, Moreover, the stochastic duration, $D\left(r, T-t\right)$, collapses to the modified duration introduced in the previous chapter (see Section 12.4) once the short-term rate is a constant.[11]

13.4.4 Some famous models

13.4.4.1 Vasicek

Vasicek (1977) derives a model of the yield curve assuming the short-term rate r_{τ} is a continuous-time and mean-reverting process with constant basis point volatility, solution to:[12]

$$dr_{\tau}=\kappa\left(\bar{r}-r_{\tau}\right)d\tau+\sigma dW_{\tau},\quad \tau\in\left(t, T\right], \tag{13.32}$$

where $\bar{r}$, κ and σ are positive constants. This model is more sensible than that of Merton (1973), where the short-term rate is an arithmetic Brownian motion. The intuition underlying the importance of mean-reversion is as follows. Suppose, first, that $\sigma=0$, in which case,

$$r_{\tau}=\bar{r}+e^{-\kappa(\tau-t)}\left(r_{t}-\bar{r}\right). \tag{13.33}$$

If the current level of the short-term rate $r_{t}=\bar{r}$, it will be "locked-in" at $\bar{r}$ forever. If, $r_{t}<\bar{r}$, the short-term rate shall steadily increase, and converge to $\bar{r}$ as $\tau\to\infty$. Likewise, the short-term rate shall converge to $\bar{r}$ when $r_{t}>\bar{r}$. The speed of convergence of r to the long-term value $\bar{r}$ depends on the magnitude of κ: the higher κ, the higher the speed of convergence to $\bar{r}$.

In the general case, $\sigma\neq 0$, the solution to Eq. (13.32) is,

$$r_{\tau}=\bar{r}+e^{-\kappa(\tau-t)}\left(r_{t}-\bar{r}\right)+\sigma e^{-\kappa\tau}\int_{t}^{\tau}e^{\kappa s}dW_{s},$$

where the integral has to be understood in the Itô's sense. The interpretation of this solution is similar to that in the determinist case, in that the short-term rate now fluctuates around its "central tendency" $\bar{r}$. In other words, shocks are absorbed at a speed depending on the magnitude of κ, leading the short-term rate to display a mean-reverting behavior. Indeed, the conditional expectation of r is the same as that in Eq. (13.33),

$$E\left(r_{\tau}|\, r_{t}\right)=\bar{r}+e^{-\kappa(\tau-t)}\left(r_{t}-\bar{r}\right). \tag{13.34}$$

[11] Indeed, note that if r is constant, then, $P\left(r, T-t\right)=e^{-r(T-t)}$, and $\Psi^{-1}\left(r, x\right)=x$.

[12] Note that Merton's (1973) seminal paper is also framed in a context with random interest rates solved in closed form. In Merton's model, the short-term rate is an arithmetic Brownian motion.

Moreover, the conditional variance of r is:

$$var\left(r_\tau|\, r_t\right) = \frac{\sigma^2}{2\kappa}\left(1 - e^{-2\kappa(\tau - t)}\right).$$

Finally, it can be shown that r is normally distributed, with expectation and variance given by the two functions given above.

To solve for the entire term-structure of interest rates, we need to make assumptions about the risk-premium, λ. A closed-form expression for the bond price obtains, once we assume λ is a constant. Indeed, by replacing a constant risk-premium λ and the functions $b(r) = \kappa\left(\bar{r} - r\right)$ and $a(r) = \sigma$ into Eq. (13.28), and denoting $r^* \equiv \bar{r} - \frac{\lambda\sigma}{\kappa}$, we obtain that the bond price P is solution to:

$$0 = \frac{\partial P}{\partial \tau} + \kappa\left(r^* - r\right)P_r + \frac{1}{2}\sigma^2 P_{rr} - rP, \quad \text{for all } (r,\tau) \in \mathbb{R} \times [t,T), \tag{13.35}$$

with the usual boundary condition. Intuitively, $\kappa\left(r^* - r\right)$ is the drift of the short-term rate under Q, which is higher than under P for $\lambda < 0$, reflecting higher Arrow-Debreu state prices for the bad states of the world arising when interest rates are high. It is instructive to see how this partial differential equation can be solved. We guess a solution of the form:

$$P\left(r,\tau,T\right) = e^{A(\tau,T) - B(\tau,T)\cdot r}, \tag{13.36}$$

for two functions A and B to be determined. Now suppose the guess in Eq. (13.36) is true. By replacing the partial derivatives of P into Eq. (13.35) leaves the equation $\phi_1\left(\tau\right) + \phi_2\left(\tau\right)r = 0$, where for all τ,

$$\phi_1\left(\tau\right) \equiv A_1\left(\tau,T\right) - \kappa r^* B\left(\tau,T\right) + \frac{1}{2}\sigma^2 B^2\left(\tau,T\right) \quad \text{and} \quad \phi_2\left(\tau\right) \equiv \kappa B\left(\tau,T\right) - B_1\left(\tau,T\right) - 1.$$

That is, $0 = \phi_1\left(\tau\right) = \phi_2\left(\tau\right)$, two ordinary differential equations subject to the boundary conditions $A\left(T,T\right) = 0$ and $B\left(T,T\right) = 0$. The solutions are:

$$A\left(\tau,T\right) = \left(\frac{1 - e^{-\kappa(T-\tau)}}{\kappa} - (T-\tau)\right)r_\infty - \frac{\sigma^2}{4\kappa^3}\left(1 - e^{-\kappa(T-\tau)}\right)^2, \quad B\left(\tau,T\right) = \frac{1}{\kappa}\left(1 - e^{-\kappa(T-\tau)}\right), \tag{13.37}$$

where

$$r_\infty = r^* - \frac{1}{2}\left(\frac{\sigma}{\kappa}\right)^2.$$

The term-structure of spot rates predicted by this model is, by the definition in Eq. (13.11),

$$R\left(r,t,T\right) = -\frac{A\left(t,T\right)}{T-t} + \frac{B\left(t,T\right)}{T-t}r. \tag{13.38}$$

The quantity, r_∞, is interpreted as the "asymptotic" spot rate, as it is the limit of $R\left(r,t,T\right)$ for large T.

The model displays a number of features that match some empirical facts—such as selected shapes of the yield-curve. Plot it.

13.4.4.2 A case study: expectations and business cycles

The Vasicek's model can illustrate aspects of the expectation hypothesis. We can also rely on a variant of this model to rationalize the predicting power of the term spread (see Section 13.3). First, we express the yield curve in Eq. (13.38) in a way that is more convenient to interpret. We mentioned earlier that the short-term rate is conditionally normally distributed. Now, the term $\int_t^T r_\tau d\tau$ is also conditionally normally distributed,[13] and, then, by Eq. (13.2) and Eq. (13.11),

$$R(r,t,T) = \mathbb{E}_t\left(\frac{1}{T-t}\int_t^T r_\tau d\tau\right) - \frac{1}{2}\frac{1}{T-t}\text{var}\left(\int_t^T r_\tau d\tau\right). \tag{13.39}$$

The second term in Eq. (13.39) reflects Jensen's inequality effects. It equals

$$\frac{1}{2}\frac{1}{T-t}\text{var}\left(\int_t^T r_\tau d\tau\right) = \frac{1}{2}\left(1 - \frac{1-e^{-\kappa(T-t)}}{(T-t)\kappa}\right)\left(\frac{\sigma}{\kappa}\right)^2 - \frac{\sigma^2}{4(T-t)\kappa^3}\left(1-e^{-\kappa(T-t)}\right)^2,$$

and is much smaller than the first term when evaluated using realistic parameter values.

We approximate the yield curve with the first term of Eq. (13.39), which is an expectation of the average short-term rate under the risk-neutral probability. By Eq. (13.34) and Eq. (13.28), this expectation can be decomposed as the sum of the expectation under the physical probability plus risk-premiums terms, as follows:

$$R(r,t,T) \approx \mathbb{E}_t\left(\frac{1}{T-t}\int_t^T r_\tau d\tau\right) = \frac{1}{T-t}\int_t^T E_t(r_\tau)\,d\tau + \frac{1}{T-t}\int_t^T \mathcal{E}_\tau d\tau, \tag{13.40}$$

where:

$$\frac{1}{T-t}\int_t^T E_t(r_\tau)\,d\tau = \bar{r} + (r_t - \bar{r})\frac{1-e^{-\kappa(T-t)}}{\kappa(T-t)}, \quad \mathcal{E}_\tau \equiv -\frac{1-e^{-\kappa(T-\tau)}}{\kappa}\sigma\lambda.$$

Eq. (13.40) says that long-term rates reflect expectations of the future short-term rate and risk-premiums terms, defined as the average expected return on the bond.

We can rely on this simple framework to describe some of the business cycle properties of the yield curve. We assume that business cycle conditions can be summarized by the realization of a single state variable, y, which is solution to the following stochastic differential equation

$$dy_\tau = \kappa(\bar{y} - y_\tau)\,d\tau + \sigma_y dW_\tau,$$

for three positive constants, b, $\bar{y}$ and σ_y. Next, suppose that: (i) the *nominal* short-term rate is procyclical, in that $r_\tau \equiv A + By_\tau$, for two positive constants A and B, and that (ii) risk-premiums are countercyclical in that, $\lambda_y(y) \equiv \lambda_0 - \lambda_1 y$, for two additional constants, where $\lambda_1 > 0$, to ensure countercyclicality, and λ_0 might in principle take any sign, although it is reasonable to assume that $\lambda_0 > 0$, which would imply that the constant portion of the risk-premium is positive anyway. We shall return to the sign of λ_0 in a moment.

Given the assumptions made so far, the short-term rate is solution to Eq. (13.32), with parameters $\bar{r} \equiv A + B\bar{y}$ and $\sigma \equiv B\sigma_y$. While the risk-premium is time-varying, being equal to $\lambda(r) \equiv \lambda_0 - \frac{\lambda_1}{B}(r - A)$, the short-term rate is still conditionally normally distributed under the risk-neutral probability, and the yield-curve can be solved in closed form, with a solution like

[13] See, for example, Lamberton and Lapeyre (1997, Chapter 6).

that in Eq. (13.36), and an approximation like that in Eq. (13.40). It is instructive to calculate the decomposition in Eq. (13.40) predicted by this model. We have,

$$R\left(r,t,T\right)\approx\mathbb{E}_{t}\left(\frac{1}{T-t}\int_{t}^{T}r_{\tau}d\tau\right)=r^{*}+\left(r_{t}-r^{*}\right)\frac{1-e^{-\kappa^{*}\left(T-t\right)}}{\kappa^{*}\left(T-t\right)},$$

where $\kappa^{*}=\kappa-\sigma_{y}\lambda_{1}$ and $r^{*}=\frac{\kappa\bar{r}-B\sigma_{y}\lambda_{0}-A\sigma_{y}\lambda_{1}}{\kappa-\sigma_{y}\lambda_{1}}$: the model clearly collapses to Vasicek's for $\lambda_{1}=0$. A countercyclical risk-premium, $\lambda_{1}>0$, leads to a tilt in the unconditional short-term rate, r^{*}, and, importantly, an increased persistence of the short-term rate, due to the fact that $\kappa^{*}<\kappa$. If risk premium effects are so large to make $\kappa^{*}<0$, the model has no solutions for large T. Finally, note that while it is reasonable to assume $\lambda_{0}>0$, we may also wish to make sure λ_{0} is small enough to ensure that for most of the time, $\lambda\left(r\right)$ is negative and, hence, expected excess returns are positive.

The term-spread (over the short-term rate) predicted by the model is

$$R\left(A+By_{t},t,T\right)-r_{t}\approx B\left(\frac{1-e^{-\kappa^{*}\left(T-t\right)}}{\kappa^{*}\left(T-t\right)}-1\right)\left(y_{t}-y^{*}\right),$$

where $y^{*}=\frac{\kappa\bar{y}-\sigma_{y}\lambda_{0}}{\kappa-\sigma_{y}\lambda_{1}}$ is the unconditional expectation of the procyclical variable y (under the risk-neutral probability). The term spread is the product of three terms. The first reflects the property that interest rates are procyclical, $B>0$. The second term is negative when $\kappa^{*}>0$, in which case the model formalizes explanations given in Section 13.3.3. That is, before a "peak," i.e., when $y_{t}<y^{*}$, the yield curve is upward sloping. After this peak is achieved and, then, $y_{t}>y^{*}$, the probability the economy is in recession is more likely, given the mean-reverting nature of y_{t}, and the yield curve becomes inverted. In other words, at the peak of an expansion, when a recession is most likely, we expect that interest will lower due to their procyclicality, $B>0$, whence the yield curve inversion. Long term interest rates are driven by expectations of future short-term rates.

Note that these effects take place once countercyclical risk-aversion is mild enough to guarantee that $\kappa^{*}>0$. If λ_{1} was so large to make $\kappa^{*}<0$, the model would generate the wrong predictions, with an inverted yield curve during the rising part of a boom, not the descending part. The mechanism would be that, during a peak, the future short-term rate would be expected to lower, but risk-aversion would be expected to increase to a level so high to dwarf expectations effects and push future prices down: future countercyclical risk aversion would dominate procyclical effects generated by the short-term rate.

Naturally, this model is very simple, driven as it is by only one factor, i.e., the business cycle variable y_{t}. Its merit is that it makes a sharp prediction regarding the slope of the yield curve: a positive slope occurs before a peak, $y_{t}<y^{*}$, and a negative slope after the peak, $y_{t}<y^{*}$, similarly as in the data. This model thus isolates the business cycle component of the yield curve that relates to its inversions. The crucial point is that the model is silent as regards the business cycle variable y_{t}. If we knew y_{t}, we could use it to forecast the business cycle in the first place.

13.4.4.3 Cox, Ingersoll and Ross

Vasicek's model suffers from two main drawbacks. First, the short-term rate is normally distributed. This circumstance might be mitigated when σ is low, compared to $\bar{r}$, in which case the probability the short-term rate takes negative values can be small. At the same time, even

a small probability of a negative interest rate might lead to severe mispricings when it comes to pricing interest rate derivatives, due to nonlinearities induced by optionality, as pointed out by Dybvig [cite reference]. The previous chapter (see Section 12.5.3) displays numerical examples where small changes in assumptions can lead to quite substantial changes in the price of derivatives.

The second drawback, related to the first, is that the short-term rate volatility is independent of the level of the short-term rate. It might be argued that short-term rates changes become more and more volatile as the level of the short-term rate increases, a phenomenon usually referred to as the *level-effect*.

Cox, Ingersoll and Ross (1985) (CIR, henceforth) propose a model that addresses these two drawbacks at once. They rely on an equilibrium model in which the short-term rate can be expressed as the solution to (see Eq. (5.102) in Chapter 5)

$$dr_\tau = \kappa(\bar{r} - r_\tau)d\tau + \sigma\sqrt{r_\tau}dW_\tau, \quad \tau \in (t, T].$$

The CIR model is also referred to as "square-root" process to emphasize that the diffusion function is proportional to the square-root of r. This feature makes the model address the level-effect phenomenon. The evidence about the level-effect is further discussed below (see Section 13.4.5). Moreover, this property prevents r from taking negative values. Intuitively, when r wanders just above zero, it is pulled back to the stricly positive region at a strength of the order $dr = \kappa\bar{r}d\tau$.[14] The transition density of r is noncentral chi-square. The stationary density of r is a gamma distribution. The expected value is as in Vasicek.[15] It is left as an exercise to show that

$$var\left(r_\tau | r_t\right) = \frac{\bar{r}\sigma^2}{2\kappa} + \left(r_0 - \bar{r}\right)\frac{\sigma^2}{\kappa}e^{-\kappa(\tau - t)} + \frac{\sigma^2}{\kappa}\left(\frac{1}{2}\bar{r} - r_0\right)e^{-2\kappa(\tau - t)}.$$

Intuitively, the variance of the short-term rate depends on the unconditional expectation, $\bar{r}$, because the instantaneous variance of the short-term rate depends on r_τ.

CIR formulate a set of assumptions (e.g., preferences), leading to a risk-premium function $\lambda = \ell\sqrt{r}$, where ℓ is a constant (see Eq. (5.103) in Chapter 5). By replacing this, $b(r) = \kappa\left(\bar{r} - r\right)$ and $a(r) = \sigma\sqrt{r}$ into the partial differential equation (13.28), one obtains that, similarly as in the Vasicek model, the bond price function has the same form as in Eq. (13.36), but with functions A and B satisfying the following differential equations

$$0 = A_1\left(\tau, T\right) - \kappa\bar{r}B\left(\tau, T\right) \quad \text{and} \quad 0 = -B_1\left(\tau, T\right) + \left(\kappa + \ell\sigma\right)B\left(\tau, T\right) + \frac{1}{2}\sigma^2 B^2\left(\tau, T\right) - 1,$$

subject to the boundary conditions, $A\left(T, T\right) = 0$ and $B\left(T, T\right) = 0$. The solution to the bond price is

$$\begin{gathered} P\left(r_t, T - t\right) = \Phi\left(T - t\right)e^{-B(T-t)r_t}, \\ \Phi\left(T - t\right) = \left(\frac{2\gamma e^{\frac{1}{2}(\kappa_o + \gamma)(T-t)}}{\left(\kappa_o + \gamma\right)\left(e^{\gamma(T-t)} - 1\right) + 2\gamma}\right)^{\frac{2\kappa_o\bar{r}_o}{\sigma^2}}, \quad B\left(T - t\right) = \frac{2\left(e^{\gamma(T-t)} - 1\right)}{\left(\kappa_o + \gamma\right)\left(e^{\gamma(T-t)} - 1\right) + 2\gamma}, \\ \gamma = \sqrt{\kappa_o^2 + 2\sigma^2}, \quad \kappa_o = \kappa + \sigma\ell, \quad \bar{r}_o = \tfrac{\kappa}{\kappa + \sigma\ell}\bar{r}. \end{gathered} \tag{13.41}$$

[14] This is only intuition. The exact condition under which the zero boundary is unattainable by r is $\kappa\bar{r} > \frac{1}{2}\sigma^2$. See Karlin and Taylor (1981, vol II chapter 15) for a general analysis of attainability of boundaries for scalar diffusion processes.

[15] The expected value of a mean-reverting processes with a linear drift is always as in Vasicek, independent of the assumption on the diffusion coefficient. This property is intuitive (heteroskedasticity does not affect expected values) can be shown through methods introduced in Chapter 8 (Appendix 5).

This model has been the reference in the industry for many years, and it still is a reference. First, there are multidimensional extensions of this basic model, as reviewed in Section 13.5.3. Second, square root processes have revealed to be useful while modeling other important variables in finance that are strictly positive, such as the probability of default of a certain name (see the next chapter).

13.4.4.4 Nonlinear drifts

Models that are analytically tractable are truly valuable. Both Vasicek and CIR models lead to closed-form solutions as they both have a *linear* drift, among other things. What does the empirical evidence say about mean-reversion in the short-term rate? This issue is controversial. In the mid 1990s, three papers by Aït-Sahalia (1996), Conley et al. (1997) and Stanton (1997) produce evidence of nonlinear mean-reverting behavior. For example, Aït-Sahalia (1996) estimates a drift function that has the following form:

$$b(r) = \beta_0 + \beta_1 r + \beta_2 r^2 + \beta_3 r^{-1}, \tag{13.42}$$

corresponding to a nonlinear diffusion function. Figure 13.3 reproduces this function using the parameter values in his Table 4, and relating to the sample period from 1983 to 1995. Similar results are reported in the other papers. To illustrate the action the short-term rate dynamics are under, Figure 13.3 also depicts a linear drift, obtained with the parameter estimates of Aït-Sahalia (1996) (Table 4), and relating to a model with a CEV diffusion.

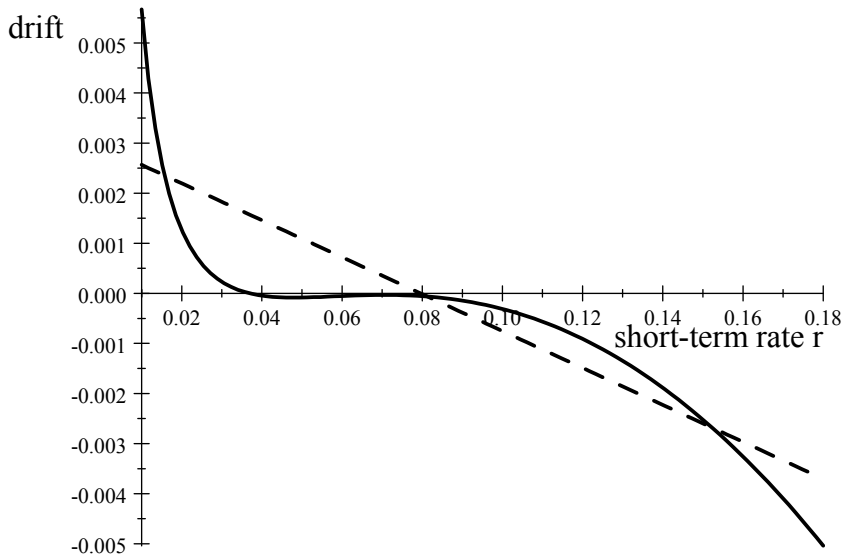


FIGURE 13.3. Nonlinear mean reversion? The solid line is the drift function in Eq. (13.42), estimated by Aït-Sahalia (1996), and relating to a parametric model with a nonlinear diffusion function. The dashed line is the estimated linear drift relating to a model with CEV diffusion.

The nonlinear drifts in Figure 13.3 might lead bond prices to exhibit unusual properties, though. As explained in Chapter 8 (see Appendix 5), bond prices are *concave* in the short-term rate if the risk-neutralized drift function is sufficiently convex (Mele, 2003). While the results in Figure 13.3 relate to the *physical* drift functions, the point is nevertheless important as risk-premiums should look like quite unusual to destroy the nonlinearities of the short-term rate under the physical probability.

The compelling lesson from Figure 13.3 is that under the "nonlinear drift dynamics," the short-term rate behaves in a way that can at least be *roughly* comparable with that it would behave under the "linear drift dynamics." However, the behavior at the extremes is dramatically different. As the short-term rate moves to the extremes, it is pulled back to the "center" in a very abrupt way. At the moment, it is not clear whether these preliminary empirical results are reliable or not. New econometric techniques are currently being developed to address this and related issues.

One possibility is that such *single factor* models of the short-term rate are simply misspecified. For example, there is strong empirical evidence that the volatility of the short-term rate is time-varying, as we shall discuss in the next section. Moreover, the term-structure implications of a single factor model are counterfactual, since we know that a single factor cannot explain the entire variation of the yield curve, as explained in Section 13.3.5. We now describe more realistic models driven by more than one factor.

13.4.5 The Monetary Experiment and interest rate volatility

One of the motivation underlying the early adoption of the CIR model in the industry was the property of the model to predict interest rate volatility to increase with the level of interest rates. Is this a robust empirical feature? It is a difficult topic, with the answer relying on particular historical episodes. Certainly the episodes of FED QE after 2010 when interest were extremely low were accompanied by a suppressed volatility. And, during the Monetary Experiment of the Federal Reserve occurred between October 1979 and October 1982, both interest rates and interest rate volatility were high.

Figure 13.4 depicts the time series behavior of the nominal short-term rate, as measured by the three month TB rate, as well as the volatility of its changes, calculated as $\text{Vol}_{r,t} \equiv \sqrt{6\pi} \cdot \bar{\sigma}_{r,t}$, where $\bar{\sigma}_{r,t} \equiv \frac{1}{12}\sum_{i=1}^{12} |r_{t+1-i} - r_{t-i}|$, and r_t is the short-term rate as of month t.[16] Figure 13.5 plots a scatterplot of the short-term rate basis point volatility, $\text{Vol}_{r,t}$, against r_t, for two sampling periods. The short-term rate r_t is expressed in percentage, and we express the basis point volatility $\text{Vol}_{r,t}$ in percentage points, such that a value of one in Figures 13.4 and 13.5 equals 100 basis points. For later use, define percentage volatility as $\text{Vol}^{\%}_{r,t} \equiv \sqrt{6\pi} \cdot \bar{\sigma}^{\%}_{r,t}$, where $\bar{\sigma}^{\%}_{r,t} \equiv \frac{1}{12}\sum_{i=1}^{12} \left|\ln \frac{r_{t+1-i}}{r_{t-i}}\right|$.

One clear property is that interest rate volatility appears to be countercyclical, spiking as it does in most of the NBER recession episodes. A natural explanation is that the FED is more aggressive decreasing interest rates in bad times than increasing them in good.

We previously discussed the "level effect," defined as the statistical relation arising when the volatility of the short-term rate increases with the short-term rate. One possible explanation of these episodes is that when liquidity is erratic, interest rates are high reflecting a liquidity risk-premium. But precisely because of erratic liquidity, interest rates are also very volatile in such periods. A simple statistical model that could help deal with these facts is one in which

[16] This calculation follows that implemented to measure aggregate equity volatility in Chapter 8 (see Section 8.2).

the short-term rate r_t has stochastic volatility, as follows:

$$\begin{cases} dr_t = \kappa\left(\mu - r_t\right)dt + \sqrt{v_t} r_t^{\eta} dW_{1t} \\ dv_t = \beta\left(\alpha - v_t\right)dt + \xi v_t^{\vartheta}(\rho dW_{1t} + \sqrt{1-\rho^2} dW_{2t}) \end{cases} \tag{13.43}$$

where W_i are standard Brownian motions under the physical probability, v_t is a factor affecting in interest rate volatility, and remaining notation refers to model's parameter. In particular, $\rho : |\rho| < 1$ is the instantaneous correlation between dr_t and dv_t. This model generalizes the one factor models in the previous section, in which the yield curve is only driven by the short-term rate. If $\eta > 0$, the instantaneous rate volatility increases with the level of the interest level. If the "correlation" coefficient $\rho > 0$, interest rate volatility is also partly related to sources of volatility not directly affected by the *level* of the interest rate.

Yet the empirical evidence underlying Figure 13.5 does not lend much support to the level effect. Rather, it seems that interest rate volatility to be quite flat against the level of rates. The exception, of course, occurred over the Monetary Experiment, when the FED target was money supply, rather than interest rates. Over this period, high volatility of money demand mechanically translated to high interest rate volatility through market clearing. Additionally, monetary base over this period was kept deliberately low as an attempt to fight against inflation. Whence, both interest rate volatility and interest rates were very high. One additional reason for the high nominal rates at the time might link to a compensation for high *inflation volatility*—not only high inflation.

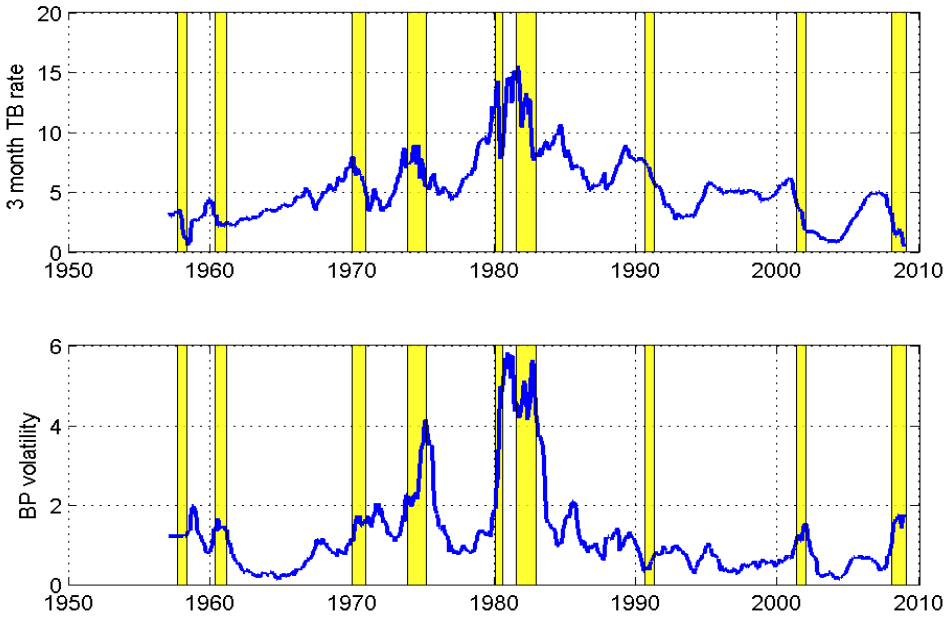


FIGURE 13.4. This picture depicts the time-series behavior of the 3 month TB rate (top panel, in percent) and the rolling, basis point volatility of the 3 month TB rate *changes*, $\text{Vol}_{r,t}$ (bottom panel), over the sampling period from 1957:01 through 2008:12. BP volatility is expressed in percentage terms, such that a value of one in the graph is the same as 100 basis points.

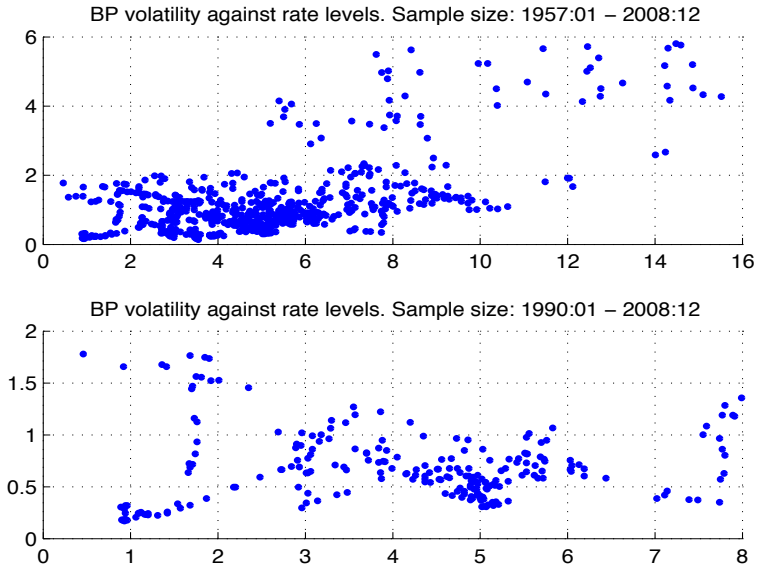


FIGURE 13.5. This picture plots the basis point volatility of the short-term rate *changes*, $\text{Vol}_{r,t}$ (on the vertical axis), against the level of the short-term rate (on the horizontal axis), for the sampling period spanning 1957:01 through 2008:12 (top panel), and for a more recent sample spanning 1990:01 through 2008:12 (bottom panel).

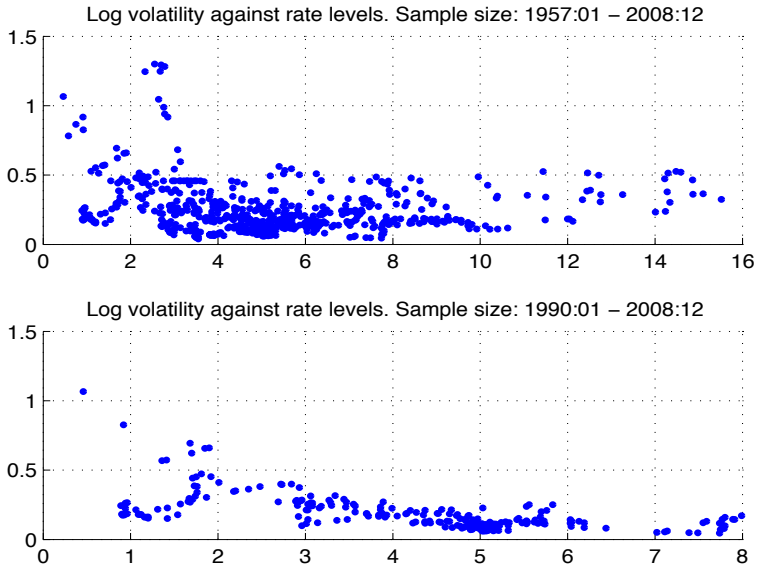


FIGURE 13.6. This picture plots the percentage volatility of the short-term rate $\text{Vol}^{\%}_{r,t}$ (on the vertical axis), against the level of the short-term rate (on the horizontal axis), for the sampling period spanning 1957:01 through 2008:12 (top panel), and for a more recent sample spanning 1990:01 through 2008:12 (bottom panel).

Figure 13.5 is also suggestive of a change in regime that possibly occurred over a more recent past. From 1990 on, interest rate volatility does not necessarily appear to positively link to rate levels, and there is evidence of the opposite. Figure 13.6 suggests that "percentage" volatility could actually be inversely related to the short-term rate, and over the more recent sample periods too.

All in all, evidence is mixed regarding the relation between the level of interest rates and interest rate volatility. The interesting property of the model in (13.43) is that it allows interest rate volatility to fluctuate, driven by its own source of risk, W_{2t}. In the next section, we review models of the yield curve with stochastic volatility in a more systematic fashion. We end the ongoing section with a succinct account of how to model the yield curve in the presence of jumps.

13.4.6 Jumps

13.4.6.1 Short-term rates as jump-diffusion processes

Ahn and Thompson (1988) extend the CIR model to one where the short-term rate is a jump-diffusion process. In general, suppose that under the risk-neutral probability, the short-term rate is a jump-diffusion process,

$$dr_\tau = b^J(r_\tau)d\tau + a(r_\tau)dW_\tau + \ell(r_\tau)\cdot\mathcal{S}\cdot dN_\tau,$$

where b^J is a jump-adjusted risk-neutral drift. The bond price $P(r,\tau,T)$ is solution to

$$0 = \left(\frac{\partial}{\partial\tau} + L - r\right)P(r,\tau,T) + v^Q\int_{\text{supp}(\mathcal{S})}\left[P\left(r+\ell\mathcal{S},\tau,T\right) - P\left(r,\tau,T\right)\right]p\left(d\mathcal{S}\right), \qquad (13.44)$$

where v^Q is the jumps intensity under Q.[17] The usual boundary condition is $P(r,T,T)=1$. Eq. (13.44) follows by the usual martingale property of the zero coupon bond price. The model can be extended to one with different quality, or types, of jumps:

$$0 = \left(\frac{\partial}{\partial\tau} + L - r\right)P(r,\tau,T) + \sum_{j=1}^{J} v_j^Q\int_{\text{supp}(\mathcal{S})}\left[P\left(r+\ell\mathcal{S},\tau,T\right) - P\left(r,\tau,T\right)\right]p^j\left(d\mathcal{S}\right),$$

where J is the number of jump types, but we simplify the exposition and take $J=1$

13.4.6.2 Defaultable bonds

Poisson processes can be used to model defaultable bonds. In particular, assume that the event of default is a Poisson process N with intensity v, and that in the event of default at τ, the bondholder receives a recovery payment $\bar{P}_\tau$, which could be deterministic or dependent on the short-term rate.[18] Let $\hat{\tau}$ be the random time of default, and define

$$g = \begin{cases} 0 & \text{if } t \le \tau < \hat{\tau} \\ 1 & \text{otherwise} \end{cases}$$

[17] It is $v^Q = v\cdot\lambda^J$, where v is the jump intensity under the *physical* probability and λ^J is the risk-premium related to the presence of jumps.

[18] The approach we are illustrating is known as "structural." Chapter 14 contains a systematic treatment regarding the pricing of defaultable bonds. The derivations in this section are based on the partial differential equations derived in Mele (2003).

such that, under the risk-neutral probability,

$$\begin{cases} dr_\tau = b(r_\tau)d\tau + a(r_\tau)dW_\tau \\ dg_\tau = \mathcal{S} \cdot dN_\tau, \end{cases} \tag{13.45}$$

Let $P^{(g)}(r,\tau,T)$ be the price at τ of a zero coupon bond maturing at T, and assume that for all $\tau \leq T$, $P^{(1)}(r,\tau,T) = \bar{P}_\tau < P^{(0)}(r,\tau,T)$ and that the recovery value is independent of v.[19] In the absence of arbitrage, the *pre-default* bond price $P^{\text{pre}}(r,\tau,T) \equiv P^{(0)}(r,\tau,T)$ satisfies

$$\begin{aligned} 0 &= \left(\frac{\partial}{\partial\tau} + L - r\right) P^{(0)}(r,\tau,T) + v(r) \cdot \left[P^{(1)}(r,\tau,T) - P^{(0)}(r,\tau,T)\right] \\ &= \left(\frac{\partial}{\partial\tau} + L - (r + v(r))\right) P^{(0)}(r,\tau,T) + v(r)\bar{P}_\tau, \end{aligned} \tag{13.46}$$

for all $\tau \in [t,T)$, with the boundary condition $P^{(0)}(r,T,T) = 1$. The solution is

$$P^{\text{pre}}(r_t,t,T) = \mathbb{E}_t^*\left(e^{-\int_t^T (r_\tau + v(r_\tau))d\tau}\right) + \mathbb{E}_t^*\left(\int_t^T e^{-\int_t^\tau (r_u + v(r_u))du} v(r_\tau)\bar{P}_\tau d\tau\right), \tag{13.47}$$

where $\mathbb{E}_t^*[\cdot]$ is the expectation taken with respect to only the first equation of system (13.45).

Eq. (13.47) can be simplified[20] by defining a percentage loss process $l_\tau \in [0,1] : \bar{P} = (1-l)P$, which inserted into Eq. (13.46) leaves a partial differential equation the solution of which is, by Feynman-Kac,

$$P^{\text{pre}}(r_t,t,T) = \mathbb{E}_t^*\left(e^{-\int_t^T (r_\tau + l_\tau \cdot v(r_\tau))d\tau}\right).$$

Finally, it is intuitive that the pre-default bond price P^{pre} is decreasing in the default intensity v, and Appendix 5 provides a few analytical details and assumptions under which this property does hold.

13.5 Multifactor models of the short-term rate

The empirical evidence reviewed in Section 13.3.4 suggests that one-factor models cannot explain the entire variation of the term-structure of interest rates. We need at least three factors. This section contains a succinct review regarding the standard approaches that deal with this multi-dimensionality.

13.5.1 Stochastic volatility

13.5.1.1 Convexity

How does volatility affect the yield curve? Consider the following two-period example. In the first period, the short-term rate is r and in the second, it is either $\tilde{r} = r+d$ or $\tilde{r} = r-d$ with equal probability, where $d > 0$. The price of a two-period bond is $P(r,d) = m(r,d)/(1+r)$, where $m(r,d) = E(1/(1+\tilde{r}))$ is the discount factor expected to prevail at next period. By Jensen's inequality, $m(r,d) > 1/(1+E(\tilde{r})) = 1/(1+r) = m(r,0)$. That is, two-period bond prices

[19] Or in case it depends on v, it satisfies $\bar{P}_\tau(v') \geq \bar{P}_v(v)$ for $v' \geq v$. This assumption guarantees that default-free bond prices are higher than defaultable, as shown in Appendix 5.

[20] See Duffie and Singleton (1999, Eq. (10) p. 696).

increase upon activation of randomness. More generally, two-period bond prices are always increasing in the "volatility" parameter d in this example, as illustrated by Figure 13.7.

The intuition underlying Figure 13.7 is standard: the bond price is decreasing and convex in the short-term rate, such that the price is increasing in the interest rate volatility—the price drop in "bad times" (i.e. when the interest rate increases), is less than the price increase in "good times."[21]

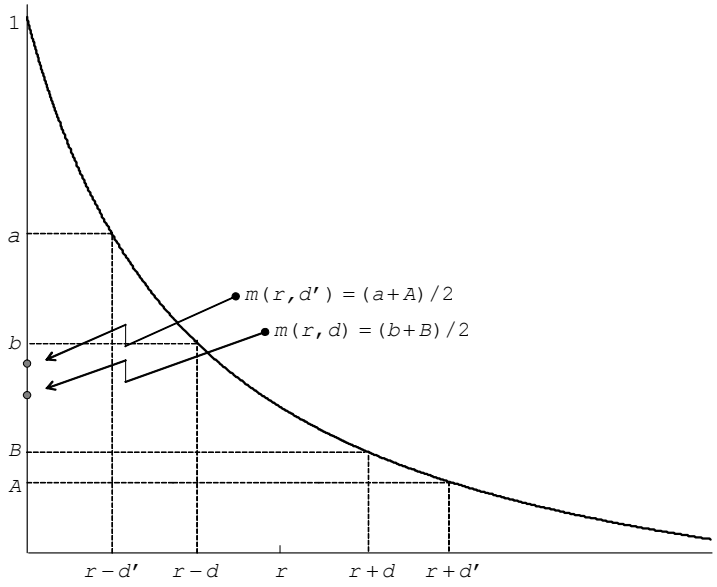


FIGURE 13.7. If the risk-neutralized interest rate of the next period is either $\tilde{r} = r + d$ or $\tilde{r} = r - d$ with equal probability, the discount factor $1/(1+\tilde{r})$ is either B or b with equal probability. Hence $m(r,d) = E[1/(1+\tilde{r})]$ is the midpoint of $\overline{bB}$. Similarly, if volatility is $d' > d$, $m(r,d')$ is the midpoint of $\overline{aA}$. Since $\overline{ab} > \overline{BA}$, it follows that $m(r,d') > m(r,d)$ such that the two-period bond price satisfies $P(r,d) = m(r,d)/(1+r)$ satisfies: $P(r,d') > P(r,d)$ for $d' > d$.

These properties are due to the assumption that the expected short-term rate is independent of d. They may well break down in alternative settings. For example, consider a market in which an upward rate movement is more likely than a downward. As a second example, consider a multiplicative setting, in which either $\tilde{r} = r(1+d)$ or $\tilde{r} = r/(1+d)$ with equal probability. It can be shown that in these two examples, bond prices are decreasing in volatility for short maturities, and increasing for longer, a property originally illustrated by Litterman, Scheinkman and Weiss (1991). Below, it is argued that, due to risk-aversion, changes in the expected short-term rate may well depend on the "volatility" parameter, d. Then, at short maturities, risk-aversion dominates the convexity effects in Figure 13.7, whereas convexities dominate over longer maturities. We now build on this intuition and explain the relation between interest rate volatility and the yield curve.

[21]This property relates to the theory of mean-preserving spreads and convex payoffs explained in Chapters 1, 8 and 11. That is, let $\tilde{m}_d(\tilde{r}) = 1/(1+\tilde{r})$ denote the random discount factor, such that $x \mapsto -\tilde{m}_d(x)$ is increasing and concave and, hence, $d' < d'' \Longrightarrow E(-\tilde{m}_{d''}(x)) < E(-\tilde{m}_{d'}(x))$, just as Figure 12.7 illustrates.

13.5.1.2 Two-factor models

In the CIR model, the instantaneous volatility of the short-term rate is stochastic, depending as it does on the level of r_τ, which is obviously stochastic. However, empirical evidence suggests that the short-term rate volatility depends on some additional factors, as discussed in Section 13.4.

A natural extension of the CIR model is one in which the instantaneous volatility of the short-term rate depends on (i) the level of the short-term rate (as in the CIR model), and (ii) some additional random component. This additional component may be what we refer to as "stochastic volatility" of the short-term rate. It is the term-structure counterpart to the stochastic volatility extension of the Black and Scholes (1973) model (see Chapter 11).

Fong and Vasicek (1991) develop the first model in which the volatility of the short-term rate is stochastic. They assume that the short-term rate r_τ is solution to

$$\begin{array}{l} dr_\tau = \kappa_r\left(\bar{r} - r_\tau\right)d\tau + \sqrt{v_\tau}r_\tau^\gamma dW_{1\tau} \\ dv_\tau = \kappa_v\left(\bar{v} - v_\tau\right)d\tau + \xi_v\sqrt{v_\tau}dW_{2\tau} \end{array} \tag{13.48}$$

where κ_r, $\bar{r}$, κ_v, $\bar{v}$ and ξ_v are constants, and $[W_1\ W_2]$ is a vector Brownian motion under the physical probability.[22]

The authors assume that $\gamma = 0$ and that the unit-risk-premia that compensate for fluctuations in the short-term rate and its volatility are both proportional to $\sqrt{v_\tau}$, being $\lambda_r\left(v_\tau\right) \equiv \lambda_1\sqrt{v_\tau}$ and $\lambda_v\left(v_\tau\right) \equiv \lambda_2\sqrt{v_\tau}$, for two constants λ_1 and λ_2.[23] These assumptions lead to a solution for the price of a zero coupon bond given by

$$P\left(r_t, v_t, T-t\right) \equiv e^{A(T-t)-B(T-t)r_t+C(T-t)v_t}, \tag{13.49}$$

for three functions $A\left(T-t\right), B\left(T-t\right), C\left(T-t\right)$ characterized in Appendix 6 (see Eqs. (13A.8)).

Due to randomness in the short-term rate volatility, bond return volatility now also fluctuates randomly, driven by sources of variation additional to r_t, i.e., v_t. It is possible to show that short-term rate loading, $B\left(\cdot\right)$, is the same as that in the Vasicek model (see Eq. (13.37)) and, thus, is always positive. Instead, the volatility loading $C\left(\cdot\right)$ can take on any sign. It is possible to show that in absence of risk-aversion corrections regarding r_τ (i.e., $\lambda_1 = 0$), C is positive, such that the whole yield curve lowers as volatility increases. However, when $-\lambda_1$ is large enough, the whole yield curve increases after an increase in volatility. Finally, for moderate values of $-\lambda_1$, $C\left(\cdot\right)$ can take both signs and is typically negative for shorter maturities and positive for longer: the short-end part of the yield curve increases in volatilty while the longer-end part decreases. The next subsection provide intuition on these properties.

Longstaff and Schwartz (1992) propose another model in which the short-term rate exhibits stochastic volatility, making bond returns exhibit stochastic volatility driven by sources of variation additional to r_t. They rely on an equilibrium model reviewed in Chapter 5, in which the short-term rate is a linear combination of two factors,

$$r_\tau = \beta_1 y_{1\tau} + \beta_2 y_{2\tau}, \tag{13.50}$$

for two constants β_i and two uncorrelated processes $y_{i\tau}$ that satisfy the following square-root dynamics:

$$dy_{i\tau} = \hat{\kappa}_i(\bar{y}_i - y_{i\tau})d\tau + \sqrt{y_{i\tau}}d\tilde{W}_{it}, \tag{13.51}$$

[22] In Appendix 5, we consider an extension of this model, in which r_τ and v_τ are instantaneously correlated.

[23] Thus, by Girsanov theorem, $d\tilde{W}_{i\tau} = dW_{i\tau} + \lambda_i\sqrt{v_\tau}d\tau$ is a Brownian motion under P.

where $\tilde{W}_{it}$ are standard Brownian motions under the risk neutral probability and $\hat{\kappa}_i$ and $\bar{y}_i$ are two constants (see Eqs. (5.104)-(5.105) in Chapter 5). It is easy to see that the bond price predicted by this model is

$$P\left(y_{1t}, y_{2t}, T-t\right) = \mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau}\right) = P^{\text{cir}}(\tilde{y}_{1t}, T-t; \theta_1) P^{\text{cir}}(\tilde{y}_{2t}, T-t; \theta_2), \tag{13.52}$$

where $\tilde{y}_{i\tau} \equiv \beta_i y_{i\tau}$ are solutions to

$$d\tilde{y}_{i\tau} = \hat{\kappa}_i(\hat{y}_i - \tilde{y}_{i\tau})d\tau + \hat{\sigma}_i\sqrt{\tilde{y}_{i\tau}}d\tilde{W}_{it},$$

with $\hat{y}_i \equiv \beta_i \bar{y}_i$, $\hat{\sigma}_i \equiv \sqrt{\beta_i}$ and $P^{\text{cir}}\left(\cdot, T-t, \theta_i\right)$ are the prices predicted by the CIR model (see Eqs. (13.41)) when the parameter vector under the risk-neutral probability is $\theta_i \equiv [\hat{\kappa}_i \ \hat{y}_i \ \hat{\sigma}_i]$.

Note that by Eqs. (13.50) and (13.51), the instantaneous variance of the short term rate is

$$v_\tau d\tau \equiv var\left(dr_\tau\right) = \left(\beta_1^2 y_{1\tau} + \beta_2^2 y_{2\tau}\right) d\tau,$$

such that we can now use the previous equation and Eq. (13.50) and express $\tilde{y}_{1\tau}$ and $\tilde{y}_{2\tau}$ in terms of r_τ and v_τ,

$$\tilde{y}_{1\tau} = \frac{\beta_2 r_\tau - v_\tau}{\beta_2 - \beta_1}, \quad \tilde{y}_{2\tau} = \frac{v_\tau - \beta_1 r_\tau}{\beta_2 - \beta_1}.$$

Therefore, Eq. (13.52) tells us that in this model, the price of a zero coupon bond is the same as that in Eq. (13.49), for some functions $A\left(\cdot\right)$, $B\left(\cdot\right)$ and $C\left(\cdot\right)$. Section 13.5.3 explains the reasons both the Fong and Vasicek model and this model predict the bond price to have the same exponential-affine functional form in Eq. (13.49).

Note, finally, the important feature of the models with random volatility in this section. They predict that the yield curve is now a function $R\left(r_t, v_t, T-t\right)$, depending on the level of the short-term rate, r_t, and one additional factor, the instantaneous variance of the short-term rate, v_t. Therefore, these models predict that we now have two factors that help explain the whole term-structure of interest rates.

13.5.1.3 Volatility and the yield curve

How does the factor v_t affect the yield curve? Consider the basic Vasicek (1977) model. Naturally, this model assumes that volatility is constant, yet it could be used to develop intuition on Eqs. (13.48) and possibly other stochastic volatility models. It is possible to show that Eq. (13.36) implies that

$$\frac{\partial R\left(r_t, T-t\right)}{\partial \sigma} = -\frac{1}{T-t}\left(\sigma \int_t^T B^2\left(T-s\right) ds + \lambda \int_t^T B\left(T-s\right) ds\right). \tag{13.53}$$

where $B\left(T-s\right)$ is as in Section 13.4.4. Eq. (13.53) shows that if $\lambda \geq 0$, the whole term-structure is decreasing in σ, the short-term rate volatility. That is, bond prices increase in σ, a conclusion that parallels that for options, where option prices are increasing in the volatility of the asset price. As explained in Chapter 11, this property arises through the optionality of the contract—say the convexity of a European call price with respect to the asset price.

But the interesting properties arise in the empirically relevant case, $\lambda < 0$.[24] In this case, the sign of $\frac{\partial R(t,T)}{\partial \sigma}$ depends on both "convexity" and "slope" effects. "Convexity" effects, those relating to the second partial $\frac{\partial^2 P(r,T-t)}{\partial r^2} = P\left(r, T-t\right) B^2\left(T-t\right)$, arise through the term $\sigma \int_t^T B^2(T-$

[24] In this simple model, the assumption that $\lambda < 0$ is reasonable, as we observe positive risk-premia more often than negative risk-premia. But in this same model, $u_r < 0$, which together with $\lambda < 0$, ensures that term-premia are positive.

$s)ds$. "Slope" effects, those relating to $\frac{\partial P(r,T-t)}{\partial r} = P(r,T-t)\,B(T-t)$, arise, instead, through the term $\int_t^T B(T-s)\,ds$. If λ is negative, and sufficiently large in absolute value, slope effects dominate convexity effects, and the term-structure can actually increase in σ. For intermediate values of λ, the term-structure can be both increasing and decreasing in σ. At short maturities, the convexity effects in Eq. (13.53) are typically dominated by slope effects, and the short-end of the term-structure can be increasing in σ. At longer maturity dates, however, convexity effects are more important and, sometimes, dominate slope effects.

More generally, changes in interest rate volatility are not mean-preserving spreads for the risk-neutral distribution, as Eq. (13.53) illustrates for the Vasicek model. In a world with complete markets, say Black-Scholes, the asset underlying the derivative contracts is traded. In the case under study, the short-term rate is not a traded risk. Therefore, its risk-neutral drift depends on volatility through risk-adjustements: to illustrate, in the Vasicek example, this dependence arises through the risk-premium parameter, λ.

While Eq. (13.53) relies on a model with constant volatility, the reasoning underlying its interpretation holds even when volatility is random.[25] In particular, suppose that the risk-premium required to bear the interest rate risk is negative and sufficiently large in absolute value. In this case, slope effects may dominate convexity ones at any maturity date, such that the whole yield curve, now, could be always increasing in volatility. Let us provide some intuition. It is reasonable to expect that in bad times (i.e., when interest rate volatility is high) risk-premium effects dominate over convexity, such that the yield curve shifts up following an increase in volatility. However, in good times, we would expect that convexity dominates, with the yield curve being decreasing in volatility. Thus, in these examples, if risk-premiums are sufficiently sensitive to volatility, we would expect that in good times, when volatility is small, convexity effects dominate and the yield curve lowers as volatility increases. In bad times, when volatility is high, we would expect that risk-premium effects dominate, such that the yield curve increases following an increase in volatility.

To illustrate, consider the Vasicek model again, and assume that the risk-premium is $\lambda = \bar{\lambda}\sigma^3$, for some constant $\bar{\lambda}$. This functional form of the risk-premium ensures that the risk-premium is quite small when σ is small, although then it substantially increases in bad times, i.e. when σ gets larger and larger. With this risk-premium, Eq. (13.53) is:

$$\frac{\partial R(r_t,T-t)}{\partial \sigma} = -\frac{1}{T-t}\sigma\left(\int_t^T B^2(T-s)\,ds + \bar{\lambda}\sigma^2\int_t^T B(T-s)\,ds\right). \tag{13.54}$$

That is, risk-premium effects become more and more relevant as σ increases. The previous equation reveals that we may also define a threshold value for σ such that convexity effects are exactly offset by risk-premium effects. Eq. (13.54) shows that for each time to maturity $T-t$, there exists a value of σ depending on $T-t$, say $\hat{\sigma}(T-t)$ such that the partial $\frac{\partial R(r_t,T-t)}{\partial \sigma} = 0$. We might go on and define an average value of $\hat{\sigma}(T-t)$, say $\hat{\sigma}_* \equiv \mathcal{T}^{-1}\int_0^{\mathcal{T}}\hat{\sigma}(u)\,du$, where $\mathcal{T}$ denotes the highest time-to-maturity we want to consider. This threshold value, $\hat{\sigma}_*$, is the one that might lead to a definition of what good or bad times can be—in terms of the term-structure implications of a volatility shock.

How do we interpret these properties in light of the factors dynamics reviewed in Section 13.3? Clearly, the very short-end of yield curve is not affected by movements in volatility, as $\lim_{T\to t} R(r_t,v_t,T-t) = r_t$, for all v_t. Moreover, these models predict that $\lim_{T\to\infty} R(r_t,v_t,T-t)$

[25] See Mele (2003).

$= \bar{R}$, where $\bar{R}$ is a constant and, hence, independent of v_t. Therefore, movements in the short-term volatility can only affect the middle portion of the yield curve. For example, if the risk-premium required to bear the interest rate risk is negative and sufficiently large, an upward movement in v_t can produce an effect on the yield curve qualitatively similar to that depicted in Figure 13.2 ("Curvature" panel), and would thus roughly mimic the "curvature" factor that we reviewed in Section 13.3.

13.5.2 Three-factor models

We need at least three factors to explain the entire variation of the yield curve. A natural extension of the model in Eqs. (13.48) is one in which the drift of the short-term rate contains some predictable component, $\bar{r}_\tau$, such that the yield curve is driven by the following three factor model:

$$\begin{cases} dr_\tau = \kappa_r \left(\bar{r}_\tau - r_\tau\right) d\tau + \sqrt{v_\tau} r_\tau^\gamma dW_{1\tau} \\ dv_\tau = \kappa_v \left(\bar{v} - v_\tau\right) d\tau + \xi_v \sqrt{v_\tau} dW_{2\tau} \\ d\bar{r}_\tau = \kappa_{\bar{r}} \left(\bar{\imath} - \bar{r}_\tau\right) d\tau + \xi_{\bar{r}} \sqrt{\bar{r}_\tau} dW_{3\tau} \end{cases} \tag{13.55}$$

where $\kappa_r, \gamma, \kappa_v, \bar{v}, \xi_v, \kappa_{\bar{r}}, \bar{\imath}$ and $\xi_{\bar{r}}$ are constants, and $[W_1\ W_2\ W_3]$ is vector Brownian motion.

Balduzzi et al. (1996) develop the first model in which the drift of the short-term rate changes stochastically, as in Eqs. (13.55). Dai and Singleton (2000) estimate a number of models that generalize that in Eqs. (13.55) (See Section 13.5.6 for details on the estimation strategy). The term-structure implications of these models can be understood very simply. First, under regularity conditions, the yield curve is now a function $R\left(r_t, \bar{r}_t, v_t, T-t\right)$. Second, and intuitively, changes in the new factor $\bar{r}_t$ should primarily affect the long-end of the yield curve. This is because empirically, the usual finding is that the short-term rate reverts relatively quickly to the long-term factor $\bar{r}_\tau$ (i.e., κ_r is relatively large), where $\bar{r}_\tau$ mean-reverts slowly (i.e., $\kappa_{\bar{r}}$ is relatively low). This mechanism makes the short-term rate quite persistent anyway. Ultimately, then, the slow mean-reversion of $\bar{r}_\tau$ means that changes in $\bar{r}_\tau$ last for the relevant part of the term-structure we are usually interested in (i.e. up to 30 years), despite the fact that $\lim_{T\to\infty} R\left(r_t, \bar{r}_t, v_t, T-t\right)$ is independent of the three factors, r_t, $\bar{r}_t$ and v_t.

However, it is difficult to see how to reconcile such a behavior of the long-end of the yield curve with the existence of any of the factors discussed in Section 13.3. First, the short-term rate cannot be taken as a "level factor," since we know its effects die off relatively quickly. Instead, a *joint* change in both the short-term rate, r_t, and the "long-term" rate, $\bar{r}_t$, should be really needed to mimic the "Level" panel of Figure 13.2. However, this interpretation is at odds with the assumption that the factors underlying the exercises in Figure 13.2 are uncorrelated![26]

Finally, to emphasize how exacerbated these puzzles are, consider the effects of changes in the short-term rate r_t. We know that the long-end of the term-structure is not affected by movements of the short-term rate. Hence, the short-term rate acts as a "steepness" factor, as in Figure 13.2 ("Slope" panel). However, this interpretation is restrictive, as factor analysis reveals that the short-end and the long-end of the yield curve move in opposite directions after a change in the steepness factor. Here, instead, a change in the short-term rate only modifies the short-end (and possibly a small portion of the middle) of the yield curve and, hence, does not produce any variation in the long-end curve.

[26] Empirical results in Dai and Singleton reveal that, if any, r_t and $\bar{r}_t$ are negatively correlated.

13.5.3 Affine and quadratic term-structure models

13.5.3.1 Affine

The Vasicek (1977) and CIR (1985) models predict that the price of a zero coupon bond is exponential-affine in the short-term rate r. Likewise, the models with stochastic volatility of Fong and Vasicek (1991) and Longstaff and Schwartz (1992) predict that the price of a zero is exponential-affine in the short-term rate r and its instantaneous variance v. These properties are the expression of a general phenomenon. For the simple one-factor case, it is indeed possible to show that bond prices are exponential-affine in r if and only if the functions b and a^2 are affine in r. Models that satisfy these conditions are known as *exponential-affine models*. These basic results extend to *multifactor* models (such as those of Fong and Vasicek or Longstaff and Schwartz), in which case bond prices are exponential-affine in the state variables.[27] In these models, the short-term rate can always be written as a function $r(y)$ such that

$$r(y) = r^0 + r^1 \cdot y, \tag{13.56}$$

where r^0 is a constant, r^1 is a vector of constants, and y is a n-dimensional diffusion, solution to

$$dy_\tau = \kappa(\mu - y_\tau)\, d\tau + \Sigma V(y_\tau)\, dW_\tau, \tag{13.57}$$

where W is a d-dimensional Brownian motion, Σ is a full rank $n \times d$ matrix, and V is a full rank $d \times d$ diagonal matrix with elements,

$$V(y)_{(ii)} = \sqrt{\alpha_i + \beta_i^\top y}, \quad i = 1, \cdots, d, \tag{13.58}$$

for some scalars α_i and vectors β_i. Langetieg (1980) appears to be the first reference in the literature containing the first multifactor model constructed along these lines (in which, notably, $\beta_i = 0$).

Next, Let $V^-(y)$ be a $d \times d$ diagonal matrix with elements

$$V^-(y)_{(ii)} = \begin{cases} \frac{1}{V(y)_{(ii)}} & \text{if } \Pr\{V(y_t)_{(ii)} > 0 \text{ all } t\} = 1 \\ 0 & \text{otherwise} \end{cases}$$

and set

$$\Lambda(y) = V(y)\lambda_1 + V^-(y)\lambda_2 y, \tag{13.59}$$

for some d-dimensional vector λ_1 and some $d \times n$ matrix λ_2. Duffie and Kan (1996) explain the benefits of this model. In their formulation $\lambda_2 = 0_{d\times n}$, and the bond price is exponential-affine in the state variables y. That is, the price of the zero has the following functional form:

$$P(y, T-t) = e^{\mathcal{A}(T-t)+\mathcal{B}(T-t)\cdot y}, \tag{13.60}$$

for some functions $\mathcal{A}(\cdot)$ and $\mathcal{B}(\cdot)$ ($\mathcal{B}(\cdot)$ is vector-valued) satisfying $\mathcal{A}(0) = 0$ and $\mathcal{B}(0)_{(i)} = 0$ and characterized in Appendix 6.

Duffee (2002) suggested the general functional form for Λ in Eq. (13.59). He explained that in bond markets, risk-premiums seem to relate to both *volatility* and *level* of the fundamentals.

[27] More generally, we say that affine models are those that make the characteristic function exponential-affine in the state variables. In the case of the multifactor interest rate models of the previous section, this condition is equivalent to the condition that bond prices are exponential affine in the state variables.

In this model, risk-premiums reduce to $\Lambda(y)V(y) = V^2(y)\lambda_1 + \lambda_2 y$. Thus, the inclusion of the term $\lambda_2 y$ allows one to model the statistical relations linking risk-premiums to fundamentals. Interestingly, bond prices still have an exponential affine form, just as in Eq. (13.60). When $\lambda_2 = 0_{d\times n}$, we say that the model is "completely affine," and "essentially affine," otherwise. The clear advantage of these affine models, then, is that they considerably simplify statistical inference, as explained in Section 13.5.5 below.

Exponential-affine models have been extensively utilized, and surveying them further will fall well beyond the scope of these lectures. Piazzesi (2010) provides an extensive survey of these models.

13.5.3.2 Quadratic

Affine models are known to impose tight conditions on the structure of the volatility of the state variables. These restrictions arise to keep the square root in Eq. (13.58) real valued. But these constraints may hinder the actual performance of the models. There exists another class of models, known as quadratic models (see Ahn, Dittmar and Gallant, 2002), which partially overcome these difficulties while preserving analytical tractability.

13.5.3.3 No-arbitrage regressions and the macroeconomy

Ang and Piazzesi (2003) and Hördahl, Tristani and Vestin (2006) (HTS, henceforth) introduce "no-arbitrage" regressions, to model the relations linking macroeconomic variables to the yield curve. In their models, the factors are taken to be a discrete-time version of Eq. (13.57), where some components of y are observable, and others are unobservable. The observables relate to macroeconomic factors such as inflation or industrial production. The authors, then, study how all these factors affect the yield curve, predicted by a pricing equation such as that in Eq. (13.60). While HTS have a structural model of the macroeconomy, Ang and Piazzesi (2003) have a reduced-form model.

Reduced-form models can be exposed to the critique that some of the parameters are not "variation-free." [**Explain what variation-free parameters are, in mathematical statistics**] Consider, for example, the simple Lucas model in Part I of the lectures, which predict that the short-term rate is $r = \rho + \gamma\mu - \frac{1}{2}\sigma^2\gamma(1+\gamma)$: a change in the risk-aversion parameter γ determines a change in the interest rate. This simple example shows that the parameters related to risk-aversion correction in Eq. (13.59) are not free, in that tilting them has an effect on the parameters of the factor dynamics in Eq. (13.57).

The advantage of reduced-form model is that they provide us with a great deal of flexibility because they don't force the model to track any theoretic-based economy. Moreover, we can always find a theoretical market supporting the no-arb market underlying the reduced-form model. No-arb regressions such as those in AP give the data the power to say which parameter constellation make the model likely to perform, without imposing theoretical restrictions which the data might, then, be likely to reject. For example, the Lucas model, while clearly illustrates that some of the parameters are not variation-free, can be simply wrong, and might impose unreasonable restrictions on the data. For no-arb models, instead, cross-equations restrictions arise through the weaker requirement of absence of arbitrage.

13.5.4 Unspanned stochastic volatility

Are fixed income markets incomplete? Mele and Obayashi (2015) argue that fixed income volatility is quite distinct from equity. Consequently, many of the investable products on the

popular gauge of equity volatility, VIX (see Chapter 11), could only be poor surrogates for exposure to fixed income volatility. The "uniqueness" of fixed income volatility has actually been widely acknowledged in the literature. It is well-known since at least Collin-Dusfresne and Goldstein (2002) that fixed income market volatility does not appear to be priced only based on existing fixed income assets.

Simply, the authors showed that straddle returns on caps and foors (see Section 13.8) cannot be explained by changes in the term structure of swap rates, but by other factors. That is, existing fixed income assets (such as bonds) do not help hedge rate volatility. The models proposed by Collin-Dusfresne and Goldstein to address these issues are known as leading to "unspanned stochastic volatility" (USV, in the sequel). The reason for this terminology is the following. Consider any of the stochastic volatility models reviewed in Chapter 11, in which a stock price S_t is solution to

$$\frac{dS_t}{S_t} = \mu dt + \sigma_t dW_t, \tag{13.61}$$

under the physical probability, for some constant μ, a Brownian motion W_t, and a volatility process $(\sigma_t)_{t\geq 0}$ possibly driven by W_t *and* other Brownian motions. One example of these models is the celebrated Heston's (1993) model, in which σ_t^2 is a square root process.

As we know from Chapter 11, these models are typically understood to describe a situation of incomplete markets. Collin-Dusfresne and Goldstein (2002) propose to extend this notion to fixed income markets, by modeling bond prices in an incomplete markets setup. The idea is simple. If bond markets have stochastic volatility and are still incomplete, bond prices should satisfy dynamics in which their instantaneous returns have stochastic volatility, similarly as in Eq. (13.61) for the equity case. At the same time, bond prices exposure to volatility should be zeroed, just as a stock price exposure to its own volatility is zero in the context of Chapter 11. In the context of the short-term rate models of this section, the latter condition is met when

$$\frac{\partial}{\partial v} P\left(r_t, \bar{r}_t, v_t, T-t\right) = 0. \tag{13.62}$$

Collin-Dufresne and Goldstein (2002) provide a characterization of the previous condition in multidimensional models. It is still an open question whether existing models with USV compare to models without. In a nutshell, restrictions such as that in Eq. (13.62) can impede standard multifactor models to perform as accurately as they could do without restrictions. However, USV and incompleteness of fixed income markets both seem to be a quite robust and well-accepted empirical feature.

13.5.5 Topics regarding estimation and trading strategies

13.5.5.1 Univariate models

How to estimate the parameter vector $\theta = [\kappa, \mu, \eta, \beta, \alpha, \xi, \vartheta, \rho]^\top$ in Eqs. (13.43)? Maximum Likelihood (ML) would be a feasible estimation method under two sets of conditions. *First*, the model in Eqs. (13.43) should not have stochastic volatility at all, viz, $\beta = \xi = 0$; in this case, the short-term rate would be solution to,

$$dr_t = \kappa\left(\mu - r_t\right) dt + \bar{\sigma} r_t^{\eta} dW_t,$$

where $\bar{\sigma}$ is now a constant. *Second*, the value of the elasticity parameter η is important. If $\eta = 0$, the short-term rate process is the Gaussian one proposed by Vasicek (1977). If $\eta = \frac{1}{2}$, we obtain

the square-root process of Cox, Ingersoll and Ross (1985). As we know, the transition density of r_t is Gaussian in the Vasicek market, and a noncentral chi-square in the CIR case. Therefore, in both Vasicek and CIR markets, we may write down the likelihood function of the diffusion process. Therefore, ML estimation is possible in these two cases. In more general cases, such those in the next section, one needs to go for simulation methods, such as those described in Chapter 6. However, we could still estimate multifactor affine models through ML.

13.5.5.2 More general models

Estimating the model in Eqs. (13.43) is certainly instructive. Yet a more important question is to examine the term-structure implications of this model. More generally, how would the estimation procedure outlined in the previous subsection change if the task is to estimate a Markov model of the term-structure of interest rates? There are three steps.

Step 1

Collect data on the term structure of interest rates. We will need to use data on three maturities, say a time series of riskless 6 month, 5 year and 10 year yields.

Step 2

Let us consider the three-factor model in Eqs. (13.55) of Section 13.5.2, where the three Brownian motions W_i are now allowed to be correlated. The bond price predicted by this model is:

$$P^j\left(r_t,v_t,\bar{r}_t\right)\equiv P\left(r_t,v_t,\bar{r}_t,N_j-t\right)=\mathbb{E}\left(e^{-\int_t^{N_j}r_s ds}\middle|\, r_t,v_t,\bar{r}_t\right), \tag{13.63}$$

where N_j is a sequence of expiration dates. Naturally, this price depends on the risk-aversion corrections needed to turn the dynamics in Eqs. (13.55) into risk-neutral ones. As discussed, one may impose analytically convenient conditions on the risk-adjustments, but we do not need to be more precise at this juncture.

No matter the risk-adjustments, we have that they entail that Eq. (13.63) depends not only on the "physical" parameter vector $\boldsymbol{\theta}=\left[\kappa_r,\kappa_v,\kappa_{\bar{r}},\gamma,\bar{v},\xi_v,\bar{\iota},\xi_{\bar{r}},\rho\right]^\top$, where ρ is a vector containing all the correlation coefficients of W_i, but also on these very same risk-adjustment parameter vector, say $\boldsymbol{\lambda}$. Precisely, the Radon-Nikodym derivative of the risk-neutral probability with respect to the physical probability is $\exp\left(-\frac{1}{2}\int\left\|\boldsymbol{\Lambda}_t\right\|^2 dt-\int\boldsymbol{\Lambda}_t d\boldsymbol{Z}_t\right)$, for some vector Brownian motion $\boldsymbol{Z}$, and $\boldsymbol{\Lambda}_t\equiv\boldsymbol{\Lambda}^{\mathrm{m}}\left(r_t,v_t,\bar{r}_t;\boldsymbol{\lambda}\right)$, for some vector-valued function $\boldsymbol{\Lambda}^{\mathrm{m}}$ and some parameter vector $\boldsymbol{\lambda}$. The function $\boldsymbol{\Lambda}^{\mathrm{m}}$ makes risk-adjustment corrections depend on the current value of the state vector $[r_t,v_t,\bar{r}_t]$, which makes the model Markov, thereby simplifying statistical inference.

To summarize, the issue is now one where we need to estimate both the physical parameter vector $\boldsymbol{\theta}$ and the "risk-adjustment" parameter vector $\boldsymbol{\lambda}$. Next, we consider the yield curve in correspondence of three maturities,

$$R^j\left(r_t,v_t,\bar{r}_t;\boldsymbol{\theta},\boldsymbol{\lambda}\right)\equiv-\frac{1}{N_j}\ln P^j\left(r_t,v_t,\bar{r}_t\right),\quad j=1,2,3, \tag{13.64}$$

where the notation $R^j\left(r,v,\bar{r};\boldsymbol{\theta},\boldsymbol{\lambda}\right)$ emphasizes that the theoretical yield curve depends on the parameter vector $(\boldsymbol{\theta},\boldsymbol{\lambda})$. We can now use actual data, $R^j_{\$}$ say, and the model predictions about the data, R^j, create moment conditions, and proceed to estimate the parameter vector $(\boldsymbol{\theta},\boldsymbol{\lambda})$ through some method of moments—provided of course the moments are enough to make $(\boldsymbol{\theta},\boldsymbol{\lambda})$ identifiable.

We encounter two difficulties. The first is that the volatility v_t and the long-term drift of the short-term rate, $\bar{r}_t$, are not observable. We can use inference methods based on simulations to cope with this issue. Very simply, we simulate Eqs. (13.55), and apply moment conditions or auxiliary models to observable variables, according to the procedure set forth in Chapter 6. For example, we simulate Eqs. (13.55) for a given value of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. For each simulation, we reconstruct a time series of interest rates R^j relying on Eq. (13.64). Then, we use these simulated data to create moment conditions or fit some auxiliary model to these artificial data that is as close as possible to the very same auxiliary model fit to real data. The parameter estimator is the value of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ minimizing some norm of these moment conditions, obtained through the simulations, with any of the methods explained in Chapter 6. By results in Chapter 6 (see Theorem 6.4), fitting a sufficiently rich auxiliary model should result in an efficient estimator.

A second difficulty is that the bond pricing formula in Eq. (13.63) does not generally admit a closed-form, an issue we can address using affine models, as explained next.

Step 3

Affine models are relatively easy to estimate. They imply that,

$$R^j\left(r_t, v_t, \bar{r}_t; \boldsymbol{\theta}, \boldsymbol{\lambda}\right) = A\left(j; \boldsymbol{\theta}, \boldsymbol{\lambda}\right) + \boldsymbol{B}\left(j; \boldsymbol{\theta}, \boldsymbol{\lambda}\right) \cdot \left[r_t, v_t, \bar{r}_t\right]^\top, \quad j = 1, 2, 3, \tag{13.65}$$

where $A\left(j; \theta, \lambda\right)$ and $\boldsymbol{B}\left(j; \theta, \lambda\right)$ are some functions of the maturity N_j ($\boldsymbol{B}$ is vector valued), and generally depend on the parameter vector $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. Once Eqs. (13.55) are simulated, a time series of yields R^j is then straightforward to determine based on Eq. (13.65).[28]

13.5.5.3 Filtering (and trading)

Once we estimate the model's parameters, we could attempt to infer the state $[r_t, v_t, \bar{r}_t]^\top$. For example, we could invert Eq. (13.65) given an observation of three yields having three different maturities. The main conceptual difficulty with this approach is that the "estimates" of the state $[r_t, v_t, \bar{r}_t]^\top$ rely on the maturities we choose. Changing maturities likely leads to different filtered states. The usual, and admittedly pragmatic, assumption is that Eq. (13.65) only holds with some additional observation/model error, and proceed, then, to find the state that minimizes the error variance using all the available maturities. This procedure delivers the state for each observation, and it is attractive, as it exploits market-based information summarized by the cross-section of yields.

Such a "pricing-model-based" procedure to filter the state has the potential to be used, in practice, to implement forecasting exercises. We can fit a VAR to the time-series of the filtered values of $[r_t, v_t, \bar{r}_t]^\top$ and, then, use this VAR to produce forecasts of the state and, then, forecasts of yields using Eq. (13.65), similarly as we could do in the case of options on equities in markets with stochastic volatility (see Chapter 11).

13.6 No-arbitrage models: early formulations

Pricing interest rate derivatives necessitates to take the initial yield as *given* rather than to *explain* it. To illustrate, consider a European option written on a bond: only "explaining" the

[28] Dai and Singleton (2000) implement this estimation strategy, although they make use of data on swap rates. The models they consider predict theoretical values for the swap rates, obtained through the formula in Eq. (13.113) of Section 13.7.5.4 below, where the bond prices in that formula are replaced by the pricing functions predicted by the models. Dai & Singleton consider three rates predicted by their models: two swap rates (with tenures of two and ten years), plus the six month Libor rate, $-\frac{1}{2}\ln P\left(t, t+\frac{1}{2}\right)$, where P is the pricing function predicted by the models they consider.

bond price is problematic as a model's error on the bond price can generate a large option price error due to the nonlinearities induced by the optionality. How can we trust an option pricing model, which is even not able to pin down the value of the underlying?

This section surveys the first attempts to deal with these issues within a continuous-time framework, with of one them actually being the continuous-time version of the Ho and Lee (1986) model in Chapter 12. In Sections 13.6.1, we explain the main issue while relying on a benchmark option pricing formula (derived in Section 13.6.2) and in Section 13.6.3, we provide details regarding two benchmark models.

13.6.1 Fitting the yield-curve, perfectly

Let again $P(r_\tau, \tau, S)$ be the price of a zero coupon bond maturing at some S. By no arbitrage, the price of a European call option on this bond, struck at K and expiring at $T < S$, is:

$$C^b(r_t, t, T, S) \equiv \mathbb{E}_t\left[e^{-\int_t^T r_\tau d\tau} \cdot (P(r_T, T, S) - K)^+\right].$$

In Section 13.8, we show that

$$C^b(r_t, t, T, S) = P(r_t, t, S) \cdot Q_F^S\left[P(r_T, T, S) \geq K\right] - KP(r_t, t, T) \cdot Q_F^T\left[P(r_T, T, S) \geq K\right], \quad (13.66)$$

where Q_F^X is a the forward probability for maturity X (see Eq. (13.100)).

The bond option price in Eq. (13.66) depends on *theoretical* prices, $P(r_t, t, T)$ and $P(r_t, t, S)$, not *market* prices. This issue is problematic to sell-side institutions while engaged in intermediating derivatives: as explained many times, the need arises in this case to simultaneously match the yield curve at the time of evaluation. This section describes models that fit the yield curve without errors, which we call "perfectly fitting models." These models are simply a more elaborated, continuous-time version of the no-arbitrage models introduced in Chapter 12. They predict that the price of any bond, say a bond expiring at some S, is, of course, random, at time $T < S$, but also exactly equal to the *current* market price, that of time t. Finally, and naturally, this price must be arbitrage-free. Aim of this and subsequent sections is to show how to achieve this task by augmenting the models seen in the previous sections with a set of "infinite dimensional parameters."

A final remark. Section 13.8 explains that at least for the Vasicek model, the option price in Eq. (13.66) does not explicitly depend on r because it only "depends" on $P(r_t, t, T)$ and $P(r_t, t, S)$. So why do we look for perfectly fitting models in the first place? Wouldn't it be enough, then, to just replace the theoretical prices $P(r_t, t, T)$ and $P(r_t, t, S)$ with the market values, say $P^\$(t, T)$ and $P^\$(t, S)$? This way, the model *is* perfectly fitting. Apart from being logically inconsistent (you would have a model predicting something generically different from prices), this way of proceeding also has practical drawbacks.

Section 13.8 reveals indeed that option pricing formulae for European options, might well agree "in notation" with those relating to perfectly fitting models. However, in Section 13.8.6, we explain that as we move towards more complex interest rate derivatives, say options on *coupon* bearing bonds and *swaption* contracts, the situation becomes dramatically different. Finally, some maturity dates might not be actually traded at some point in time. For example, $P^\$(t, T)$ might not be observed and still, we might well be interested in the pricing of exotic products requiring knowledge of $P^\$(t, T)$. An intuitive procedure to deal with this difficulty is to interpolate across the traded maturities. In fact, the objective of perfectly fitting models is to allow for such an "interpolation" while preserving absence of arbitrage.

The next two sections discuss two specific, old, and yet very famous examples of perfectly fitting models: (i) the Ho and Lee (1986) model, and (ii) one generalization of it, introduced by Hull and White (1990). In Section 13.7, we move on towards a general model-building principle that includes these two models as special cases.

13.6.2 Ho & Lee

Ho and Lee (1986) originally set their model in discrete-time, which is analyzed in the context of Chapter 12 along with alternative models. The model below, represents the "diffusion limit" of the original Ho & Lee model, as put forward in Chapter 12 (see Section 12.6.7), in which the short-term rate r_τ is solution to

$$dr_\tau = \theta_\tau d\tau + \sigma d\tilde{W}_\tau, \quad \tau \geq t, \tag{13.67}$$

where $\tilde{W}$ is a Brownian motion under Q, σ is a constant, and θ_τ is an "infinite dimensional" parameter, which we need to pin down the initial, observed yield curve, as we now explain. The reason we refer to θ_τ as infinite dimensional is that θ_τ is taken to be a continuous function of calendar time $\tau \geq t$. We assume this function is known at t—whence, "parameter."

Clearly, Eq. (13.67) is an affine model. Therefore, the bond price takes the following form,

$$P(r_\tau, \tau, T) = e^{A(\tau,T) - B(\tau,T)\cdot r_\tau}, \tag{13.68}$$

for two functions A and B to be determined below. It is easy to show that,

$$A(\tau, T) = \int_\tau^T \theta_s (s - T)\, ds + \frac{1}{6}\sigma^2 (T - \tau)^3, \quad B(\tau, T) = T - \tau.$$

Let $f_\$(t, \tau)$ denote the instantaneous, observed forward rate. By matching the instantaneous forward rate $f(\tau, T)$ predicted by the model to $f_\$(\tau, T)$ yields:

$$f_\$(t, T) = f(t, T) = -\frac{\partial \ln P(r_t, t, T)}{\partial T} = \int_t^T \theta_s ds - \frac{1}{2}\sigma^2 (T - t)^2 + r_t. \tag{13.69}$$

Because $P(t, T) = \exp(-\int_t^T f(t, \tau)\, d\tau)$, the drift term θ_s satisfying Eq. (13.69) guarantees an exact fit of the yield curve. By differentiating Eq. (13.69) with respect to T, leaves $\theta_T = \frac{\partial}{\partial T} f_\$(t, T) + \sigma^2 (T - t)$, or:[29]

$$\theta_\tau = \frac{\partial}{\partial \tau} f_\$(t, \tau) + \sigma^2 (\tau - t). \tag{13.70}$$

By Eqs. (13.67) and (13.70), the short-term rate is, then:

$$r_t = f_\$(0, t) + \frac{1}{2}\sigma^2 t^2 + \sigma \tilde{W}_t, \tag{13.71}$$

such that both the conditional expectation and conditional variance of the short-term rate explode over time. Moreover, by Eq. (13.69), and Eq. (13.67), the instantaneous forward rate satisfies:

$$d_\tau f(\tau, T) = -\theta_\tau d\tau + \sigma^2 (T - \tau)\, d\tau + dr_\tau = \sigma^2 (T - \tau)\, d\tau + \sigma d\tilde{W}_t. \tag{13.72}$$

[29] To verify that θ is indeed the fitting parameter we are searching for, we replace Eq. (13.70) into Eq. (13.69) and verify indeed that Eq. (13.69) holds as an identity.

Eq. (13.71) fully describes the essence of this model: the probability distribution of the short-term rate at any point in time is uniquely determined by absence of arbitrage: we did not make any assumptions regarding the drift of r_τ in Eq. (13.67). This is in sharp contrast with the models of the short-term rate, in which the risk-adjusted dynamics of the short-term rate under inherits much of the assumptions made regarding the dynamics under the physical probability.

The approach by which we infer the short-term rate dynamics from market prices is the continuous time analogue of that in Chapter 12, and for completeness, we mention again an interesting parallel already pointed out in Chapter 12: in terms of this particular model, we are searching for the function θ_τ that returns the initial yield curve in similarly to when in Chapter 11, we were calibrating a *local volatility* of the asset price to return the observed option prices. Section 13.7 generalizes this way of proceeding to account for much richer market dynamics, leading to the celebrated HJM framework. We now proceed to an extension of Ho & Lee model that makes the market stationary, in that interest rates are mean-reverting and conditional variances are bounded over time.

13.6.3 Hull & White

We now consider a model relatively more complex than that Ho & Lee. Why did we need a more complex model in the first place? After all, the Ho & Lee model can pin down the entire yield curve. The reason is that the yield curve may not be the only object that we want to fit. In general, we may want to price a large variety of derivatives, and fitting the yield curve may only be a sufficient working condition. For example, our goal could be to price complex interest rate derivatives while ensuring model predictions in line with some empirical evidence.

Hull and White (1990) consider a variety of models that can fit the initial yield curve without errors, and amongst them, their simplest is

$$dr_\tau = \kappa\left(\frac{\theta_\tau}{\kappa} - r_\tau\right)d\tau + \sigma d\tilde{W}_\tau, \tag{13.73}$$

where $\tilde{W}$ is a Q-Brownian motion, and κ, σ are constants. The model generalizes the Ho and Lee model (1986) in Eq. (13.67) and the Vasicek (1977) model in Eq. (13.32). In the original formulation of Hull and White, κ and σ are both time-varying, but the main points of this model can be learnt by working out this particularly simple case.

Eq. (13.73) also gives rise to an affine model. Therefore, the solution for the bond price is given by Eq. (13.68). It is easy to show that the functions A and B are given by

$$A(\tau, T) = \frac{1}{2}\sigma^2 \int_\tau^T B^2(s, T)\, ds - \int_\tau^T \theta_s B(s, T)\, ds, \tag{13.74}$$

and

$$B(\tau, T) = \frac{1}{\kappa}\left(1 - e^{-\kappa(T-\tau)}\right). \tag{13.75}$$

By reiterating the same reasoning produced to show (13.70), one shows (see Appendix 7 for details) that the solution for θ is

$$\theta_\tau = \frac{\partial}{\partial\tau} f_\$(t, \tau) + \kappa f_\$(t, \tau) + \frac{\sigma^2}{2\kappa}\left(1 - e^{-2\kappa(\tau-t)}\right). \tag{13.76}$$

That is, and contrary to the Ho & Lee model, mean-reversion ($\kappa > 0$) allows this model to keep the short-term rate stationary.

We now turn to a class of models that encompasses those in this section, in which the focus is to model forward rates in a way that prevents arbitrage, similarly to what we have done in this section, but without being restrictive regarding the particular relations linking bond prices to short-term rates (e.g., affine).

13.7 The Heath-Jarrow-Morton framework

13.7.1 Framework

The bond price representation in Eq. (13.10),

$$P(\tau,T) = e^{-\int_\tau^T f(\tau,\ell)d\ell}, \quad \text{all } \tau \in [t,T], \tag{13.77}$$

underlies the modeling approach started by Heath, Jarrow and Morton (1992) (HJM, henceforth). Given Eq. (13.77), this approach takes as a primitive the stochastic evolution of the entire structure of forward rates, not only the special case of the short-term rate, $r_t = \lim_{\ell \downarrow t} f(t,\ell) \equiv f(t,t)$. The goal is to start with Eq. (13.77), take the initial observed forward rates $(f(t,\ell))_{\ell \in [t,T]}$ as given, and, then, find the no-arb, "cross-equation" restrictions on the stochastic behavior of $(f(\tau,\ell))_{\tau \in (t,\ell]}$, for any $\ell \in [t,T]$.

By construction, the HJM approach allows for a perfect fit of the initial term-structure. This point can be illustrated quite simply, as the bond price $P(\tau,T)$ is

$$\begin{aligned} P(\tau,T) &= e^{-\int_\tau^T f(\tau,\ell)d\ell} \\ &= \frac{P(t,T)}{P(t,\tau)} \cdot \frac{P(t,\tau)}{P(t,T)} e^{-\int_\tau^T f(\tau,\ell)d\ell} \\ &= \frac{P(t,T)}{P(t,\tau)} \cdot e^{-\int_t^\tau f(t,\ell)d\ell + \int_t^T f(t,\ell)d\ell - \int_\tau^T f(\tau,\ell)d\ell} \\ &= \frac{P(t,T)}{P(t,\tau)} \cdot e^{\int_\tau^T f(t,\ell)d\ell - \int_\tau^T f(\tau,\ell)d\ell} \\ &= \frac{P(t,T)}{P(t,\tau)} \cdot e^{-\int_\tau^T (f(\tau,\ell) - f(t,\ell))d\ell}. \end{aligned} \tag{13.78}$$

The previous relation is valid no matter how we decide to model bond prices, but the first restriction we want to make is that the price at τ is consistent with the yield curve at t, such that Eq. (13.78) can be written as

$$P(\tau,T) = \frac{P^{\$}(t,T)}{P^{\$}(t,\tau)} \cdot e^{-\int_\tau^T (f(\tau,\ell) - f(t,\ell))d\ell}, \tag{13.79}$$

where $P^{\$}(t,\cdot)$ denote market prices observed at time t. Indeed, evaluating Eq. (13.79) for $\tau = t$ leaves $P(t,T) = P^{\$}(t,T)$. Note that Eq. (13.79) summarizes the modeling perspective in a continuous time setting, which is the continuous-time counterpart to that underlying the discrete time Ho & Lee model (see Chapter 12). Indeed, below, we shall show that the continuous-time version of the Ho & Lee model (see Eq. (13.72)) is a special case of HJM framework.

Models of the short-term rate are obviously consistent with Eq. (13.78) but cannot be consistent with Eq. (13.79), because we know these models cannot fit the initial yield curve. Models

of the short-term have implications regarding forward rates. Instead, models in the HJM framework do make assumptions regarding forward rates. In particular, the key points of the HJM methodology are (i) to take the current forward rates $f(t,\ell)$ as given (i.e., equal to those in the market) and, then, (ii) to model the future forward rate movements,

$$f(\tau,\ell)-f(t,\ell).$$

Therefore, in the HJM methodology, the current yield curve is perfectly fitted (see Eq. (13.79). In contrast, the approach to interest rate modeling in Section 13.4 is to *model* the *current* bond price $P(t,T)$ through assumptions regarding the dynamics of the short-term rate. We know from previous sections and Chapter 12 that fitting the initial term-structure is critical for market making purposes.

13.7.2 The model

13.7.2.1 Primitives

We assume information is Brownian, such that for any given T, the instantaneous forward rate, $(f(\tau,T))_{\tau\in[t,T]}$, satisfies

$$d_\tau f(\tau,T)=\alpha(\tau,T)\,d\tau+\sigma(\tau,T)\,dW_\tau,\quad \tau\in(t,T], \tag{13.80}$$

for some processes α and σ, and where $f(t,T)$ is given. Note that in this model, W doesn't obviously "depend" on T. Strings models, analyzed in Section 13.7.5, allow us to "index" W by T, in a sense to be discussed below.

13.7.2.2 No-arb restrictions

The next step is to derive restrictions on α that rule out arbitrage. Let $X_\tau\equiv-\int_\tau^T f(\tau,\ell)d\ell$. We have

$$dX_\tau=f(\tau,\tau)d\tau-\int_\tau^T\left(d_\tau f(\tau,\ell)\right)d\ell=\left(r_\tau-\alpha^I(\tau,T)\right)d\tau-\sigma^I(\tau,T)dW_\tau,$$

where the second equality follows by Eq. (13.80) and the following defintions:

$$\alpha^I(\tau,T)\equiv\int_\tau^T\alpha(\tau,\ell)d\ell,\quad \sigma^I(\tau,T)\equiv\int_\tau^T\sigma(\tau,\ell)d\ell.$$

By Eq. (13.77), $P=e^X$. By Itô's lemma,

$$\frac{d_\tau P(\tau,T)}{P(\tau,T)}=\left(r_\tau-\alpha^I(\tau,T)+\frac{1}{2}\left\|\sigma^I(\tau,T)\right\|^2\right)d\tau-\sigma^I(\tau,T)dW_\tau.$$

By the FTAP, we have that:

$$\frac{d_\tau P(\tau,T)}{P(\tau,T)}=\left(r_\tau-\alpha^I(\tau,T)+\frac{1}{2}\left\|\sigma^I(\tau,T)\right\|^2+\sigma^I(\tau,T)\lambda_\tau\right)d\tau-\sigma^I(\tau,T)d\tilde{W}_\tau,$$

where $\tilde{W}_\tau=W_\tau+\int_t^\tau\lambda_s ds$ is a Q-Brownian motion, and λ satisfies:

$$\alpha^I(\tau,T)=\frac{1}{2}\left\|\sigma^I(\tau,T)\right\|^2+\sigma^I(\tau,T)\lambda_\tau. \tag{13.81}$$

By differentiating the previous relation with respect to T gives us the arbitrage restriction that we were looking for:

$$\alpha(\tau,T)=\sigma(\tau,T)\int_\tau^T\sigma(\tau,\ell)^\top d\ell+\sigma(\tau,T)\lambda_\tau. \tag{13.82}$$

13.7.3 The dynamics of the short-term rate

By Eq. (13.80), the short-term rate satisfies

$$r_\tau \equiv f(\tau,\tau) = f(t,\tau) + \int_t^\tau \alpha(s,\tau)\,ds + \int_t^\tau \sigma(s,\tau)\,dW_s, \quad \tau \in (t,T]. \tag{13.83}$$

Differentiating with respect to τ yields

$$dr_\tau = \left(f_2(t,\tau) + \sigma(\tau,\tau)\lambda_\tau + \int_t^\tau \alpha_2(s,\tau)ds + \int_t^\tau \sigma_2(s,\tau)dW_s \right) d\tau + \sigma(\tau,\tau)dW_\tau, \tag{13.84}$$

where

$$\alpha_2(s,\tau) = \underbrace{\sigma_2(s,\tau)\int_s^\tau \sigma(s,\ell)^\top d\ell + \sigma(s,\tau)\sigma(s,\tau)^\top}_{\equiv\tilde{\alpha}_2(s,\tau)} + \sigma_2(s,\tau)\,\lambda_s. \tag{13.85}$$

Eq. (13.83) reveals that the short-term rate is in general non-Markov. A special case of Eq. (13.83) is the Ho and Lee (1986) model, where $\sigma(s,\tau) = \sigma$, a constant, such that, by Eq. (13.82), $\alpha(s,\tau) = \sigma^2(\tau - s) + \sigma\lambda_\tau$, consistently with Eq. (13.71). The Hull and White (1990) model is dealt with in the next section.

Note that regarding the main objective of these models—pricing—we do not need to be concerned with "estimating" any risk-premium, i.e., $\lambda(\cdot)$. We only need to consider the risk-neutral dynamics of the "left corner" in the forward rate surface, that is, those of the short-term rate r_τ. By Eq. (13.84) and Girsanov theorem, these are given by

$$dr_\tau = \left(f_2(t,\tau) + \int_t^\tau \tilde{\alpha}_2(s,\tau)ds + \int_t^\tau \sigma_2(s,\tau)d\tilde{W}_s \right) d\tau + \sigma(\tau,\tau)d\tilde{W}_\tau, \tag{13.86}$$

where $\tilde{\alpha}_2(s,\tau)$ is defined in Eq. (13.85), and $\tilde{W}_\tau$ denotes as usual a Brownian motion under the risk-neutral probability.

Eq. (13.86) can be easily simulated for the purpose of the evaluation of exotic products.

13.7.4 Embedding

Naturally, HJM models are not distinct from the short-term rate models of Section 13.4. Under "embeddability" conditions, HJM can be turned into short-term rate models—a property known as "universality" of HJM models.

13.7.4.1 Markovianity

Under which conditions do HJM models predict the short-term rate to be Markov? This question naturally links to the early literature reviewed in Section 13.4, where the whole yield curve is driven by a scalar Markov process—the short-term rate. Carverhill (1994) and Ritchken and Sankarasubramanian (1995) study conditions under which the original state vector can be enlarged such that the resulting "augmented" state vector is Markov and at the same time, includes the short-term rate as a component. The resulting model quite resembles some of the short-term rate models surveyed in Section 13.4. In these models, the short-term rate is not Markov, yet it is part of a system that is Markov. We now illustrate these points within the simple Markov scalar case.

Assume the forward-rate volatility is deterministic and takes the following form:

$$\sigma(t,T) = g_1(t)g_2(T) \quad \text{all } t,T. \tag{13.87}$$

By Eq. (13.86), then, r_τ satisfies

$$\begin{aligned} dr_\tau &= \left(f_2(t,\tau) + \int_t^\tau \tilde{\alpha}_2(s,\tau)ds + g_2'(\tau)\int_t^\tau g_1(s)d\tilde{W}_s \right) d\tau + \sigma(\tau,\tau)d\tilde{W}_\tau \\ &= \left(f_2(t,\tau) + \int_t^\tau \tilde{\alpha}_2(s,\tau)ds + \frac{g_2'(\tau)}{g_2(\tau)}g_2(\tau)\int_t^\tau g_1(s)d\tilde{W}_s \right) d\tau + \sigma(\tau,\tau)d\tilde{W}_\tau \\ &= \left[f_2(t,\tau) + \int_t^\tau \tilde{\alpha}_2(s,\tau)ds + \frac{g_2'(\tau)}{g_2(\tau)}\left(r_\tau - f(t,\tau) - \int_t^\tau \tilde{\alpha}(s,\tau)ds \right) \right] d\tau + \sigma(\tau,\tau)d\tilde{W}_\tau. \end{aligned}$$

Done. This is Markov. Precisely, the condition in Eq. (13.87) ensures the HJM model predicts the short-term rate is Markov. Mean reversion, then, obtains assuming that $g_2'(T) < 0$ for all T. For example, take λ to be a constant, and:

$$g_1(t) = \sigma \cdot e^{\kappa t}, \quad \sigma > 0, \quad g_2(t) = e^{-\kappa t}, \quad \kappa \geq 0,$$

such that $\sigma(t,T) = \sigma e^{-\kappa(T-t)}$, and the price volatility is $\int_\tau^T \sigma(\ell,T)\,d\ell = \sigma\frac{1-e^{-\kappa(T-\tau)}}{\kappa}$. This is the Hull-White model discussed in Section 13.4, of which the Ho and Lee model is a particular case, namely for $\kappa = 0$.

13.7.4.2 Short-term rate reductions

We prove everything in the Markov case. Let the short-term rate be solution to:

$$dr_\tau = \bar{b}(\tau, r_\tau)d\tau + a(\tau, r_\tau)d\tilde{W}_\tau,$$

where $\tilde{W}$ is a Q-Brownian motion, and $\bar{b}$ is some risk-neutralized drift function. The rational bond price function is $P(r_t,t,T)$, and the forward rate implied by the model is:

$$f(r_t,t,T) = -\frac{\partial}{\partial T}\ln P(r_t,t,T).$$

By Itô's lemma,

$$df = \left(\frac{\partial}{\partial t}f + \bar{b}f_r + \frac{1}{2}a^2 f_{rr} \right) d\tau + af_r d\tilde{W}.$$

But for $f(r,t,T)$ to be consistent with the solution to Eq. (13.80), it must be that

$$\begin{aligned} \alpha(t,T) - \sigma(t,T)\lambda_t &= \frac{\partial}{\partial t}f(r,t,T) + \bar{b}(t,r)f_r(r,t,T) + \frac{1}{2}a(t,r)^2 f_{rr}(r,t,T) \\ \sigma(t,T) &= a(t,r)f_r(t,r) \end{aligned} \tag{13.88}$$

and

$$f(t,T) = f(r,t,T). \tag{13.89}$$

In particular, the last condition can only be satisfied if the short-term rate model under consideration is of the perfectly fitting type.

13.7.5 Stochastic string shocks models

The first papers are Kennedy (1994, 1997), Goldstein (2000) and Santa-Clara and Sornette (2001). Heaney and Cheng (1984) are also very useful to read.

13.7.5.1 Stochastic singularity

Let $\sigma(\tau,T)=[\sigma_1(\tau,T),\cdots,\sigma_N(\tau,T)]$ in Eq. (13.80). For any $T_1<T_2$,

$$E[df(\tau,T_1)\,df(\tau,T_2)]=\sum_{i=1}^{N}\sigma_i(\tau,T_1)\,\sigma_i(\tau,T_2)\,d\tau,$$

and,

$$c(\tau,T_1,T_2)\equiv corr[df(\tau,T_1)\,df(\tau,T_2)]=\frac{\sum_{i=1}^{N}\sigma_i(\tau,T_1)\,\sigma_i(\tau,T_2)}{\|\sigma(\tau,T_1)\|\cdot\|\sigma(\tau,T_2)\|}. \tag{13.90}$$

By replacing this result into Eq. (13.82),

$$\begin{aligned}\alpha(\tau,T)&=\int_{\tau}^{T}\sigma(\tau,T)\cdot\sigma(\tau,\ell)^{\top}d\ell+\sigma(\tau,T)\lambda_{\tau}\\&=\int_{\tau}^{T}\|\sigma(\tau,\ell)\|\,\|\sigma(\tau,T)\|\,c(\tau,\ell,T)\,d\ell+\sigma(\tau,T)\lambda_{\tau}.\end{aligned}$$

One drawback of this model is that the correlation matrix of any $(N+M)$-dimensional vector of forward rates is degenerate for $M\geq 1$. Stochastic string models overcome this difficulty by *modeling* the correlation structure $c(\tau,\tau_1,\tau_2)$ for all τ_1 and τ_2 in an independent way, rather than *implying* it from a given N-factor model (as in Eq. (13.90)). In other terms, within the HJM methodology, one uses the functions σ_i to model both volatility and correlation structure of forward rates. The outcome might not be a good model, in practice. Instead, stochastic string models have two separate functions with which to model volatility and correlation.

The starting point is a model where the forward rate is solution to,

$$d_{\tau}f(\tau,T)=\alpha(\tau,T)\,d\tau+\sigma(\tau,T)\,d_{\tau}Z(\tau,T),$$

where the *string* Z satisfies the following five properties:

(i) *For all* τ, $Z(\tau,T)$ *is continuous in* T;

(ii) *For all* T, $Z(\tau,T)$ *is continuous in* τ;

(iii) $Z(\tau,T)$ *is a* τ*-martingale and, hence, a local martingale i.e.* $E(d_{\tau}Z(\tau,T))=0$;

(iv) $var(d_{\tau}Z(\tau,T))=d\tau$;

(v) $cov(d_{\tau}Z(\tau,T_1)\,d_{\tau}Z(\tau,T_2))=\psi(T_1,T_2)$ *(say)*.

Properties (iii), (iv) and (v) make Z Markovian. The functional form for ψ is crucially important to guarantee this property. Given the previous properties, one can easily derive a key property of forward rates. We have

$$\sqrt{var(df(\tau,T))}=\sigma(\tau,T)$$

$$c(\tau,T_1,T_2)\equiv corr(df(\tau,T_1)\,df(\tau,T_2))=\frac{\sigma(\tau,T_1)\,\sigma(\tau,T_2)\,\psi(T_1,T_2)}{\sigma(\tau,T_1)\,\sigma(\tau,T_2)}=\psi(T_1,T_2)$$

As claimed before, we now have two separate functions with which to model volatility and correlation.

13.7.5.2 No-arbitrage restrictions

Similarly as in the HJM-Brownian case, let $X_\tau \equiv -\int_\tau^T f(\tau,\ell)\,d\ell$. We have

$$dX_\tau = f(\tau,\tau)\,d\tau - \int_\tau^T d_\tau f(\tau,\ell)\,d\ell = \left(r_\tau - \alpha^I(\tau,T)\right)d\tau - \int_\tau^T \left[\sigma(\tau,\ell)\,d_\tau Z(\tau,\ell)\right]d\ell,$$

where as usual, $\alpha^I(\tau,T) \equiv \int_\tau^T \alpha(\tau,\ell)\,d\ell$. But $P(\tau,T) = \exp(X_\tau)$. Therefore,

$$\begin{aligned}\frac{dP(\tau,T)}{P(\tau,T)} &= dX_\tau + \frac{1}{2}var(dX_\tau)\\ &= \left(r_\tau - \alpha^I(\tau,T) + \frac{1}{2}\int_\tau^T\int_\tau^T \sigma(\tau,\ell_1)\,\sigma(\tau,\ell_2)\,\psi(\ell_1,\ell_2)\,d\ell_1 d\ell_2\right)d\tau\\ &- \int_\tau^T \left(\sigma(\tau,\ell)\,d_\tau Z(\tau,\ell)\right)d\ell.\end{aligned}$$

Next, suppose that the pricing kernel ξ satisfies:

$$\frac{d\xi_\tau}{\xi_\tau} = -r_\tau d\tau - \int_{\mathbb{T}} \phi(\tau,T)\,d_\tau Z(\tau,T)\,dT,$$

where $\mathbb{T}$ denotes the set of all "risks" spanned by the string Z, and ϕ is the corresponding family of "unit risk-premia."

In absence of arbitrage,

$$0 = E\left[d(P\xi)\right] = E\left[P\xi\cdot\left(\text{drift}\left(\frac{dP}{P}\right) + \text{drift}\left(\frac{d\xi}{\xi}\right) + cov\left(\frac{dP}{P},\frac{d\xi}{\xi}\right)\right)\right].$$

By exploiting the dynamics of P and ξ,

$$\alpha^I(\tau,T) = \frac{1}{2}\int_\tau^T\int_\tau^T \sigma(\tau,\ell_1)\,\sigma(\tau,\ell_2)\,\psi(\ell_1,\ell_2)\,d\ell_1 d\ell_2 + cov\left(\frac{dP}{P},\frac{d\xi}{\xi}\right),$$

where

$$\begin{aligned}cov\left(\frac{dP}{P},\frac{d\xi}{\xi}\right) &= E\left(\int_{\mathbb{T}}\phi(\tau,S)\,d_\tau Z(\tau,S)\,dS\cdot\int_\tau^T \sigma(\tau,\ell)\,d_\tau Z(\tau,\ell)\,d\ell\right)\\ &= \int_\tau^T\int_{\mathbb{T}}\phi(\tau,S)\,\sigma(\tau,\ell)\,\psi(S,\ell)\,dS d\ell.\end{aligned}$$

By differentiating α^I with respect to T we obtain,

$$\alpha(\tau,T) = \int_\tau^T \sigma(\tau,\ell)\,\sigma(\tau,T)\,\psi(\ell,T)\,d\ell + \sigma(\tau,T)\int_{\mathbb{T}}\phi(\tau,S)\,\psi(S,T)\,dS. \tag{13.91}$$

A proof of Eq. (13.91) is in the Appendix.

13.8 Interest rate derivatives

This section provides the main definitions and an evaluation framework for the main interest rate derivatives, such as bond options, callable and puttable bonds, interest rate swaps, caps, floors and swaptions.

We begin by addressing a simple fundamental question (in Section 13.8.1). How come interest rate derivatives are so pervasive, when in fact, we know very well that the short-term rate displays little volatility? We explain this apparent puzzle by proposing that the bulk of interest rate volatility occurs for deals relating to long maturities. Granted, the short-term rate is not very volatile in the short term. However, it is very persistent. This persistence can be at the origins of wild volatility of interest rate deals that have larger maturities.

The remaining sections develop evaluation models of interest rate derivatives, which mainly rely on models of the short-term rate, and their perfectly fitting extensions. Section 13.9 contains details of the so-called "market model," which is a HJM-style model intensively used by practitioners to evaluate these derivatives.

13.8.1 Fixed income market volatility and the persistence of the short-term rate

The implied vol on bond options is typically high for large maturities, in fact comparable to that on stocks. Why is it that this implied vol is so large, when in fact, the volatility of the short-term rate is one order of magnitude less than that in equity markets? We tackle this issue by analyzing the realized volatility, not implied volatility. To anticipate, the answer is that the short-term rate is very persistent, and it is "a risk for the long-run," pretty much in the same spirit of the explanations attempting to rationalize the equity premium puzzle reviewed in Chapter 9.

Define, first, the *term-structure of volatility*. It is the function, $\tau \mapsto \text{Vol}\left(R\left(\tau\right)\right)$, where $R\left(\tau\right)$ is the spot rate for the maturity τ, and $\text{Vol}\left(R\left(\tau\right)\right)$ is the standard deviation of this spot-rate. By the definition of $R\left(\tau\right)$, the term-structure of volatility can also be written as the function

$$\tau \mapsto \text{Vol}\left(-\frac{1}{\tau}\ln P\left(\tau\right)\right),$$

where $P\left(\tau\right)$ is the price of a zero with maturity equal to τ. It is instructive to see what this volatility looks like, for a concrete model. Consider the Vasicek model in Eq. (13.32). We know from Eq. (13.38), that the yields predicted by this model are:

$$R\left(\tau\right) = \frac{A\left(\tau\right)}{\tau} + \frac{1}{\tau}B\left(\tau\right)r, \quad B\left(\tau\right) = \frac{1-e^{-\kappa\tau}}{\kappa}.$$

for some function $A\left(\tau\right)$. Therefore, we have that,

$$\text{Vol}\left(R\left(\tau\right)\right) = \frac{1}{\tau}B\left(\tau\right)\text{Vol}_{\infty}\left(r\right), \tag{13.92}$$

where $\text{Vol}_{\infty}\left(r\right) = \sqrt{\frac{\sigma^2}{2\kappa}}$ is the long-run volatility of the short-term rate. For example, if $\kappa = 0.2$ and $\sigma = 0.03$, then $\text{Vol}_{\infty}\left(r\right) \approx 4.7\%$. Given the previous values for κ and σ, Figure 13.5 depicts the term-structure of volatility, i.e. Eq. (13.92).

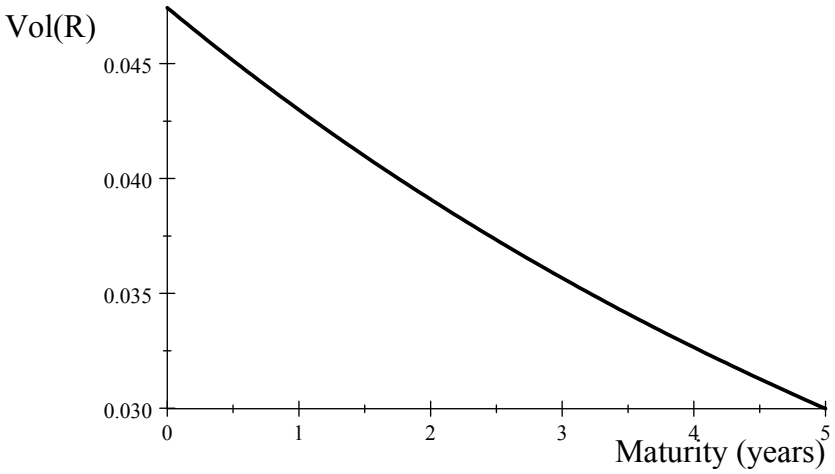


FIGURE 13.8. The term-structure of volatility predicted by the Vasicek model.

Figure 13.8 illustrates how the term-structure of volatility decreases over the maturity of the zero, attaining its maximum at $\text{Vol}_{\infty}(r) \approx 4.7\%$. It is natural, as the yield curve in this model flattens out, converging towards a constant long-term value, the asymptotic interest rate, as we say sometimes.

Despite this, the *volatility of bond returns* can be much higher, as we now illustrate. We need to figure out the dynamics of the bond price, for the Vasicek model. By Itô's lemma,

$$\frac{dP_t(\tau)}{P_t(\tau)} = (\cdots)\, dt + (-\sigma \cdot B(\tau))\, dW_t.$$

Therefore, the volatility of bond returns is,

$$\text{Vol}\left(\frac{dP}{P}\right) = \sigma B(\tau). \tag{13.93}$$

Compare Eq. (13.93) with Eq. (13.92). The main difference between these two equations is that the right hand side of Eq. (13.92) is divided by τ, which makes $\text{Vol}(R(\tau))$ decreasing in τ. (Otherwise, $\text{Vol}_{\infty}(r)$ and σ have roughly the same order of magnitude.) The main point is that the yield $R(\tau)$ is, simply, an *average return* achieved once a bond is purchased and held until its expiry. This average return is progressively less volatile as time to maturity gets larger, and becomes constant, eventually. The return $\frac{dP}{P}$ does, instead, measure the capital gains achieved while trading the bond. The volatility of these capital increases with time to maturity. Even if σ is very small, the bond returns volatility in Eq. (13.93) can be quite high. Suppose, for example, that κ is close to zero, in which case $\text{Vol}\left(\frac{dP}{P}\right) \approx \sigma \cdot \tau$, which is 15% for a five year zero. These properties are illustrated by Figure 13.6, which depicts Eq. (13.93) when the parameter values are $\kappa = 0.2$ and $\sigma = 0.03$.

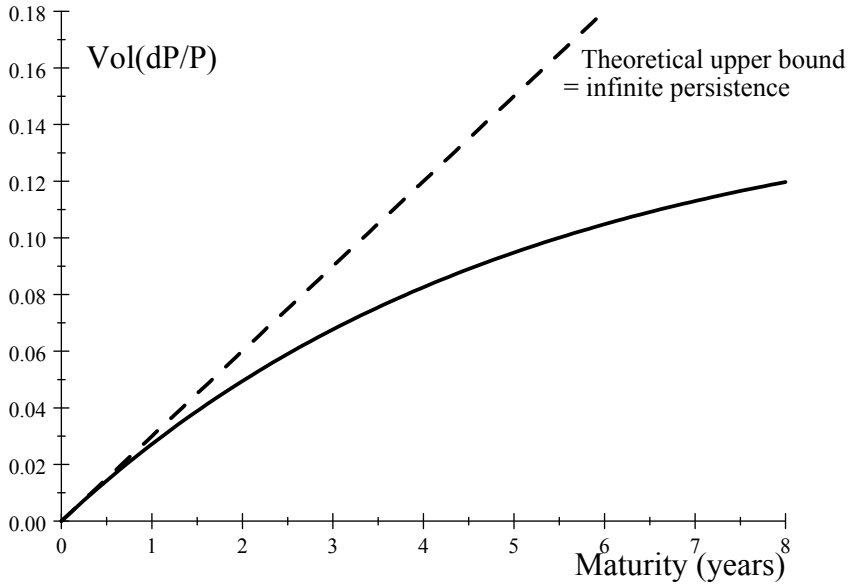


FIGURE 13.9. The dashed line depicts the bond return volatility, Vol $\left(\frac{dP}{P}\right)$, arising when the persistence parameter $\kappa = 0$, and the solid line is the bond return volatility for $\kappa = 0.2$.

The high persistence of the short-term rate, as measured by a low value of κ, makes long maturity bond returns quite volatile. Intuitively, this high persistence implies that a shock in the short-term rate has long lasting effects on the future path of the short-term rate. This makes the short-term rate very volatile in the long-run, which makes the value of long maturity zeros very volatile as a result. Intuitively, interest rates exhibit inertia: (i) it takes a number of shocks to move interest rates away from their equilibrium paths and so, short-term bonds are not volatile; and (ii) it takes time for interest rates to absorbe shocks and so, medium/long -term bonds are volatile. For example, Figure 13.7 depicts the dynamics of the three month rate and those of the three months into five years forward swap rate, an interest rate that refers to relatively higher maturities, as explained in Section 13.8.7 below (see Eq. (13.113)). The forward swap rate is orders of magnitude more volatile than the short-term rate.

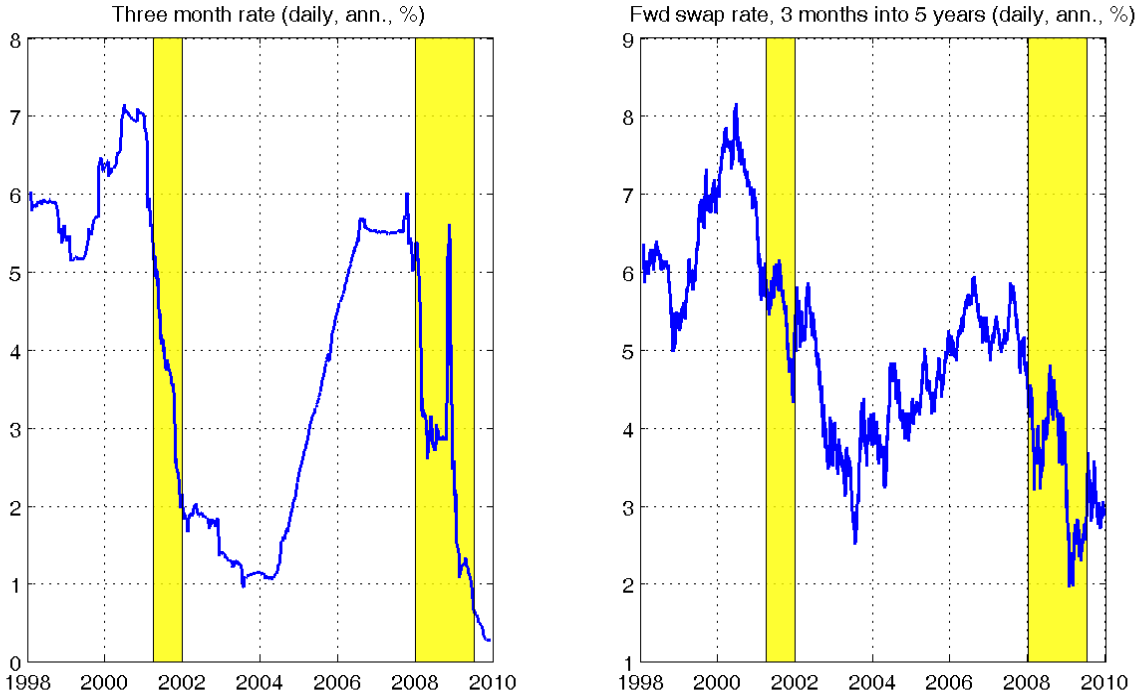


FIGURE 13.10. Interest rate volatility increases with maturity.

These facts are confirmed by the "implicit" (not implied) option-based volatility. In Section 13.8.4, Eq. (13.104), we show that this volatility is,

$$v = \sqrt{T-t} \cdot \text{Vol}^O, \quad \text{Vol}^O \equiv \sigma \sqrt{\frac{1-e^{-2\kappa(T-t)}}{2\kappa\,(T-t)}} \frac{1-e^{-\kappa(S-T)}}{\kappa}.$$

As κ gets small, Vol^O tends to $\sigma \times (S-T)$, which increases with the bond's time to maturity left at its expiration, $S-T$.

The previous reasoning does, of course, still hold in the more realistic case of a three-factor model, such as that in Eqs. (13.55). In that case, as explained, κ_r is large and $\kappa_{\bar{r}}$ is small: the short-term rate is quite persistent because it mean-reverts, quickly, to a persistent process, which we denoted as $\bar{r}_\tau$. Naturally, in such as a three-factor model, Eq. (13.93) does not hold anymore, as we should add two more volatility components, related to stochastic volatility, v_τ, and the persistent process $\bar{r}_\tau$. However, the bond return volatility would be boosted by the high persistence of $\bar{r}_\tau$.

13.8.2 Hypothetical continuous payoffs

Interest rate derivatives could be priced in a very elegant fashion once we assume payoffs are paid out continuously. Let w denote the price of any such derivative, and π be the instantaneous payoff paid by it, a function of calendar time and r. Consider any model of the short-term rate in Section 13.4, and to simplify, assume that $d = 1$, such that r_t in Eq. (13.25) carries all information. By the FTAP, w is solution to the following partial differential equation:

$$\frac{\partial w}{\partial \tau} + \bar{b} w_r + \frac{1}{2} a^2 w_{rr} + \pi = rw, \quad \text{for all } (r,\tau) \in \mathbb{R} \times [t,T), \tag{13.94}$$

subject to the appropriate boundary conditions.

In Eq. (13.94), $\bar{b}$ is some risk-neutralized drift function of r_t, with the payoff π term that needs to be added to the expected basis point appreciation of w under Q, $\frac{\partial w}{\partial \tau} + \bar{b}w_r + \frac{1}{2}a^2 w_{rr}$, such that overall, derivative yields rw, which prevents arbitrage. While the payoff π could approximate many of the interest rate derivatives payoffs dealt with below, standard market practice typically relies on payoffs defined in terms of LIBOR-type discretely-compounded rates and, intermediate payments obviously only take place discretely, not continuously.

We now proceed with explaining how to proceed while pricing interest rate derivatives based on more realistic assumptions than the continuous payoffs leading to the evaluation equation (13.94). We begin by introducing probably the most famous framework of analysis regarding interest rate derivation evaluation, based on the so called forward probabilities.

13.8.3 Forward martingale probabilities

We know from Chapter 5 that asset prices normalized by the money market account are martingales, yet the money market account is not the only numéraire in markets where interest rates are random. Chapter 5 also explains we may use the price of zero-coupon bonds as an alternative numéraire, which leads to the notion of *forward probabilities* (see Section 5.4). This section develops this notion more closely, and hinges upon it to demonstrate useful results applying to the dynamic of interest rates, not only prices, which will be quite useful whilst pricing interest rate derivatives.

Consider a forward contract agreed at time t, and let $\varphi_T(t)$ be the T-forward price for a claim S_T at $T > t$. That is, and consistently with notation in Chapter 11, $\varphi_T(t)$ is the price agreed at t, which will be paid at T for delivery of the claim at T, such that the agreement is worthless at t. Clearly,

$$0 = \mathbb{E}_t\left(e^{-\int_t^T r_u du} \cdot (S_T - \varphi_T(t))\right).$$

But since $\varphi_T(t)$ is known at time t,

$$\mathbb{E}_t\left(e^{-\int_t^T r_u du} \cdot S_T\right) = \varphi_T(t) \cdot \mathbb{E}_t\left(e^{-\int_t^T r_u du}\right). \tag{13.95}$$

For example, assume S is the price of a *traded* asset, such that $\mathbb{E}_t(e^{-\int_t^T r_u du} S_T) = S_t$: Eq. (13.95) then collapses to the well-known formula: $\varphi_T(t)\, P(t,T) = S_t$ as explained in Chapter 11. Yet and as also explained in Chapter 11, entering the forward contract originated at t, at a later date $\tau > t$, costs. To calculate the marking-to-market of the forward at time τ, note that the final payoff at time T is $S_T - \varphi_T(t)$. Discounting this payoff at $\tau \in [t,T]$ delivers $P(\tau,T) \cdot [\varphi_T(\tau) - \varphi_T(t)]$.

Next, let us elaborate on Eq. (13.95). We can use the basic bond pricing equation (13.2) in Section 13.2, and rearrange terms in Eq. (13.95), to obtain:

$$\varphi_T(t) = \mathbb{E}_t\left(\frac{e^{-\int_t^T r_u du}}{P(t,T)} \cdot S_T\right) = \mathbb{E}_t\left(\eta_T(T) \cdot S_T\right), \tag{13.96}$$

where

$$\eta_T(T) \equiv \frac{e^{-\int_t^T r_u du}}{P(t,T)}.$$

Eq. (13.96) suggests that we can define a new probability Q_F^T, as follows,

$$\eta_T(T) = \frac{dQ_F^T}{dQ} \equiv \frac{e^{-\int_t^T r_u du}}{\mathbb{E}_t\left(e^{-\int_t^T r_u du}\right)}. \tag{13.97}$$

Naturally, $\mathbb{E}_t[\eta_T(T)] = 1$. Moreover, if the short-term rate process is deterministic, $\eta_T(T)$ equals one and Q and Q_F^T are the same.

In terms of this new probability Q_F^T, the forward price $\varphi_T(t)$ is:

$$\varphi_T(t) = \mathbb{E}_t\left(\eta_T(T) \cdot S_T\right) = \int \left(\eta_T(T) \cdot S_T\right) dQ = \int S_T dQ_F^T = \mathbb{E}_{Q_F^T}\left(S_T\right), \tag{13.98}$$

where $\mathbb{E}_{Q_F^T}(\cdot)$ is conditional expectation at t under Q_F^T. For reasons clarified in a moment, Q_F^T is referred to as the T-forward probability. Clearly, $\varphi_T(T) = S_T$. Therefore, Eq. (13.98) is, also,

$$\varphi_T(t) = \mathbb{E}_{Q_F^T}\left(\varphi_T(T)\right).$$

That is, forward prices are martingales under the forward probability—whence, the expression, forward *martingale* probability, which we shall shorten to forward probability to simplify the presentation. Naturally, and as usual, future prices are martingales under the risk-neutral probability, not the forward, as explained in Chapter 11.

The forward probability is a useful tool, which helps pricing interest-rate derivatives, as we shall explain in detail below. It was introduced by Geman (1989) and Jamshidian (1989), and further analyzed by Geman, El Karoui and Rochet (1995). The appendix provides additional details. Appendix 2 relates forward prices to their certainty equivalent, and Appendix 3 deals with additional technical details. We now rely on this probability to facilitate the calculation of options on bonds.

13.8.4 European options on bonds

13.8.4.1 A bond option pricing formula

Let T be the expiration date of a European call option on a zero-coupon bond, and $S > T$ the expiration date of the bond. We consider a simple model of the short-term rate with $d = 1$, such that the price of a zero is $P(r, \tau, S)$, with the usual notation. Consider the price of an option on this bond, maturing at T and with strike equal to K. It equals:

$$C^b(r, \tau, T, S) = \mathbb{E}_t\left[e^{-\int_t^T r_\tau d\tau}\left(P(r_T, T, S) - K\right)^+\right]. \tag{13.99}$$

Finding a closed-form solution for this price looks formidable. Note indeed that C^b is solution to a partial differential equation, subject to the boundary condition $C^b(r, T, T, S) = (P(r, T, S) - K)^+$, where $P(r, \tau, S)$ is also solution to another partial differential equation. Relying on the forward probability allows simplifying this problem.

Let us, then, elaborate on Eq. (13.99). Note that the main issue we encounter is that the payoff $\left(P(r_T, T, S) - K\right)^+$ depends on r_T, and yet the discounting factor, $e^{-\int_t^T r_\tau d\tau}$, would also obviously depend on the realization of the short-term rate. As discussed in Chapter 5, it is a general issue arising whilst evaluating fixed-income instruments, because interest rates are obviously random in this context. These issues are overcome by turning the risk-neutral expectation in (13.99) into the forward.

Let $\mathbb{I}_{\cdot}$ be the indicator function of the event the option is exercized, i.e $P_T \equiv P(r_T, T, S) \geq K$. We have:

$$\begin{aligned}
&C^b(r_t, t, T, S) \\
&= \mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau} P(r_T, T, S) \cdot \mathbb{I}_{P_T \geq K}\right) - K \cdot \mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau} \cdot \mathbb{I}_{P_T \geq K}\right) \\
&= P(r_t, t, S) \cdot \mathbb{E}_t\left(\frac{e^{-\int_t^S r_\tau d\tau}}{P(r_t, t, S)} \cdot \mathbb{I}_{P_T \geq K}\right) - KP(r_t, t, T) \cdot \mathbb{E}_t\left(\frac{e^{-\int_t^T r_\tau d\tau}}{P(r_t, t, T)} \cdot \mathbb{I}_{P_T \geq K}\right) \\
&= P(r_t, t, S) \cdot \mathbb{E}_{Q_F^S}(\mathbb{I}_{P_T \geq K}) - KP(r_t, t, T) \cdot \mathbb{E}_{Q_F^T}(\mathbb{I}_{P_T \geq K}) \\
&= P(r_t, t, S) \cdot Q_F^S[P(r_T, T, S) \geq K] - KP(r_t, t, T) \cdot Q_F^T[P(r_T, T, S) \geq K], \qquad (13.100)
\end{aligned}$$

where the second equality follows by a standard calculation,[30] Q_F^i is the i-forward probability, $i = T, S$; and, finally, $\mathbb{E}_{Q_F^i}(\cdot)$ is the expectation taken under the i-forward martingale probability. Section 13.8.4 explains how the two probabilities in Eq. (13.100) are computed in the case of a specific model.

Eq. (13.100) is the bond option counterpart to Black & Scholes formula, in that: (i) the underlying asset is a zero-coupon bond expiring at S, the current price of which, $P(r, t, S)$, multiplies the probability Q_F^S; and (ii) the present value of the strike price, $KP(r, t, T)$, multiplies a second probability, Q_F^T. Below, we provide an important instance in which the probabilities in Eq. (13.100) can be calculated in closed-form, based on the assumption that the short-term rate is Gaussian. We actually present two models with a closed-form solution, one developed in a seminal paper by Jamshidian (1989), and the second one, simply, its perfectly fitting extension. We shall primarily deal with call options although the pricing of put options will be easily derived through a put-call parity, given next.

13.8.4.2 A put-call parity arising in government bond markets

Consider the identity,

$$(K - P(T, S))^+ \equiv (P(T, S) - K)^+ + K - P(T, S), \quad T \leq S.$$

Taking the risk-neutral, discounted expectations of both sides of this equation leaves,

$$\begin{aligned}
&\mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau}(K - P(T, S))^+\right) \\
&= \mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau}(P(T, S) - K)\right)^+ + P(t, T)K - \mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau} P(T, S)\right) \\
&= \mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau}(P(T, S) - K)\right)^+ + P(t, T)K - P(t, S),
\end{aligned}$$

[30] By the Law of Iterated Expectations, we have, similarly as in Section 12.2.2.2,

$$\mathbb{E}_t\left[e^{-\int_t^T r(\tau)d\tau} P(r(T), T, S)\mathbb{I}_{\text{exe}}\right] = \mathbb{E}_t\left[e^{-\int_t^T r(\tau)d\tau}\mathbb{I}_{\text{exe}}\mathbb{E}_T\left(e^{-\int_T^S r(\tau)d\tau}\right)\right] = \mathbb{E}_t\left[e^{-\int_t^S r(\tau)d\tau}\mathbb{I}_{\text{exe}}\right].$$

where the last equality follows by the same argument leading to Eq. (13.100). Therefore, we have the put-call parity relation:

$$\text{Put}\left(t,T;P\left(t,S\right),K\right)=\text{Call}\left(t,T;P\left(t,S\right),K\right)+P\left(t,T\right)K-P\left(t,S\right), \tag{13.101}$$

where $\text{Put}\left(t,T;P\left(t,S\right),K\right)$ is the price of a European put written on a zero expiring at time S, expiring at time $T<S$, and struck at K, and $\text{Call}\left(\cdot\right)$ denotes the corresponding call price.

13.8.4.3 Jamshidian & Vasicek model

Suppose that the short-term rate is solution to the Vasicek model considered in Section 13.4 (see Eq. (13.32)), such that under the risk-neutral probability,

$$dr_{\tau}=\kappa\left(r^{*}-r_{\tau}\right)d\tau+\sigma d\tilde{W},$$

where $\tilde{W}$ is a Q-Brownian motion and $r^{*}\equiv\bar{r}-\frac{\lambda\sigma}{\kappa}$, and remaining notation is as in Section 13.4.4. As we know from Section 13.4, Eq. (13.36), the price of a zero predicted by the model is $P\left(r_{\tau},\tau,S\right)=e^{A(\tau,S)-B(\tau,S)r_{\tau}}$, for some function A, and $B\left(t,T\right)=\frac{1}{\kappa}\left(1-e^{-\kappa(T-t)}\right)$. In Section 13.4, it was also argued that the price of a European option on a zero is,

$$\begin{aligned}&\mathbb{E}_{t}\left(e^{-\int_{t}^{T}r_{\tau}d\tau}\left(P(r_{T},T,S)-K\right)^{+}\right)\\&=P(r_{t},t,S)\cdot Q_{F}^{S}\left[P(r_{T},T,S)\geq K\right]-KP(r_{t},t,T)\cdot Q_{F}^{T}\left[P(r_{T},T,S)\geq K\right],\end{aligned} \tag{13.102}$$

where Q_{F}^{T} denotes the T-forward probability.

In Appendix 8, we show that the two probabilities in Eq. (13.102) can be evaluated by the *changes of numéraire* described in Section 13.8.3, such that the solution for $P\left(r,T,S\right)$ is:

$$\begin{aligned}P\left(r,T,S\right)&=\frac{P\left(r,t,S\right)}{P\left(r,t,T\right)}e^{-\frac{1}{2}\sigma^{2}\int_{t}^{T}[B(\tau,S)-B(\tau,T)]^{2}d\tau-\sigma\int_{t}^{T}[B(\tau,S)-B(\tau,T)]dW_{\tau}^{Q_{F}^{T}}}\quad\text{under }Q_{F}^{T}\\P\left(r,T,S\right)&=\frac{P\left(r,t,S\right)}{P\left(r,t,T\right)}e^{\frac{1}{2}\sigma^{2}\int_{t}^{T}[B(\tau,S)-B(\tau,T)]^{2}d\tau-\sigma\int_{t}^{T}[B(\tau,S)-B(\tau,T)]dW_{\tau}^{Q_{F}^{S}}}\quad\text{under }Q_{F}^{S}\end{aligned} \tag{13.103}$$

where $W^{Q_{F}^{T}}$ is a Brownian motion under the forward probability Q_{F}^{T}. Therefore, simple algebra now reveals that:

$$Q_{F}^{S}\left[P\left(T,S\right)\geq K\right]=\Phi\left(d_{1}\right),\quad Q_{F}^{T}\left[P\left(T,S\right)\geq K\right]=\Phi\left(d_{1}-v\right),\quad d_{1}=\frac{\ln\frac{P(r_{t},t,S)}{KP(r_{t},t,T)}+\frac{1}{2}v^{2}}{v},$$

where

$$v^{2}=\sigma^{2}\int_{t}^{T}\left[B(\tau,S)-B(\tau,T)\right]^{2}d\tau=\sigma^{2}\frac{1-e^{-2\kappa(T-t)}}{2\kappa}B(T,S)^{2}. \tag{13.104}$$

It is a very elegant formula. It resembles Black & Scholes, although the inputs to the volatility function link to both the instantaneous volatility, σ, and the speed of mean-reversion, κ, of the short-term rate. We provided the economic interpretation of this dependence in Section 13.8.1.

13.8.4.4 Perfectly fitting extension

Brigo and Mercurio (2006) survey a number of perfectly fitting models that go well beyond that in the previous section. The simplest relies on the Hull and White (1990) model in Eq. (13.73) of Section 13.6.3. Note that while, formally, the solution to Eq. (13.99) is the same as in the previous section, the value of more complex derivatives depends on whether we use or not a perfectly fitting extension, as Section 13.8.6 explains in further detail.

13.8.5 Callable and puttable bonds

This section relies on a continuous time model to illustrate a few, key properties of callable and puttable bonds. As explained in Chapter 12 (see Section 12.8) *callable* bonds are assets that give the issuer the right to buy them back at certain times and predetermined prices; *puttable bonds*, instead, give the investor the right to sell them back to the issuer at certain times and strikes. Chapter 12 explains the pricing mechanims of these assets within a discrete time framework in which the option exercise is of the American type, relying on binomial trees. This section only considers European-style options (and zero coupon bonds), but leads to clear predictions and analytical evaluation formule. For simplicity, we consider *non-defaultable*, and zero coupon, bonds.

Consider, first, callable bonds maturing at S, and let K be the strike at which they can be called. Suppose that the date of exercise, if any, is some future time $T < S$. Repeating some of the reasoning in Chapter 12, assume exercise, in which case the issuer can buy its bonds back at K and re-issue a zero-coupon bond at better market conditions, P, where P denotes as usual the price of a *non-callable* bond. The difference, $P - K$, is just a net gain for the issuer. Therefore, the callable bond is worth just K when $K < P$. Instead, if $K > P$, the issuer does not have any incentives to exercise and, then, the value of the callable bond is just that of a non-callable bond. Therefore, the callable bond is worth P when $K > P$. To sumup, the value at T of a callable bond is $\min\{K, P(T,S)\}$. It easy to see that,

$$\min\{P, K\} = P - \max\{P - K, 0\}.$$

Therefore, we see that the price of a *callable* bond with maturity date S, equals the price of a *non-callable* bond with the same maturity date S, minus the value to call the bond, which equals the price of an hypothetical option on the *non-callable* bond, struck at K.

We can apply these insights to price a callable option in a concrete example. Consider, for example, the short-term rate in the Vasicek model. Then, if the short-term rate is r_t at time t, the value as of time t of the non-defaultable zero coupon bond maturing at time S, callable at time $T < S$, at a strike price equal to K, is,

$$P^{\text{callable}}(r_t, t, T, S) = P(r_t, t, S) - \text{Call}^b(r_t, t, T, S), \tag{13.105}$$

where $P(r_t, t, S)$ is the value of the *non-callable* zero maturing at time S, and $\text{Call}^b(r_t, t, T, S)$ is the value of a call option on the *non-callable* S-zero, maturing at time T and having a strike price equal to K.

Eq. (13.105) shows that the presence of the option to call the bond raises the cost of capital for the issuer.

In the context of the Vasicek model, the solution to $C^b(r_t, t, T, S)$ in Eq. (13.105) is given by the Jamshidian's (1989) formula in Eq. (13.102), which we now use below. Figure 13.8 depicts the behavior of the price of the callable bond in Eq. (13.105), $P^{\text{callable}}(r, 0, T, S)$, as a function of the short-term rate, r, when the exercise price $K = 0.65$, option maturity is $T = 0.5$, and the bond maturity is $S = 10$. Finally, to evaluate Eq. (13.105), we make use of the closed-form solution in Eq. (13.36), and the parameter values $\kappa = 0.2$, $\bar{r} = 0.06$, $\sigma = 0.03$, $\lambda = -1.7146 \times 10^{-2}$.

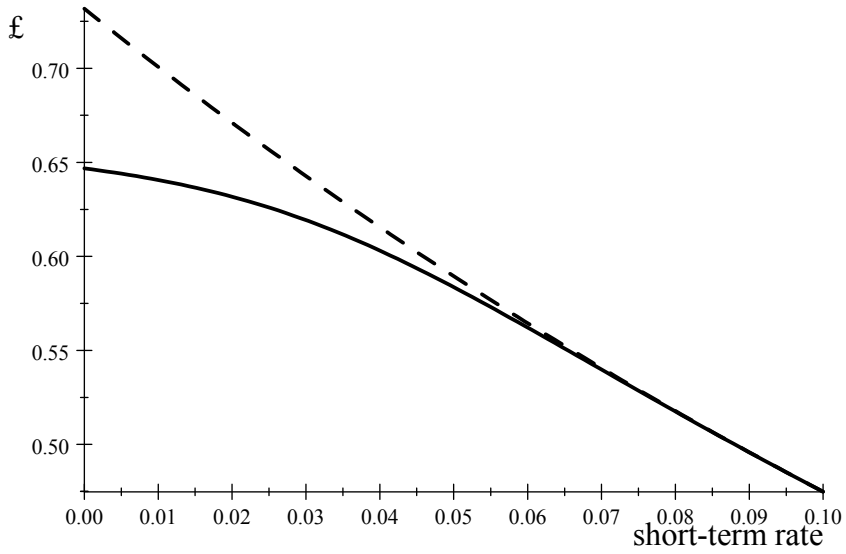


FIGURE 13.11. "Negative convexity." Solid line: the price of a callable bond. Dashed line: the price of a non-callable bond. The price of a callable bond exhibits negative convexity with respect to the short-term rate.

As Figure 13.8 illustrates, the convexity of the non-callable bond price is destroyed by the convexity of the price of the option embedded in the callable bond. Intuitively, as the short-term rate lowers, callable and non-callable bond prices increase. However, the price of callable bonds increases less because as the short-term rate decreases, bond prices increase and, then, the probability the issuer will exercise the option increases. As a result, the risk-neutral distribution of the callable bond price becomes markedly shifted towards the strike price, $K = 0.65$, implying a progressively lower decay rate for the bond price as the short-term rate gets small.

What is the duraton of a callable bond? Roughly, a five year bond with fixed coupons issued when interest rates are relatively high might resemble, so to speak, a three year conventional bond, as a likely decrease in the interest rates would lead the bond-issuer to reedem its debt at the strike price. To formalize this intuition, we determine the stochastic duration of the callable bond predicted by this model, using Eq. (13.31). For the Vasicek model, the semi-elasticity of the non-callable bond price with respect to r is $\Psi(r, T-t) = \frac{1}{\kappa}\left(1 - e^{-\kappa(T-t)}\right)$; its inverse with respect to time-to-maturity is

$$\Psi^{-1}(x) = -\frac{1}{\kappa}\ln(1 - \kappa x).$$

Therefore, the stochastic duration for the callable bond predicted by the Vasicek model is, by Eq. (13.31):

$$D(r, S-t) = -\frac{1}{\kappa}\ln\left(1 + \kappa\frac{P_r(r,t,S) - C_r^b(r,t,T,S)}{P(r,t,S) - C^b(r,t,T,S)}\right),$$

where subscripts denote partial derivatives.

[In progress]

Next, we proceed with pricing the puttable bond. As explained in the previous chapter (see Section 12.8), the payoff at the expiration of the bondholders right to tender the bonds is:

$$\max\{P, K\} = P + \max\{K - P, 0\},$$

where P is the price of a non-puttable bond. We can use, again, the Vasicek model to price the previous payoff. The price at t of a non-defaultable zero-coupon bond maturing at time S, puttable at time $T < S$, at a strike price equal to K, when the short-term rate is r_t, is:

$$P^{\text{puttable}}(r_t, t, T, S) = P(r_t, t, S) + \text{Put}^b(r_t, t, T, S) = \text{Call}^b(r_t, t, T, S) + P(r_t, t, T) K,$$

where $P(r_t, t, S)$ is the value of the *non-puttable* zero maturing at time S; $\text{Put}^b(r_t, t, T, S)$ is the value of a put option on the *non-puttable* zero maturing at S, maturing at T, struckable at K; and the second equality follows by the put-call parity of Eq. (13.101), with $\text{Call}^b(r_t, t, T, S)$ defined as in Eq. (13.105).

[In progress]

13.8.6 Options on fixed coupon bonds

For simplicity, we shall ignore issues regarding coupon accruals, and assume the expiration date of these options occurs at any of the reset dates. Therefore, the payoff of an option maturing at T_0 on a fixed coupon bond paying off at dates $T_1, \cdots, T_n$ is

$$(P_{\text{fcb}}(T_0, T_n) - K)^+ = \left(P(T_0, T_n) + \sum_{i=1}^{n} c_i P(T_0, T_i) - K \right)^+. \tag{13.106}$$

Evaluating the expectation of the payoff in Eq. (13.106) is somehow problematic: the maximum between zero and a sum is obviously not in general the same as the sum of the maxima between zero and each element of the sum. Even with a model in which bond prices are log-normal, the sum of log-normals is not log-normal. However, this issue can be dealt through a very well-know trick, described next.

Consider any of the models of the short-term rate reviewed in Section 13.4, in which the price of a zero is some function $P(t, T) = P(r, t, T)$. Assume that $P(r, T, T)$ is decreasing in r,[31] such that (under additional conditions) there exists a unique value of r, say $r^*(K)$, which solves the following equation

$$P(r^*(K), T_0, T_n) + \sum_{i=1}^{n} c_i P(r^*(K), T_0, T_i) = K. \tag{13.107}$$

The payoff in Eq. (13.106) can then be written as:

$$\left(\sum_{i=1}^{n} \bar{c}_i P(r_{T_0}, T_0, T_i) - K \right)^+ = \left(\sum_{i=1}^{n} \bar{c}_i \left(P(r_{T_0}, T_0, T_i) - P(r^*(K), T_0, T_i) \right) \right)^+,$$

where $\bar{c}_i = c_i$, $i = 1, \cdots, n-1$, and $\bar{c}_n = 1 + c_n$.

Next, note that the terms $P(r_{T_0}, T_0, T_i) - P(r^*(K), T_0, T_i)$ have all the same sign for all i.[32] Therefore, the payoff in Eq. (13.106) is

$$\left(\sum_{i=1}^{n} \bar{c}_i P(r_{T_0}, T_0, T_i) - K \right)^+ = \sum_{i=1}^{n} \bar{c}_i \cdot (P(r_{T_0}, T_0, T_i) - \mathcal{K}_i^*(K))^+, \quad \mathcal{K}_i^*(K) \equiv P(r^*(K), T_0, T_i), \tag{13.108}$$

[31] Bond prices are indeed always decreasing in the short-term rate in all one-factor stationary, Markov models of the short-term rate. However, this is not a general property in multi-factor models (see Mele, 2003).

[32] Suppose that $P(r_{T_0}, T_0, T_1) > P(r^*, T_0, T_1)$. Because $P(\cdot, T_0, T_1)$ is decreasing, $r_{T_0} < r^*$. Hence, $P(r_{T_0}, T_0, T_i) > P(r^*, T_0, T_i)$ for each i.

Each term of the sum in Eq. (13.108) can be evaluated as an option on a pure discount bond with strike price equal to $P(r^*(K), T_0, T_i)$, where the threshold $r^*(K)$ is found numerically. The device to reduce the problem of an option on a fixed coupon bond to a problem involving the sum of options on zero coupon bonds was invented by Jamshidian (1989).

The price of the call on the fixed coupon bond is, therefore

$$\overline{\text{Call}}\left(t, T_0; P_{\text{fcb}}\left(t, T_n\right), K, v\right) = \sum_{i=1}^{n} \bar{c}_i \cdot \text{Call}\left(t, T_0; P\left(t, T_i\right), \mathcal{K}_i^*\left(K\right), v_i\right), \tag{13.109}$$

where r^* solves Eq. (13.107), and

$$\text{Call}\left(t, T_0; P_i, \mathcal{K}_i^*\left(K\right), v_i\right) = P_i \Phi\left(d_{1,i}\right) - \mathcal{K}_i^*\left(K\right) P\left(t, T_0\right) \Phi\left(d_{1,i} - v_i\right),$$

$$d_{1,i} = \frac{\ln \frac{P_i}{\mathcal{K}_i^*(K) P(t,T_0)} + \frac{1}{2} v_i^2}{v_i}, \quad v_i = \sigma \sqrt{\frac{1 - e^{-2\kappa(T_0 - t)}}{2\kappa}} B\left(T_0, T_i\right), \quad B\left(t, T\right) = \frac{1}{\kappa}\left(1 - e^{-\kappa(T-t)}\right).$$

The price of a put can then be determined through the put-call parity in Eq. (13.101).

Why are perfectly fitting models so important, in practice? Suppose that in Eq. (13.107), the critical value r^* is determined through Vasicek's model. This assumption is attractive because it leads to evaluate the payoff in Eq. (13.108) through the Jamshidian's formula of Section 13.8.4. However, this way to proceed does not ensure that the yield curve is perfectly fitted.

The natural alternative is to use the corresponding perfectly fitting extension, the Hull and White model in Section 13.8.4, i.e. Eq. (13.73), and use this price to calibrate r^* in Eq. (13.107). Note, now, the importance of a perfectly fitting model. As mentioned in Section 13.8.4, both Jamshidian and its perfectly fitting extension agree regarding the price of an option on a zero. However, Jamshidian and its perfectly fitting extension would assign different values to options on coupon bearing bonds, because they would lead to different values for r^* in Eq. (13.107) and, hence, different values for the fictitious strikes $P\left(r^*, T_0, T_i\right)$ in Eq. (13.108).

13.8.7 Interest rate swaps

Savings and Loan associations (also known as "thrifts"; S&L, henceforth) are types of cooperative ventures that extend mortgage, car and personal loans to individual members, and are typically financed through retail savings; their governance is marked by the presence of voting and maganement rights extended to their depositors and borrowers members.

During the 1980s and through the beginning of the 1990s, S&Ls entered into a deep and persistent crisis, leading to a painful government bailout of about $125b under George H.W. Bush administration. There are many causes of this crisis, and one of them may link to the rise in short-term rates arising as an attempt of the FED to fight against inflation—the so-called Monetary Experiment discussed in Section 13.4.5. Banking is risky precisely because it involves lending at horizons longer than those for borrowing. S&L "banking" is not immune to the risks inherent in such modus operandi. Interest rate swaps are derivatives instruments that could have helped copying with the yield curve compressions of the time. We now examine the pricing details of these instruments.

13.8.7.1 FRA again

Interest rate swaps are baskets of FRAs dealt with in Section 13.2.2. Consider a FRA in which the fixed rate does not clear its value at origination: its payoff is then $\pi_{\text{fra}}\left(T, S\right)$ in Eq. (13.8).

Denote its value at origination, t, with $\text{FRA}(t,T,S;K)$, where $S-T$ is the debt-servicing period, and K is the fixed rate, viz

$$\begin{aligned}\text{FRA}(t,T,S;K) &= \mathbb{E}_t\left(e^{-\int_t^S r_\tau d\tau}(S-T)(L(T,S)-K)\right)\\ &= \mathbb{E}_t\left(e^{-\int_t^S r_\tau d\tau}(S-T)L(T,S)\right)-(S-T)P(t,S)K\\ &= \mathbb{E}_t\left(\frac{e^{-\int_t^S r_\tau d\tau}}{P(T,S)}\right)-(1+(S-T)K)P(t,S)\\ &= P(t,T)-(1+(S-T)K)P(t,S), \end{aligned} \tag{13.110}$$

where the third line holds by the definition of L and the fourth follows by arguments set forth in Section 13.2 (see Eq. (13.5)).

We may achieve the conclusions underlying Eq. (13.110) through a different route. Note that the LIBOR, $L(T,S)$, while known at T, is only paid off at S, such that the value of $1+(S-T)L(T,S)$ (to be delivered at S) is simply one at T and, obviously, $P(t,T)$ at t. That is, the value of $(S-T)L(T,S)$ (to be delivered at time S) is $P(t,T)-P(t,S)$ at t, whence, the fourth equality.[33]

We can achieve a more intuitive expression for $\text{FRA}(t,T,S;K)$. Replacing the definition of the forward rate in Eq. (13.6) into Eq. (13.110) leaves

$$\text{FRA}(t,T,S;K) = (S-T)(F(t,T,S)-K)P(t,S). \tag{13.111}$$

That is, the value of the FRA can take on any sign, and is exactly zero when $K=F(t,T,S)$, where $F(t,T,S)$ solves Eq. (13.6).

Interest rate swaps are baskets of FRAs, as mentioned: they consist in payment exchanges that occur repeatedly and over a given time horizon known as the swap tenor. We proceed with details.

13.8.7.2 Forward starting swaps

An interest rate swap is an exchange of interest rate payments. Two counterparties exchange a fixed against a floating interest rate payment. The floating payment is, typically, a short-term interest rate. For example, the party receiving a floating interest rate payment has "good" (or only) access to markets for "variable" interest rates, but wishes to pay fixed interest rates. Alternatively, the party receiving a floating interest rate is exposed to interest rate risks in the short-term segment, which it wants to hedge, as it could have been the case with the S&Ls during the 1980s.

The party receiving a floating interest rate payment and paying a fixed interest rate K_{irs} has access to a payoff equal to

$$\delta_{i-1}(L(T_{i-1},T_i)-K_{\text{irs}})$$

at time T_i, $i=1,\cdots,n$, where $\delta_{i-1}\equiv T_i-T_{i-1}$ as usual. Each of these payments is a FRA really, and can be evaluated as in the previous section. By convention, we say that the *swap*

[33] We are assuming, as it is standard, that settlements occur at S. When settlements occur at $T<S$, the value of $1+(S-T)L(T,S)$ at T is obviously higher than one, and the calculations leading to the last line of Eq. (13.110) and the "model-free" expression for F would not go through. The technical issue is a "convexity effect" by which a payoff of $1+(S-T)L(T,S)$ at time T is equivalent to a payoff of $(1+(S-T)L(T,S))^2$ at time S. Brigo and Mercurio (2006, Chapter 13) and Veronesi (2010, Chapter 21) explain the standard market practice to deal with this issue.

payer is the counterparty who pays the fixed interest rate K_{irs}, and that the *swap receiver* is the counterparty paying the variable interest rates.

With a dedicated interest swap of this kind, a S&L could have "locked-in the yield curve." Specifically, in this stylized example, the payoff for the S&L could have been

$$\delta_{i-1}\left(L^{\text{long}}\left(T_{i-1},T_i\right)-L\left(T_{i-1},T_i\right)\right)+\delta_{i-1}\left(L\left(T_{i-1},T_i\right)-K_{\text{irs}}\right)=\delta_{i-1}\left(L^{\text{long}}\left(T_{i-1},T_i\right)-K_{\text{irs}}\right),$$

where $L^{\text{long}}\left(T_{i-1},T_i\right)$ is the interest rate gained over long-term assets. Naturally, if short-term interest rates had to go down, relative to K_{irs}, the S&L would not have benefited from the increased long-term/short-term spread, $\delta_{i-1}\left(L^{\text{long}}\left(T_{i-1},T_i\right)-L\left(T_{i-1},T_i\right)\right)$. Yet, insuring against yield curve inversions is the thing to do, if yield curve inversions lead to bankruptcy and bankruptcy is costly. Below, we explain that additional products exist (caps, swaptions), which ensure against the upside while at the same time freeing up the downside.

In turn, there are institutions that may like to be on the opposite side of these deals. For example, pension funds may be interested in swap receivers, as anticipated in Chapter 10 (see Section 10.10). First, pension funds want to have access to long-term assets that distribute fixed payments, ideally, "coupons" (needed to honour payments to past funds' contributors). For all intents and purposes, a high-tenor receiver swap may play such a role: a fund may simply pay a variable interest rate (obtained by investing current contributions into a variable interest rate market) against the payment of fixed interest rates. Second, and as explained in the previous chapter, pension funds are quite adverse to decreasing interest rates, a scenario consistent with losses; a pension fund is particularly averse to such losses.

How do we evaluate an interest rate swap? The value at t of a forward starting swap (i.e. a swap starting at $T>t$) *payer*, $\upsilon_{\text{irs}}\left(t\right)$ say, is

$$\upsilon_{\text{irs}}\left(t\right)=\sum_{i=1}^{n}\mathbb{E}_t\left(e^{-\int_t^{T_i}r_\tau d\tau}\delta_{i-1}\left(L\left(T_{i-1},T_i\right)-K_{\text{irs}}\right)\right)=\sum_{i=1}^{n}\text{FRA}(t,T_{i-1},T_i;K_{\text{irs}}),\qquad(13.112)$$

where, by Eq. (13.111),

$$\text{FRA}\left(t,T_{i-1},T_i;K_{\text{irs}}\right)=\delta_{i-1}\left(F\left(t,T_{i-1},T_i\right)-K_{\text{irs}}\right)P\left(t,T_i\right).$$

The *forward swap rate* R_{sw} is the value of K_{irs} such that $\upsilon_{\text{irs}}(t)=0$. Simple calculations yield

$$R_{\text{sw}}\left(t\right)=\frac{\sum_{i=1}^{n}\delta_{i-1}F\left(t,T_{i-1},T_i\right)P\left(t,T_i\right)}{\sum_{i=1}^{n}\delta_{i-1}P\left(t,T_i\right)}=\frac{P\left(t,T_0\right)-P\left(t,T_n\right)}{\sum_{i=1}^{n}\delta_{i-1}P\left(t,T_i\right)},\qquad(13.113)$$

where the last equality holds by Eq. (13.6) in Section 13.2: $\delta_{i-1}F\left(t,T_{i-1},T_i\right)P\left(t,T_i\right)=P\left(t,T_{i-1}\right)-P\left(t,T_i\right)$.[34] This expression collapses to the par coupon rate derived in Chapter 12 (see Section 12.2), once we set $t=T_0$. That is, the *spot* swap rate is a par yield.

The first equality of Eq. (13.113) indicates that the forward swap rate is a linear combination of forward rates. This property has been used by Rebonato (1998) to deal with approximations of swaptions values based on the market model (see Section 13.9.4), as discussed in length by Brigo and Mercurio (2006, pp. 247-249).

[34]To cast this problem in terms of continuous time swap exchanges and, then, PDEs, we set $\upsilon_{\text{irs}}(T)\equiv 0$ as a boundary condition, and $\pi(r)=r-k$, where k plays the same role as K_{irs} above. Then, if the bond price $P(\tau)$ is solution to Eq. (13.94), the following function, $\upsilon_{\text{irs}}(\tau)=1-P(\tau)-k\int_\tau^T P(s)ds$, does also satisfy Eq. (13.94).

By plugging the expression for the forward swap rate in Eq. (13.113) into Eq. (13.112), we obtain the following intuitive expression for the forward swap payer:

$$\begin{aligned}\upsilon_{\text{irs}}(t) &= \sum_{i=1}^{n} \delta_{i-1} F(t, T_{i-1}, T_i) P(t, T_i) - K_{\text{irs}} \sum_{i=1}^{n} \delta_{i-1} P(t, T_i) \\ &= \sum_{i=1}^{n} \delta_{i-1} P(t, T_i) (R_{\text{sw}}(t) - K_{\text{irs}}) \\ &\equiv \text{PVBP}_t(T_1, \cdots, T_n) (R_{\text{sw}}(t) - K_{\text{irs}}), \end{aligned} \tag{13.114}$$

where $\text{PVBP}_T(T_1, \cdots, T_n)$ is the so-called swap "Present Value of the Basis Point," i.e. the present value impact of one basis point move in the forward swap rate at T.

13.8.7.3 Marking to market

While a forward starting swap is costless at origination, its marking to market updates are calculated as follows. Suppose to enter at τ into a forward starting swap originated at time t. The value of this position is,

$$\begin{aligned}&\upsilon_{\text{irs}}^{\text{m-t-m}}(\tau) \\ &\equiv \sum_{i=1}^{n} \mathbb{E}_\tau \left(e^{-\int_\tau^{T_i} r_s ds} \delta_{i-1} (L(T_{i-1}, T_i) - R_{\text{sw}}(t)) \right) \\ &= \sum_{i=1}^{n} \mathbb{E}_\tau \left(e^{-\int_\tau^{T_i} r_s ds} \delta_{i-1} (L(T_{i-1}, T_i) - R_{\text{sw}}(\tau) + R_{\text{sw}}(\tau)) \right) - \text{PVBP}_t(T_1, \cdots, T_n) \cdot R_{\text{sw}}(t) \\ &= \text{PVBP}_\tau(T_1, \cdots, T_n) \cdot (R_{\text{sw}}(\tau) - R_{\text{sw}}(t)), \end{aligned}$$

where the third equality follows by the definition of a forward swap at τ.

13.8.8 Caps & floors

A cap works as an interest rate swap, with the exception that the exchange of interest rates payments takes place only if actual interest rates are higher than K. A cap protects against upward movements in the interest rates, while freeing up the downside. By going long a cap, the S&L institution in the example of the previous section, then, would benefit from the downside in the short-term interest rates through a cap on them, literally. Precisely, a cap is made up of *caplets*. The payoff as of time T_i of a caplet is:

$$\delta_{i-1} (L(T_{i-1}, T_i) - K)^+, \quad i = 1, \cdots, n.$$

Floors are defined in a similar way. They are baskets of single *floorlets* that pay off $\delta_{i-1} (K - L(T_{i-1}, T_i))^+$ at time T_i, $i = 1, \cdots, n$.

We only discuss caps. By the FTAP, the a cap at time t is valued

$$p_{\text{cap}}(t) \equiv \sum_{i=1}^{n} \mathbb{E}_t \left(e^{-\int_t^{T_i} r_\tau d\tau} \delta_{i-1} (L(T_{i-1}, T_i) - K)^+ \right). \tag{13.115}$$

We can rely on models of the short-term rate to address this problem analytically. Consider the definition of simply compounded rates (see Eq. (13.1) in Section 13.2), viz $\delta_{i-1} L(T_{i-1}, T_i) =$

$\frac{1}{P(T_{i-1},T_i)}-1$, and re-write the caplet payoff as follows

$$\left(\delta_{i-1}L\left(T_{i-1},T_i\right)-\delta_{i-1}K\right)^+=\frac{1}{P\left(T_{i-1},T_i\right)}\left(1-(1+\delta_{i-1}K)P(T_{i-1},T_i)\right)^+.$$

We have

$$\begin{aligned}p_{\text{cap}}\left(t\right)&=\sum_{i=1}^{n}\mathbb{E}_t\left(\frac{e^{-\int_t^{T_i}r_\tau d\tau}}{P(T_{i-1},T_i)}\left(1-(1+\delta_{i-1}K)P(T_{i-1},T_i)\right)^+\right)\\&=\sum_{i=1}^{n}\mathbb{E}_t\left(e^{-\int_t^{T_{i-1}}r_\tau d\tau}\frac{1}{\mathcal{K}_i}\left(\mathcal{K}_i-P(T_{i-1},T_i)\right)^+\right),\quad \mathcal{K}_i=\left(1+\delta_{i-1}K\right)^{-1},\end{aligned}\tag{13.116}$$

where the last equality follows by a simple calculation.[35] Jamshidian and Hull & White models predict closed-form expressions bond prices and, then, for the cap price in Eq. (13.116). Precisely, Eq. (13.116) suggests that a cap is a basket of puts on zero coupon bonds, with strikes $\mathcal{K}_i$. As such, it can be priced in closed form, using the models in Sections 13.8. We have

$$p_{\text{cap}}\left(t\right)=\sum_{i=1}^{n}\frac{1}{\mathcal{K}_i}\text{Put}\left(t,T_{i-1};P\left(t,T_i\right),\mathcal{K}_i,v\right),\tag{13.117}$$

where $\text{Put}\left(\cdot\right)$ satisfies the put-call parity in Eq. (13.101), and, by the pricing formulae in Section 13.8.4,

$$\begin{aligned}&\text{Call}\left(t,T_{i-1};P\left(t,T_i\right),\mathcal{K}_i,v\right)=P\left(t,T_i\right)\Phi\left(d_{1,i}\right)-\mathcal{K}_iP\left(t,T_{i-1}\right)\Phi\left(d_{1,i}-v\right),\\&d_{1,i}=\frac{\ln\frac{P(t,T_i)}{\mathcal{K}_iP(t,T_{i-1})}+\frac{1}{2}v^2}{v},\quad v=\sigma\sqrt{\frac{1-e^{-2\kappa(T_{i-1}-t)}}{2\kappa}}B\left(T_{i-1},T_i\right),\quad B\left(t,T\right)=\tfrac{1}{\kappa}\left(1-e^{-\kappa(T-t)}\right).\end{aligned}\tag{13.118}$$

Naturally, caps on interest rates, which are nothing but baskets of calls, are portfolios of puts on fixed coupon bonds, due to the inverse relation between prices and interest rates.[36]

13.8.9 Swaptions

Let us elaborate on the S&L example of the previous sections. The benefits for the S&L to be long a cap is to be protected against upward movements in the short-term rates while ensuring the downside is freed up. These benefits arise period per period, so to speak: a cap is a basket

[35] By the law of iterated expectations,

$$\begin{aligned}\mathbb{E}_t\left[\frac{e^{-\int_t^{T_i}r_\tau d\tau}}{P(T_{i-1},T_i)}\left[1-\mathcal{K}_iP(T_{i-1},T_i)\right]^+\right]&=\mathbb{E}_t\left[\mathbb{E}\left(\frac{e^{-\int_t^{T_i}r_\tau d\tau}}{P(T_{i-1},T_i)}\left(1-\mathcal{K}_iP(T_{i-1},T_i)\right)^+\middle|\mathcal{F}_{T_i}\right)\right]\\&=\mathbb{E}_t\left[\mathbb{E}\left(e^{-\int_t^{T_i}r_\tau d\tau}e^{\int_{T_{i-1}}^{T_i}r_\tau d\tau}\left(1-\mathcal{K}_iP(T_{i-1},T_i)\right)^+\middle|\mathcal{F}_{T_i}\right)\right]\\&=\mathbb{E}_t\left[\mathbb{E}\left(e^{-\int_t^{T_{i-1}}r(\tau)d\tau}\left(1-\mathcal{K}_iP(T_{i-1},T_i)\right)^+\middle|\mathcal{F}_{T_i}\right)\right]\\&=\mathbb{E}_t\left[e^{-\int_t^{T_{i-1}}r_\tau d\tau}\left(1-\mathcal{K}_iP(T_{i-1},T_i)\right)^+\right]\end{aligned}$$

[36] We might also price caps and floors through the partial differential equation (13.94), after setting $\pi\left(r\right)=\left(r-k\right)^+$ (caps) and $\pi\left(r\right)=\left(k-r\right)^+$ (floors), for some strike k. However, this type of contracts, where payoffs are paid continuously in time, is highly stylized, and does not exist in the markets.

of options with different maturities. A swaption works differently, in that the optionality kicks in "all together."

The S&L is now concerned about *future* inversions of the yield curve; therefore, it anticipates its need to go long a swap payer at some future date. At the same time, the S&L fears that in the future, swap rates will be lower relative to some reference strike. Swaptions allow to free up such a downside risk, because they offer the option to enter a swap contract on a future date. Let T_0 be the maturity date of this option. Come time T_0, the payoff for a *payer* swaption is the maximum between zero and the value of a *payer* interest rate swap at T_0, $\upsilon_{\text{irs}}(T_0)$, viz

$$\left(\upsilon_{\text{irs}}\left(T_0\right)\right)^+ = \left(\sum_{i=1}^{n} \text{FRA}\left(T_0, T_{i-1}, T_i; K_{\text{irs}}\right)\right)^+ = \left(\sum_{i=1}^{n} \delta_{i-1}\left(F\left(T_0, T_{i-1}, T_i\right) - K_{\text{irs}}\right) P\left(T_0, T_i\right)\right)^+. \tag{13.119}$$

By the FTAP, the value of the payer swaption at time t is

$$\begin{aligned} \text{SW}_p\left(t\right) &= \mathbb{E}_t\left[e^{-\int_t^{T_0} r_\tau d\tau}\left(\sum_{i=1}^{n} \delta_{i-1}\left(F(T_0, T_{i-1}, T_i) - K_{\text{irs}}\right) P(T_0, T_i)\right)^+\right] \\ &= \mathbb{E}_t\left[e^{-\int_t^{T_0} r_\tau d\tau}\left(1 - P(T_0, T_n) - K_{\text{irs}}\sum_{i=1}^{n} \delta_{i-1} P(T_0, T_i)\right)^+\right], \end{aligned} \tag{13.120}$$

where we used the relation $\delta_{i-1}F(T_0, T_{i-1}, T_i) = \frac{P(T_0,T_{i-1})}{P(T_0,T_i)} - 1$.

Eq. (13.120) is the expression for the price of a *put* option *on a fixed coupon bond* struck at one. Therefore, we can price this contract in closed-form, through the models in Section 13.8.4, similarly to that we did in the previous section while pricing caps. We have:

$$\text{SW}_p\left(t\right) = \overline{\text{Put}}\left(t, T_0; P_{\text{fcb}}\left(t, T_n\right), 1, v\right),$$

where $\overline{\text{Put}}\left(\cdot\right)$ satisfies the put-call parity in Eq. (13.101). By the pricing formulae in Section 13.8.4,

$$\overline{\text{Call}}\left(t, T_0; P_{\text{fcb}}\left(t, T_n\right), 1, v\right) = K_{\text{irs}}\sum_{i=1}^{n} \delta_{i-1}\text{Call}\left(t, T_0; P\left(t, T_i\right), P_i^*, v\right) + \text{Call}\left(t, T_0; P\left(t, T_n\right), P_n^*, v\right),$$

where $\text{Call}\left(t, T_0; P\left(t, T_i\right), P_i^*, v\right)$ is as in Eq. (13.118), with $P_i^* = P\left(r^*, T_0, T_i\right)$, and r^* solution to Eq. (13.107) for $K = 1$.

13.9 Market models

13.9.1 Models and market practice

The previous section illustrates that models of the short-term rate (along woth their perfectly fitting extensions) lead to closed-form expressions for the price of important interest rate sensitive products. Yet practitioners evaluate caps, floors or swaptions through the Black's (1976) formula, as explained in detail below. The assumption underlying market practice is that the simply-compounded forward rate is lognormally distributed. As it turns out, the analytically tractable (Gaussian) short-term rate models are *not* consistent with this assumption. Clearly,

the (Gaussian) Vasicek model does not predict that the simply-compounded forward rates are Geometric Brownian motions.[37]

How to address these issues while relying on a non-Markovian HJM? Some qualifications are in order. A practical difficulty with HJM is that *instantaneous* forward rates are not observed, which at a first sight seems to be an hindrance to realistic pricing of caps, floors and swaptions. Brace, Gatarek and Musiela (1997), Jamshidian (1997) and Miltersen, Sandmann and Sondermann (1997) address this issue, and note that the HJM framework can be somehow "forced" to produce models ready to be used consitsently with market practice.

The key feature of the market models is the emphasis on the dynamics of the *simply-compounded* forward rates. One additional, and technical, assumption is that these simply-compounded forward rates are lognormal under the risk-neutral probability Q. That is, given a non-decreasing sequence of reset times $\{T_i\}_{i=0,1,\cdots}$, each simply-compounded rate, F_i, is solution to the following stochastic differential equation:[38]

$$\frac{dF_{i,\tau}}{F_{i,\tau}} = m_i(\tau)d\tau + \gamma_i(\tau)d\tilde{W}_\tau, \quad \tau \in [t, T_i], \quad i = 0, \cdots, n-1, \tag{13.121}$$

where to simplify notation, we have set, $F_{i\tau} \equiv F(\tau, T_i, T_{i+1})$, and m_i and γ_i are some deterministic functions of time (γ_i is vector valued). On a mathematical point of view, that assumption that F_i follows Eq. (13.121) is innocuous.[39]

As we shall show, this simple framework can be used to use the simple Black's (1976) formula to price caps and floors. However, we need to emphasize that there is nothing wrong with the short-term rate models analyzed in previous sections. The real advance of the so-called market model is to give a rigorous foundation to the standard market practice to price caps, floors and swaptions by means of the Black's (1976) formula.

13.9.2 Simply-compounded forward rate dynamics, and no-arb restrictions

By the definition of the simply-compounded forward rates in Eq. (13.6),

$$\ln\left(\frac{P(\tau, T_i)}{P(\tau, T_{i+1})}\right) = \ln(1 + \delta_i F_{i,\tau}). \tag{13.122}$$

The logic we follow, now, is the same as that underlying the HJM representation of Section 13.7. We wish to express the volatility of bond prices in terms of the volatility of forward rates. To achieve this task, we first assume that bond prices are driven by Brownian motions and expand the L.H.S. of Eq. (13.122) (step 1). Then, we expand the R.H.S. of Eq. (13.122) (step 2). Finally, we identify the two diffusion terms derived from the previous two steps (step 3).

Step 1: Let $P_i \equiv P(\tau, T_i)$, and assume that under the risk-neutral probability Q, P_i is solution to:

$$\frac{dP_i}{P_i} = rd\tau + \sigma_{bi}d\tilde{W}.$$

[37] Indeed, $1 + \delta_i F_i(\tau) = \frac{P(\tau,T_i)}{P(\tau,T_{i+1})} = \exp[\Delta A_i(\tau) - \Delta B_i(\tau) r(\tau)]$, where $\Delta A_i(\tau) = A(\tau, T_i) - A(\tau, T_{i+1})$, and $\Delta B_i(\tau) = B(\tau, T_i) - B(\tau, T_{i+1})$. Hence, $F_i(\tau)$ is not a Geometric Brownian motion, despite the fact that the short-term rate r is Gaussian and, hence, the bond price is log-normal. Black '76 can not be applied in this context.

[38] Brace, Gatarek and Musiela (1997) formulate a model in terms of the spot simply-compounded LIBOR interest rates. Because $F_i(T_i) = L(T_i)$, the two derivations are essentially the same.

[39] It is well-known that lognormal *instantaneous* forward rates are problematic as they imply the money market account is ill-behaved. Sandmann and Sondermann (1997) provide a succinct overview on how this problem is handled with *simply-compounded* forward rates.

In terms of the HJM framework in Section 13.7,

$$\sigma_{bi}(\tau) = -\sigma^I(\tau, T_i) = -\int_\tau^{T_i} \sigma(\tau, \ell) d\ell, \tag{13.123}$$

where $\sigma(\tau, \ell)$ is the instantaneous volatility of the instantaneous ℓ-forward rate as of time τ. By Itô's lemma,

$$d \ln \left(\frac{P(\tau, T_i)}{P(\tau, T_{i+1})} \right) = -\frac{1}{2} \left(\|\sigma_{bi}\|^2 - \|\sigma_{b,i+1}\|^2 \right) d\tau + (\sigma_{bi} - \sigma_{b,i+1}) \, d\tilde{W}. \tag{13.124}$$

Step 2: Applying Itô's lemma to $\ln(1 + \delta_i F_{i,\tau})$, and using Eq. (13.121), yields:

$$\begin{aligned} d \ln (1 + \delta_i F_{i,\tau}) &= \frac{\delta_i}{1 + \delta_i F_i} dF_i - \frac{1}{2} \frac{\delta_i^2}{(1 + \delta_i F_i)^2} (dF_i)^2 \\ &= \left(\frac{\delta_i m_i F_i}{1 + \delta_i F_i} - \frac{1}{2} \frac{\delta_i^2 F_i^2 \|\gamma_i\|^2}{(1 + \delta_i F_i)^2} \right) d\tau + \frac{\delta_i F_i}{1 + \delta_i F_i} \gamma_i d\tilde{W}. \end{aligned} \tag{13.125}$$

Step 3: By Eq. (13.122), the diffusion terms in Eqs. (13.124) and (13.125) have to be the same. Therefore,

$$\sigma_{bi}(\tau) - \sigma_{b,i+1}(\tau) = \frac{\delta_i F_{i,\tau}}{1 + \delta_i F_{i,\tau}} \gamma_i(\tau), \quad \tau \in [t, T_i] \, .$$

By summing over i, we get the following no-arbitrage restriction applying to the volatility of the bond prices:

$$\sigma_{bi}(\tau) - \sigma_{b,0}(\tau) = -\sum_{j=0}^{i-1} \frac{\delta_j F_{j,\tau}}{1 + \delta_j F_{j,\tau}} \gamma_j(\tau). \tag{13.126}$$

Eq. (13.126) is, thus, a restriction to the general HJM framework. In other words, assume the instantaneous forward rates are as in Eq. (13.80) of Section 13.7. As shown in Section 13.7, the bond price volatility is, then, given by Eq. (13.123). But if we also assume that simply-compounded forward rates are solution to Eq. (13.121), then, the bond price volatility needs to equal that in Eq. (13.126). Comparing Eq. (13.123) with Eq. (13.126) produces,

$$\int_{T_0}^{T_i} \sigma(\tau, \ell) d\ell = \sum_{j=0}^{i-1} \frac{\delta_j F_{j,\tau}}{1 + \delta_j F_{j,\tau}} \gamma_j(\tau).$$

The practical interest to restrict the forward-rate volatility dynamics in this way lies in the possibility to obtain closed-form solutions for some of the interest rates derivatives surveyed in Section 13.8.

13.9.3 Applications to derivative evaluation

13.9.3.1 Forward rates as martingales

Forward rates are martingales under the forward probability. Regarding the continuously compounded notion, we have, by the definition of the instantaneous forward rate, the usual change

of probability and standard regularity conditions,

$$\begin{aligned} f(t,T) &= -\frac{\partial \ln P(t,T)}{\partial T} \\ &= -\frac{\partial P(t,T)}{\partial T} \Big/ P(t,T) \\ &= \mathbb{E}_t\left(\frac{e^{-\int_t^T r_\tau d\tau}}{P(t,T)} \cdot r_T\right) \\ &= \mathbb{E}_t\left(\eta_T(T) \cdot r_T\right) \\ &= \mathbb{E}_{Q_F^T}\left(r_T\right) \\ &= \mathbb{E}_{Q_F^T}\left(f(T,T)\right), \end{aligned}$$

where the last equality holds as $r_t = f(t,t)$ as usual.

Similarly, and regarding simply-compounded forward rates, we have that for a given sequence of dates $\{T_i\}_{i=0,1,\cdots}$,

$$F(\tau, T_i, T_{i+1}) = \mathbb{E}_{Q_F^{T_{i+1}}}\left(L(T_i, T_{i+1})\right) = \mathbb{E}_{Q_F^{T_{i+1}}}\left(F(T_i, T_i, T_{i+1})\right), \quad \tau \in [t, T_i], \tag{13.127}$$

where the second equality follows by Eq. (13.7), i.e. $L(T_i, T_{i+1}) = F(T_i, T_i, T_{i+1})$. To show Eq. (13.127), note that by definition, the simply-compounded forward rate $F(t,T,S)$ satisfies,

$$F(t,T,S) : 0 = \mathbb{E}_t\left(e^{-\int_t^S r_\tau d\tau} \cdot \left(L(T,S) - F(t,T,S)\right)\right),$$

as explained in Section 13.7.1. That is, by rearranging terms,

$$F(t,T,S) = \mathbb{E}_t\left(\frac{e^{-\int_t^S r_\tau d\tau}}{P(t,S)} L(T,S)\right) = \mathbb{E}_{Q_F^S}\left(L(T,S)\right). \tag{13.128}$$

We use this martingale property whilst pricing caps and floors.

13.9.3.2 Caps & floors

We provide analytical results for the price of caps only. We have:

$$\begin{aligned} p_{\text{cap}}(t) &= \sum_{i=1}^{n} \mathbb{E}_t\left(e^{-\int_t^{T_i} r_\tau d\tau} \delta_{i-1}\left(L(T_{i-1}, T_i) - K\right)^+\right) \\ &= \sum_{i=1}^{n} \mathbb{E}_t\left(e^{-\int_t^{T_i} r_\tau d\tau} \delta_{i-1}\left(F(T_{i-1}, T_{i-1}, T_i) - K\right)^+\right) \\ &= \sum_{i=1}^{n} \delta_{i-1} P(t, T_i) \cdot \mathbb{E}_{Q_F^{T_i}}\left(F(T_{i-1}, T_{i-1}, T_i) - K\right)^+, \end{aligned} \tag{13.129}$$

where $\mathbb{E}_{Q_F^{T_i}}(\cdot)$ denotes, as usual, the expectation taken under the T_i-forward martingale probability $Q_F^{T_i}$; the first equality is Eq. (13.115); and the second equality follows by a change of probability, from the risk-neutral to the forward.

A key point explained above is that $F_{i-1,\tau} \equiv F_{i-1}(\tau, T_{i-1}, T_i)$ is a martingale under $Q_F^{T_i}$ for $\tau \in [t, T_{i-1}]$ (see Eq. (13.128)), such that, by Eq. (13.121), $F_{i-1,\tau}$ satisfies

$$\frac{dF_{i-1,\tau}}{F_{i-1,\tau}} = \gamma_{i-1}(\tau)\, dW^{Q_F^{T_i}}(\tau), \quad \tau \in [t, T_{i-1}], \quad i = 1, \cdots, n,$$

under $Q_F^{T_i}$. Therefore, the cap price in Eq. (13.129) reduces to the Black's (1976) formula discussed in Chapter 11,[40] once we assume γ is deterministic

$$\mathbb{E}_{Q_F^{T_i}}\left(F(T_{i-1}, T_{i-1}, T_i) - K\right)^+ = F_{i-1,t}\Phi(d_{1,i-1}) - K\Phi(d_{1,i-1} - s_i), \tag{13.130}$$

where

$$d_{1,i-1} = \frac{\ln \frac{F_{i-1,t}}{K} + \frac{1}{2}s_i^2}{s_i}, \quad s_i^2 = \int_t^{T_{i-1}} \gamma_{i-1}^2(\tau)d\tau.$$

We may now price floors based on a put-call parity relation. Finding such a relation is left as an exercise.

13.9.3.3 Swaptions

By Eq. (13.114), the payoff of a *payer* swaption expiring at time T_0 is

$$\left(\upsilon_{\text{irs}}(T_0)\right)^+ = \text{PVBP}_{T_0}(T_1, \cdots, T_n)\left(R_{\text{sw}}(T_0) - K_{\text{irs}}\right)^+, \ \text{PVBP}_{T_0}(T_1, \cdots, T_n) = \sum_{i=1}^{n} \delta_{i-1} P(T_0, T_i).$$

Therefore, by the FTAP, and a change of probability,

$$\begin{aligned} \text{SW}_p(t) &= \mathbb{E}_t\left(e^{-\int_t^{T_0} r_\tau d\tau}\text{PVBP}_{T_0}(T_1, \cdots, T_n)\left(R_{\text{sw}}(T_0) - K_{\text{irs}}\right)^+\right) \\ &= \text{PVBP}_t(T_1, \cdots, T_n) \cdot \mathbb{E}_{Q_{\text{sw}}}\left(R_{\text{sw}}(T_0) - K_{\text{irs}}\right)^+, \end{aligned} \tag{13.131}$$

where $\mathbb{E}_{Q_{\text{sw}}}$ denotes the expectation taken under the so-called *forward swap probability*, defined by:

$$\left.\frac{dQ_{\text{sw}}}{dQ}\right|_{\mathbb{F}_{T_0}} = e^{-\int_t^{T_0} r_\tau d\tau}\frac{\text{PVBP}_{T_0}(T_1, \cdots, T_n)}{\text{PVBP}_t(T_1, \cdots, T_n)}.$$

It is easy to see that $\mathbb{E}\left(\left.\frac{dQ_{\text{sw}}}{dQ}\right|_{\mathbb{F}_{T_0}}\right) = 1$, by using the definition of $\text{PVBP}_{T_0}(T_1, \cdots, T_n)$, the pricing equation, $P(t, T_i) = \mathbb{E}_t\left(e^{-\int_t^{T_0} r_\tau d\tau} P(T_0, T_i)\right)$, as anticipated in Chapter 5. As also mentioned in Chapter 5, Q_{swap} is also sometimes referred to as *annuity probability*.

The key point underlying this change of probability is that the forward swap rate R_{swap} is a Q_{swap}-martingale,[41] and clearly, positive. Therefore, it must satisfy:

$$\frac{dR_{\text{sw}}(\tau)}{R_{\text{sw}}(\tau)} = \gamma_{\text{sw}}(\tau)\, dW_\tau^{\text{sw}}, \quad \tau \in [t, T_0], \tag{13.132}$$

[40] See Section 11.4 and Appendix 2 to Chapter 11.

[41] By Eq. (13.113), and one change of measure,

$$\mathbb{E}_{Q_{\text{swap}}}\left[R_{\text{swap}}(\tau)\right] = \mathbb{E}_{Q_{\text{swap}}}\left[\frac{P(\tau, T_0) - P(\tau, T_n)}{\text{PVBP}_\tau(T_1, \cdots, T_n)}\right] = \mathbb{E}_t\left[\frac{e^{-\int_t^\tau r(\tau)d\tau}\left(P(\tau, T_0) - P(\tau, T_n)\right)}{\text{PVBP}_t(T_1, \cdots, T_n)}\right] = \frac{P(t, T_0) - P(t, T_n)}{\text{PVBP}_t(T_1, \cdots, T_n)} = R_{\text{swap}}(t).$$

where W^{sw} is a Q_{sw}-Brownian motion, and $\gamma_{\text{sw}}(\tau)$ is adapted.

If the vol $\gamma_{\text{sw}}(\tau)$ in Eq. (13.132) is deterministic, we can use Black 76 to price the payer swaption in Eq. (13.131) in closed-form. We have:

$$\text{SW}_p(t) = \text{PVBP}_t(T_1, \cdots, T_n) \cdot \text{Black76}(R_{\text{sw}}(t); T_0, K_{\text{irs}}, \sqrt{\bar{V}}), \tag{13.133}$$

where $\text{Black76}(\cdot)$ is given by Black's (1976) formula:

$$\text{Black76}(R_{\text{sw}}(t); T_0, K_{\text{irs}}, \sqrt{\bar{V}}) = R_{\text{sw}}(t)\,\Phi(d_t) - K_{\text{irs}}\Phi(d_t - \sqrt{\bar{V}}),$$
$$d_t = \frac{\ln \frac{R_{\text{sw}}(t)}{K_{\text{irs}}} + \frac{1}{2}\bar{V}}{\sqrt{\bar{V}}}, \quad \bar{V} = \int_t^{T_0} \gamma_{\text{sw}}^2(\tau)\, d\tau.$$

13.9.3.4 A put-call parity in the swap space

We derive a put-call parity relation for swaptions. The reasoning is standard. Consider the following identity

$$(R_{\text{sw}}(T_0) - K_{\text{irs}})^+ = (K_{\text{irs}} - R_{\text{sw}}(T_0))^+ + R_{\text{sw}}(T_0) - K_{\text{irs}}.$$

Taking the expectations under Q_{sw} of both sides of the previous relation leaves

$$\text{SW}_p(t) = \text{SW}_r(t) + R_{\text{sw}}(t) - K_{\text{irs}}$$

where $\text{SW}_r(t)$ denotes a swaption receiver, struck at the same strike K_{irs}, and where we have used the martingale property, $\mathbb{E}_{Q_{\text{sw}}}(R_{\text{sw}}(T_0)) = R_{\text{sw}}(t)$.

13.9.3.5 Inconsistencies

If the forward rate is solution to Eq. (13.121), γ_{sw} cannot be deterministic. Unfortunately, if forward swap rates are lognormal, then, Eq. (13.121) does not hold. Therefore, we may use Black's formula to price either caps or swaptions, not both. This might limit the importance of market models. A couple of tricks that seem to work in practice. The best known is based on a suggestion by Rebonato (1998), to replace the true pricing problem with an approximating pricing problem where γ_{sw} is deterministic. That works in practice, but in a world with stochastic volatility, we should expect that trick to generate unstable things in periods experiencing highly volatile volatility. See, also, Rebonato (1999) for an essay on related issues. The next section suggests to use numerical approximation based on Montecarlo techniques.

13.9.3.6 Numerical approximations

Suppose forward rates are lognormal. Then, we can price caps using Black's formula and we proceed to price swaptions relying on the general HJM framework, as summarized by its restrictions in Eq. (13.126), and Montecarlo integration, as follows. By a change of probability,

$$\begin{aligned}\text{SW}_p(t) &= \mathbb{E}_t\left[e^{-\int_t^{T_0} r_\tau d\tau}\left(\sum_{i=1}^n \delta_{i-1}\left(F(T_0, T_{i-1}, T_i) - K\right) P(T_0, T_i)\right)^+\right] \\ &= P(t, T_0)\mathbb{E}_{Q_F^{T_0}}\left(\sum_{i=1}^n \delta_{i-1}\left(F(T_0, T_{i-1}, T_i) - K\right) P(T_0, T_i)\right)^+,\end{aligned}$$

where $F(T_0, T_{i-1}, T_i)$, $i = 1, \cdots, n$, can be simulated under $Q_F^{T_0}$.

Details are as follows. We know that $F_{i-1,\tau} \equiv F(\tau, T_{i-1}, T_i)$ is solution to,

$$\frac{dF_{i-1,\tau}}{F_{i-1,\tau}} = \gamma_{i-1}(\tau) dW_\tau^{Q_F^{T_i}}, \quad \tau \leq T_{i-1}, \tag{13.134}$$

and that by results in Appendix 3,

$$\begin{aligned} dW_\tau^{Q_F^{T_i}} &= dW_\tau^{Q_F^{T_0}} - (\sigma_{bi}(\tau) - \sigma_{b0}(\tau))\, d\tau \\ &= dW_\tau^{Q_F^{T_0}} + \sum_{j=0}^{i-1} \frac{\delta_j F_{j,\tau}}{1+\delta_j F_{j,\tau}} \gamma_j(\tau) d\tau, \end{aligned}$$

where the second line follows from Eq. (13.126). Replacing this into Eq. (13.134) leaves:

$$\frac{dF_{i-1,\tau}}{F_{i-1,\tau}} = \gamma_{i-1\tau} \sum_{j=0}^{i-1} \frac{\delta_j F_{j,\tau}}{1+\delta_j F_{j,\tau}} \gamma_j(\tau) d\tau + \gamma_{i-1}(\tau) dW_\tau^{Q_F^{T_0}}, \quad \tau \leq T_{i-1}, \quad i = 1, \cdots, n.$$

These can easily be simulated with the methods described in any standard textbook of this kind, such as that of Kloeden and Platen (1992).

13.10 Volatility surfaces

13.10.1 Implied volatilities

13.10.1.1 Caps & floors

Market practice quoting conventions rely on volatility surfaces stemming from the market models of the previous section, not on the approach in Section 13.8.7-13.8.9. In principle, the models in Section 13.8.7-13.8.9 may be used to produce volatility surfaces, albeit indirectly, after calibration of the two parameters κ and σ, as Eq. (13.117) indicates. However, it is easier to provide volatility surfaces in the first place, through the models of this section. Quite simply, practitioners use Eq. (13.130) and quote volatilities such that the market price of a cap equals to the value predicted by Eq. (13.130) using the desired implied volatility s_i. In Eq. (13.130),

$$s_i = \sqrt{T_{i-1} - t} \cdot \gamma(i),$$

for some $\gamma(i)$, although, then, practitioners simply quote the value of $\hat{\gamma}_n$ that satisfies:

$$\hat{\gamma}_n : p_{\text{cap}}^{\$}(t; n) = \sum_{i=1}^{n} \delta_{i-1} P(t, T_i) \cdot \text{Black76}\left(F_{i-1,t}; K, \hat{s}_{i,n}\right),$$

where $p_{\text{cap}}^{\$}(t; n)$ is the market price of the cap, and:

$$\text{Black76}\left(F_{i-1,t}; K, \hat{s}_{i,n}\right) = F_{i-1,t}\Phi\left(d_{1,i-1}^n\right) - K\Phi\left(d_{1,i-1}^n - \hat{s}_{i,n}\right),$$

$$d_{1,i-1}^n = \frac{\ln \frac{F_{i-1,t}}{K} + \frac{1}{2}\hat{s}_{i,n}^2}{\hat{s}_{i,n}}, \quad \hat{s}_{i,n} = \sqrt{T_{i-1} - t} \cdot \hat{\gamma}_n$$

Given n, we can bootstrap $\hat{\gamma}(i)$, i.e. we can recursively solve for $\hat{\gamma}(i)$, as follows:

$$0 = \sum_{i=1}^{n} \delta_{i-1} P(t, T_i) \cdot \left[\text{Black76}\left(F_{i-1,t}; K, \hat{s}_{i,n}\right) - \text{Black76}\left(F_{i-1,t}; K, \hat{s}_i\right)\right], \quad n = 1, \cdots, N,$$

where N is the latest available maturity, and $\hat{s}_i = \sqrt{T_{i-1} - t} \cdot \hat{\gamma}(i)$. The values of $\hat{\gamma}(i)$ amount to what is typically referred to as the term structure of caps volatilities.

13.10.1.2 Swaptions

As for swaptions, the situation is much simpler. The market practice is to quote swaptions through standard implied vols, i.e. those vols IV_t such that, once inserted into Eq. (13.133), delivers the swaption market price:

$$\text{SW}_p(t) = \text{PVBP}_t\left(T_1, \cdots, T_n\right) \cdot \text{Black76}\left(R_{\text{sw}}\left(t\right); T_0, K_{\text{irs}}, \text{IV}_t\right).$$

13.10.2 Local volatilities and SABR models

How to evaluate illiquid derivatives while making sure that the existing products are fitted without error? When the "existing products" are derivatives, we can apply the same methodology relying on local volatilities, and put forward in Chapter 11 (see Section 11.7).

Local volatility models suffer from a drawback pointed out by Hagan, Kumar, Lesnieki and Woodward (2002), which we now illustrate in the case of the modeling of swaptions. For simplicity, assume that the local volatility does not depend on calendar time, and denote it with $\sigma_{\text{loc}}\left(R_{n,\tau}\right)$, where $R_{n,\tau}$ is the forward swap rate at time τ for a given tenor T_n. Hagan, and Woodward (1999) show that for any maturity, the implied Black's volatility of a European-style option is:

$$\sigma_{\text{iv}}\left(K, R_n\right) = \sigma_{\text{loc}}\left(\tfrac{1}{2}\left(K + R_n\right)\right)\left(1 + \frac{1}{24}\frac{\sigma''_{\text{loc}}\left(\frac{1}{2}\left(K + R_n\right)\right)}{\sigma_{\text{loc}}\left(\frac{1}{2}\left(K + R_n\right)\right)}\left(R_n - K\right)^2 + \cdots\right.,$$

where the omitted terms are likely to be numerically negligible for practical purposes.

What are the dynamics of the market smile implied by this local volatility model? To illustrate, consider what happens to the first term in the previous expansion, say $\hat{\sigma}_{\text{iv}}\left(K, R_n\right)$, when the forward increases from R_n to $R_n + \Delta R_n$,

$$\hat{\sigma}_{\text{iv}}\left(K, R_n + \Delta R_n\right) \equiv \sigma_{\text{loc}}\left(\tfrac{1}{2}\left(K + R_n + \Delta R_n\right)\right) = \hat{\sigma}_{\text{iv}}\left(K + \Delta R_n, R_n\right).$$

In other words, provided σ_{loc} is decreasing, the local volatility model predicts that, as the forward R_n increases, the skew moves to the left, which might contradict market behavior. For example, let us assume the local volatility function is, $\sigma_{\text{loc}}\left(x\right) = 0.04 \cdot x^{-1/2}$. The left panel of Figure 13.12 plots the implied volatility $\hat{\sigma}_{\text{iv}}\left(K, R_n\right)$ for $R_n = 3\%$ (solid line) and $R_n = 4\%$ (dashed line).

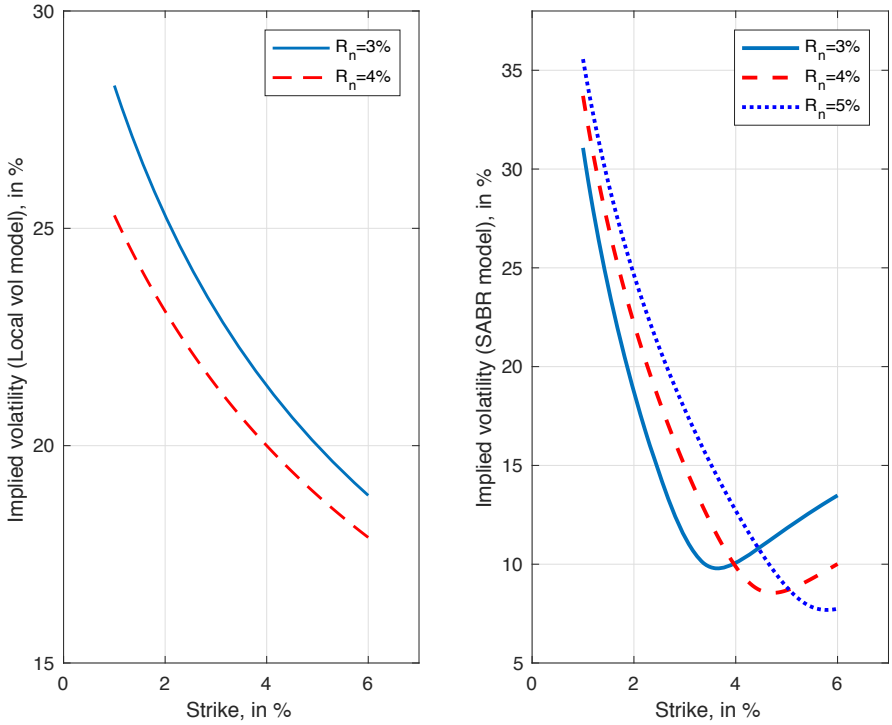


FIGURE 13.12. The left panel depicts the approximated implied volatility $\hat{\sigma}_{\text{iv}}(K, R_n)$ predicted by a local volatility model with $\sigma_{\text{loc}}(x) = 0.04 \cdot x^{-1/2}$; solid and dashed lines correspond to values of $R_{n,t}$ equal to 3% and 4%, respectively. The right panel depicts the approximated implied volatility $\sigma_{\text{iv}}(K, R_{n,t}; \alpha_t)$ in Eq. (13.136) predicted by the SABR model in Eq. (13.135), with $\alpha_t = 0.02$, $\rho = -0.5$, $\beta = 0.5$, $\xi = 0.5$, and time-to-maturity $T = 1$; solid, dashed and dotted lines correspond to values of $R_{n,t}$ equal to 3%, 4% and 5%, respectively.

Hagan, Kumar, Lesnieki and Woodward (2002) (HKLW in the sequel) consider a richer model, which they call SABR for "Stochastic-$\alpha\beta\rho$," in which $R_{n,t}$ satisfies,

$$\left\{ \begin{array}{lcl} dR_{n,\tau} & = & v_\tau R_{n,\tau}^{\beta} dW_{1\tau} \\ dv_\tau & = & \xi v_\tau \left(\rho dW_{1\tau} + \sqrt{1-\rho^2} dW_{2\tau} \right), \quad v_t \equiv \alpha_t \end{array} \right. \tag{13.135}$$

where $W_{i\tau}$ are two standard Brownian motions under the market probability, β, ρ and ξ are constants, and α_t is interpreted as the initial condition for the unobserved stochastic volatility component of the forward. Note that the model allows the forward and its volatility to be conditionally correlated with instantaneous correlation equal to ρ.

HKLW show that the implied volatility predicted by this model is,

$$\sigma_{\text{iv}}(K, R_{n,t}; \alpha_t) = \frac{\alpha_t}{(R_{n,t}K)^{(1-\beta)/2}} \frac{1 + \left(\frac{(1-\beta)^2}{24} \frac{\alpha_t^2}{(R_{n,t}K)^{1-\beta}} + \frac{1}{4} \frac{\rho\beta\xi}{(R_{n,t}K)^{(1-\beta)/2}} \alpha_t + \frac{2-3\rho^2}{24} \xi^2 \right) T + \cdots}{\left(1 + \frac{(1-\beta)^2}{24} \ln^2 \frac{R_{n,t}}{K} + \frac{(1-\beta)^4}{1920} \ln^4 \frac{R_{n,t}}{K} + \cdots \right.} \frac{z_t}{u(z_t)}, \tag{13.136}$$

where,

$$z_t \equiv \frac{\xi}{\alpha_t}\left(R_{n,t}K\right)^{(1-\beta)/2}\ln\frac{R_{n,t}}{K}, \quad u\left(z_t\right) \equiv \ln\frac{\sqrt{1-2\rho z_t+z_t^2}+z_t-\rho}{1-\rho}.$$

The right panel of Figure 13.12 depicts the approximated implied volatility predicted by the SABR model obtained with hypothetical parameter values. The model can fix the counterfactual behavior of the skew predicted by a local vol model: as $R_{n,t}$ increases, the implied volatility shifts to the right while at the same time generating a downward-sloping "backbone" (the curve traced by the at-the-money volatility as the forward varies).

The reason for a downward-sloping backbone is that the coefficient $\beta < 1$. HKLW also show the origins of the skew (i.e., the asymmetric smile) predicted by their model, due to (i) a coefficient $\beta < 1$, which makes the instantaneous volatility in Eq. (13.135), $vR_n^{\beta-1}$, decreasing in R_n, and (ii) a $\rho < 0$, which makes the transition density of the log-changes in $R_{n,t}$ skewed towards the left, as in classical explanations of Heston (1993) given in the equity case (see Chapter 11). Finally, the volatility of volatility parameter ξ helps determine the curvature of the skew. The implied volatility shifts up as α_t increases (option prices increase with volatility in this model (see Chapter 11), and so does implied volatility.

The SABR model is widely used in the market practice, especially while modeling the swaption skew. Note, however, that the model does not allow for a perfect matching of all available swaption prices, which by construction the local volatility can, at least theoretically. Finally, the comparative statics exercise in Figure 13.12 regards a change in $R_{n,t}$. Because $R_{n,t}$ is correlated with volatility, an alternative comparative statics exercise is one in which both the forward changes and volatility change in accordance with their assumed correlation (see Bartlett, 2006). Figure 13.13 shows that in this case with negative correlation, an increase in the forward accompanied by a decrease in the volatility (consistent with the negative correlation) implies the skew shifts toward the left although the backbone is still downward sloping.

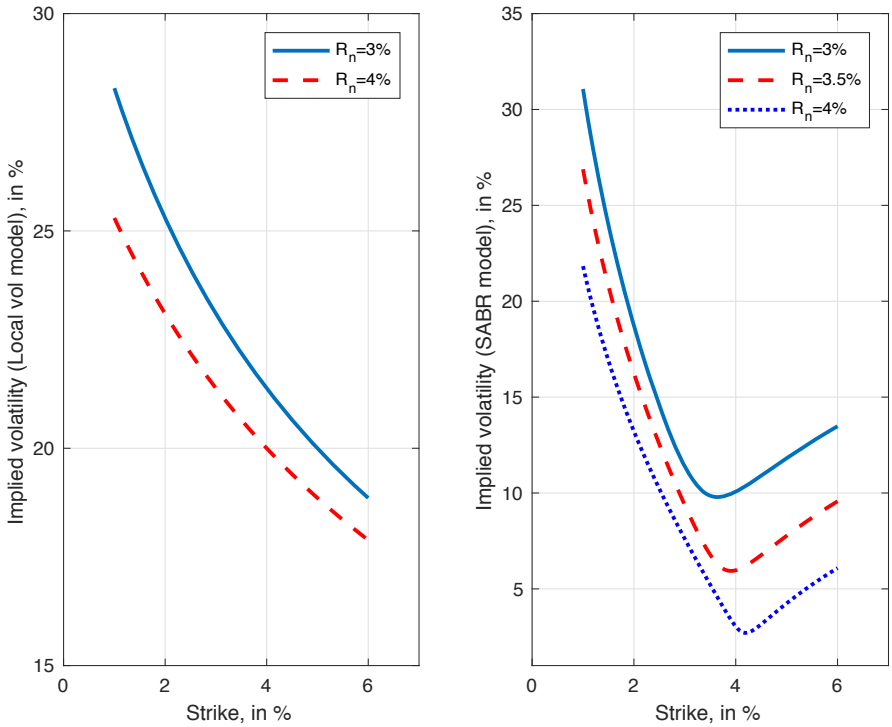


FIGURE 13.13. The left panel is the same as the left panel in Figure 13.12. The right panel depicts the approximated implied volatility $\sigma_{\text{iv}}(K, R_{n,t}; \alpha_t)$ in Eq. (13.136) predicted by the SABR model in Eq. (13.135), obtained with the parameters values used in Figure 13.12. Solid, dashed and dotted lines correspond to values of $R_{n,t}$ equal to 3%, 3.5% and 4%, respectively. The values of α_t in corrispondence of the three values of $R_{n,t}$ are $\alpha_t = 0.02$, and then, two increments obtained consistently with the negative correlation in Eq. (13.135), $\Delta\alpha = \frac{\rho\xi}{R_n^\beta}\Delta R_n$.

13.11 Appendix 1: The FTAP for bond prices

We assume that available for trading are m pure discount bonds with prices $P_i \equiv P(\tau, T_i)$ satisfying:

$$\frac{dP_i}{P_i} = \mu_{bi} \cdot d\tau + \sigma_{bi} \cdot dW, \;\; i = 1, \cdots, m, \tag{13A.1}$$

where W is a Brownian motion in $\mathbb{R}^d$, and μ_{bi} and σ_{bi} (σ_{bi} is vector-valued) are such that there exists a strong solution to the previous system. The value of a self-financing portfolio in these m bonds and a money market account satisfies:

$$dV = \left(\pi^\top(\mu_b - \mathbf{1}_m r) + rV\right) d\tau + \pi^\top \sigma_b dW,$$

where π is some portfolio, $\mathbf{1}_m$ is a m-dimensional vector of ones, and

$$\mu_b = [\mu_{b1}, \cdots, \mu_{b2}]^\top, \quad \sigma_b = [\sigma_{b1}, \cdots, \sigma_{b2}]^\top.$$

Now suppose there exists a portfolio $\underline{\pi}$ such that $\underline{\pi}^\top \sigma_b = 0$. This is an arbitrage opportunity if there exist events for which at some time, $\mu_b - \mathbf{1}_m r \neq 0$. (Use as usual, $\underline{\pi}$ when $\mu_b - \mathbf{1}_m r > 0$, and $-\underline{\pi}$ when $\mu_m - \mathbf{1}_d r < 0$: the drift of V will then be appreciating at a deterministic rate that is strictly greater than r.) Therefore, arbitrage is ruled out if:

$$\pi^\top(\mu_b - \mathbf{1}_m r) = 0 \text{ whenever } \pi^\top \sigma_b = 0.$$

In other terms, there is no arbitrage as soon as every vector in the null space of σ_b is orthogonal to $\mu_b - \mathbf{1}_m r$, or when there exists a λ in $\mathbb{R}^d$ and satisfying a few integrability conditions, and such that $\mu_b - \mathbf{1}_m r = \sigma_b \lambda$, or

$$\mu_{bi} - r = \sigma_{bi}\lambda, \;\; i = 1, \cdots, m. \tag{13A.2}$$

In this case,

$$\frac{dP_i}{P_i} = (r + \sigma_{bi}\lambda) \cdot d\tau + \sigma_{bi} \cdot dW, \;\; i = 1, \cdots, m.$$

Next, define $\tilde{W} = W + \int \lambda d\tau$, $\frac{dQ}{dP} = \exp(-\int_t^T \lambda^\top dW - \frac{1}{2}\int_t^T \|\lambda\|^2 d\tau)$. It is easy to show, now, that by Girsanov's theorem the discounted bond price is a martingale under Q. Indeed, define for a generic i, $P(\tau, T) \equiv P(\tau, T_i) \equiv P_i$, and:

$$g(\tau) \equiv e^{-\int_t^\tau r_u du} \cdot P(\tau, T), \quad \tau \in [t, T].$$

By Girsanov's theorem, and Itô's lemma,

$$\frac{dg}{g} = \sigma_{bi} \cdot d\tilde{W}, \quad \text{under } Q.$$

Therefore, for all $\tau \in [t, T]$, $g(\tau) = \mathbb{E}_t(g(T))$, implying that:

$$g(\tau) \equiv e^{-\int_t^\tau r_u du} \cdot P(\tau, T) = \mathbb{E}_t(g(T)) = \mathbb{E}_t[e^{-\int_t^T r_u du} \cdot \underbrace{P(T, T)}_{=1}] = \mathbb{E}_t\left(e^{-\int_t^T r_u du}\right),$$

or

$$P(\tau, T) = e^{\int_t^\tau r_u du} \cdot \mathbb{E}_t\left(e^{-\int_t^T r_u du}\right) = \mathbb{E}_t\left(e^{-\int_\tau^T r_u du}\right), \quad \text{all } \tau \in [t, T],$$

which is Eq. (13.2).

Note that no assumptions have been made regarding the number of zero-coupon bonds, m. For example, suppose that $m < d$, and that there are no other assets traded in this market. Then, there

exists an infinite number of risk-neutral proabilities Q. If $m = d$, there exists one and only one risk-neutral probability Q. If $m > d$, there exists one and only one risk-neutral probability but then, the various bond prices have to satisfy some basic no-arbitrage restrictions. For example, take $m = 2$ and $d = 1$, such that Eq. (13A.2) reduces to,

$$\frac{\mu_{b1} - r}{\sigma_{b1}} = \lambda = \frac{\mu_{b2} - r}{\sigma_{b2}}.$$

In other terms, the Sharpe ratio of any two bonds must be identical. Eq. (13A.2) is used several times in this chapter. In Section 13.4, the market primitive is the short-term rate, solution of a multidimensional diffusion process, and μ_{bi} and σ_{bi} are derived via Itô's lemma. In Section 13.7, μ_{bi} and σ_{bi} are restricted by a model for the forward rates.

13.12 Appendix 2: Certainty equivalent interpretation of forward prices

Multiply both sides of the pricing equation (13.2) by the amount S_T:

$$P(t,T)\cdot S_T=\mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau}\right)\cdot S_T.$$

Suppose momentarily that S_T is known at T. In this case, we have:

$$P(t,T)\cdot S_T=\mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau}\cdot S_T\right).$$

But in the context of this chapter, S_T is random. Define then its certainty equivalent by the number $\overline{S_T}$ that solves:

$$P(t,T)\cdot \overline{S_T}=\mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau}\cdot S_T\right),$$

or

$$\overline{S_T}=\mathbb{E}_t\left(\eta_T(T)\cdot S_T\right), \tag{13A.3}$$

where $\eta_T(T)$ has been defined in (13.97).

Comparing Eq. (13A.3) with Eq. (13.96) reveals that forward prices can be interpreted in terms of the previously defined certainty equivalent.

13.13 Appendix 3: Additional results on forward probabilities

Eq. (13.97) defines $\eta_T(T)$ as:

$$\eta_T(T) = \frac{e^{-\int_t^T r_\tau d\tau} \cdot 1}{\mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau}\right)}$$

More generally, define a *density process* as:

$$\eta_T(\tau) \equiv \frac{e^{-\int_t^\tau r_u du} \cdot P(\tau, T)}{\mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau}\right)}, \quad \tau \in [t, T]\,.$$

By the FTAP, the numerator of $\eta_T(\tau)$ is a martingale under Q (see Appendix 1 to this chapter). Therefore, $\mathbb{E}[\frac{dQ_F^T}{dQ}\Big|\,\mathcal{F}_\tau] = \mathbb{E}[\eta_T(T)|\,\mathcal{F}_\tau] = \eta_T(\tau)$ all $\tau \in [t, T]$, and in particular, $\eta_T(t) = 1$.

We demonstrate these claims under a slightly different angle. Let us consider the price dynamics of a zero-coupon bond in Eq. (13A.1), $P(\tau, T) \equiv P(\tau, T_i)$,

$$\frac{dP}{P} = \mu \cdot d\tau + \sigma \cdot dW,$$

where we have defined $\mu \equiv \mu_{bi}$ and $\sigma \equiv \sigma_{bi}$.

Under the risk-neutral probability Q,

$$\frac{dP}{P} = r \cdot d\tau + \sigma \cdot d\tilde{W},$$

where $\tilde{W} = W + \int \lambda$ is a Q-Brownian motion.

By Itô's lemma,

$$\frac{d\eta_T(\tau)}{\eta_T(\tau)} = -\left(-\sigma\left(\tau, T\right)\right) \cdot d\tilde{W}_\tau, \;\; \eta_T(t) = 1.$$

The solution is:

$$\eta_T(\tau) = \exp\left(-\frac{1}{2}\int_t^\tau \|\sigma(u, T)\|^2\, du - \int_t^\tau \left(-\sigma(u, T)\right) \cdot d\tilde{W}_u\right).$$

Under the usual integrability conditions, we can use the Girsanov's theorem and conclude that

$$W_\tau^{Q_F^T} \equiv \tilde{W}_\tau + \int_t^\tau \left(-\sigma(u, T)^\top\right) du \tag{13A.4}$$

is a Brownian motion under the T-forward probability Q_F^T.

Assuming for example that the driving state variable is the short-term rate, we have that the drift under Q_F^T of the same short-term rate is lower and that of the bond price is higher, due to negative bond price volatility, $\sigma < 0$.

Finally, note that for all i and non-decreasing sequences of dates $\{T_i\}_{i=0,1,\cdots,}$,

$$W_\tau^{Q_F^{T_i}} = \tilde{W}_\tau + \int_t^\tau \left(-\sigma(u, T_i)^\top\right) du, \;\; i = 0, 1, \cdots.$$

Therefore,

$$W_\tau^{Q_F^{T_i}} = W_\tau^{Q_F^{T_{i-1}}} - \int_t^\tau \left(\sigma(u, T_i)^\top - \sigma(u, T_{i-1})^\top\right) du, \;\; i = 1, 2, \cdots, \tag{13A.5}$$

is a Brownian motion under the T_i-forward martingale probability $Q_F^{T_i}$. Eqs. (13A.5) and (13A.4) are used in Section 13.8 on interest rate derivatives.

13.14 Appendix 4: Principal components analysis

Principal component analysis transforms the original data into a set of uncorrelated variables, the principal components, with variances arranged in descending order. Consider the following program,

$$\max_{C_1} \left(var\left(Y_1 \right) \right) \text{ s.t. } C_1^\top C_1 = 1,$$

where $var\left(Y_1 \right) = C_1^\top \Sigma C_1$, and the constraint is an identification constraint. The first order conditions lead to,

$$\left(\Sigma - \lambda I \right) C_1 = 0,$$

where λ is a Lagrange multiplier. The previous condition tells us that λ must be one eigenvalue of the matrix Σ, and that C_1 must be the corresponding eigenvector. Moreover, we have $var\left(Y_1 \right) = C_1^\top \Sigma C_1 = \lambda$ which is clearly maximized by the largest eigenvalue. Suppose that the eigenvalues of Σ are distinct, and let us arrange them in descending order, i.e. $\lambda_1 > \cdots > \lambda_p$. Then,

$$var\left(Y_1 \right) = \lambda_1.$$

Therefore, the first principal component is $Y_1 = C_1^\top \left(R - \bar{R} \right)$, where C_1 is the eigenvector corresponding to the largest eigenvalue, λ_1.

Next, consider the second principal component. The program is, now,

$$\max_{C_2} \left(var\left(Y_2 \right) \right) \text{ s.t. } C_2^\top C_2 = 1 \text{ and } C_2^\top C_1 = 0,$$

where $var\left(Y_2 \right) = C_2^\top \Sigma C_2$. The first constraint, $C_2^\top C_2 = 1$, is the usual identification constraint. The second constraint, $C_2^\top C_1 = 0$, is needed to ensure that Y_1 and Y_2 are orthogonal, i.e. $E\left(Y_1 Y_2 \right) = 0$. The first order conditions for this problem are,

$$0 = \Sigma C_2 - \lambda C_2 - \nu C_1$$

where λ is the Lagrange multiplier associated with the first constraint, and ν is the Lagrange multiplier associated with the second constraint. By pre-multiplying the first order conditions by $C_1^\top$,

$$0 = C_1^\top \Sigma C_2 - \nu,$$

where we have used the two constraints $C_1^\top C_2 = 0$ and $C_1^\top C_1 = 1$. Post-multiplying the previous expression by $C_1^\top$, one obtains, $0 = C_1^\top \Sigma C_2 C_1^\top - \nu C_1^\top = -\nu C_1^\top$, where the last equality follows by $C_1^\top C_2 = 0$. Hence, $\nu = 0$. So the first order conditions can be rewritten as,

$$\left(\Sigma - \lambda I \right) C_2 = 0.$$

The solution is now λ_2, and C_2 is the eigenvector corresponding to λ_2. (Indeed, this time we cannot choose λ_1 as this choice would imply that $Y_2 = C_1^\top \left(R - \bar{R} \right)$, implying that $E\left(Y_1 Y_2 \right) \neq 0$.) It follows that $var\left(Y_2 \right) = \lambda_2$.

In general, we have,

$$var\left(Y_i \right) = \lambda_i, \qquad i = 1, \cdots, p.$$

Let Λ be the diagonal matrix with the eigenvalues λ_i on the diagonal. By the spectral decomposition of Σ, $\Sigma = C \Lambda C^\top$, and by the orthonormality of C, $C^\top C = I$, we have that $C^\top \Sigma C = \Lambda$ and, hence,

$$\sum_{i=1}^p var\left(R_i \right) = \text{Tr}\left(\Sigma \right) = \text{Tr}\left(\Sigma C C^\top \right) = \text{Tr}\left(C^\top \Sigma C \right) = \text{Tr}\left(\Lambda \right).$$

Hence, Eq. (13.23) follows.

13.15 Appendix 5: Comparison results applying to defaultable bonds

Consider two markets A and B where the default intensities are v^A and v^B, and assume that the coefficients of L are independent of v^i. The pre-default bond price in market i is $P^i(r,\tau,T)$, and by Eq. (13.47), satisfies

$$0=\left(\frac{\partial}{\partial\tau}+L-r\right)P^i+v^i(\bar{P}^i-P^i),\quad i=A,B,$$

with the usual boundary condition. Assuming that $\bar{P}^A=\bar{P}^B$, subtracting these two equations, and rearranging terms, leads to the result that the price difference $\Delta P(r,\tau,T)\equiv P^A(r,\tau,T)-P^B(r,\tau,T)$ satisfies

$$0=\left(\frac{\partial}{\partial\tau}+L-(r+v^A)\right)\Delta P\left(r,\tau,T\right)+\left(v^A-v^B\right)\left[\bar{P}^B\left(\tau\right)-P^B\left(r,\tau,T\right)\right],$$

with the boundary condition $\Delta P\left(r,T,T\right)=0$ for all r. Because $\bar{P}^B<P^B$, one has that $\Delta P\left(r,\tau,T\right)<0$ whenever $v^A>v^B$, by an application of the maximum principle reviewed in Appendix 3 of Chapter 8.

13.16 Appendix 6: Details regarding exponential-affine models

13.16.1 Derivation of the closed-form solution

First, note that in this model, the vector Brownian motion under the risk-neutral probability is defined as $d\tilde{W}_\tau = dW_\tau + \Lambda(y_\tau)\,d\tau$, where the risk-premium vector Λ is as in Eq. (13.59), such that, assuming that $\Pr\{V(y_t)_{(ii)} > 0 \text{ all } t\} = 1$, the factor dynamics under the risk-neutral probability are

$$\begin{aligned} dy_\tau &= \left(\kappa(\mu - y_\tau) - \Sigma V^2(y_\tau)\lambda_1 - \Sigma\lambda_2 y_\tau\right)d\tau + \Sigma V(y_\tau)\,d\tilde{W}_\tau \\ &= \kappa_Q\left(\mu_Q - y_\tau\right)d\tau + \Sigma V(y_\tau)\,d\tilde{W}_\tau, \end{aligned} \tag{13A.6}$$

for some matrix κ_Q and vector μ_Q. Then, by Itô's lemma, Eq. (13A.6), and the assumption that the price function P is as in Eq. (13.60), P satisfies

$$\begin{aligned} \mathbb{E}_\tau\left(\frac{dP}{P}\right) &= \frac{1}{P}\left(\frac{\partial P}{\partial \tau} + P_y\kappa_Q\left(\mu_Q - y\right) + \frac{1}{2}\mathrm{Tr}\left([\Sigma V(y)][\Sigma V(y)]^\top P_{yy}\right)\right)d\tau \\ &= \left(-\dot{\mathcal{A}} - \dot{\mathcal{B}}y + \mathcal{B}\kappa_Q\left(\mu_Q - y\right) + \frac{1}{2}\sum_{i=1}^{n}(\mathcal{B}\Sigma)_i^2\,\alpha_i + \frac{1}{2}\sum_{i=1}^{n}(\mathcal{B}\Sigma)_i^2\,\beta_i y\right)d\tau \\ &= \left(r^0 + r^1 y\right)d\tau, \end{aligned}$$

where dots indicate differentiation to time-to-maturity $x \equiv T - \tau$, and the last equality follows by the standard martingale property of the discounted bond price, and by Eq. (13.56). Therefore, $\mathcal{A}(\cdot)$ and $\mathcal{B}(\cdot)$ are solutions to the following Riccati's equations

$$\begin{cases} \dot{\mathcal{A}}(x) = -r^0 + \mathcal{B}(x)\,\kappa_Q\mu_Q + \frac{1}{2}\sum_{i=1}^{n}(\mathcal{B}(x)\,\Sigma)_i^2\,\alpha_i \\ \dot{\mathcal{B}}(x) = -r^1 - \mathcal{B}(x)\,\kappa_Q + \frac{1}{2}\sum_{i=1}^{n}(\mathcal{B}(x)\,\Sigma)_i^2\,\beta_i^\top \end{cases} \tag{13A.7}$$

subject to the boundary conditions $\mathcal{A}(0) = 0$ and $\mathcal{B}(0) = 0$.

13.16.2 Application to Fong and Vasicek model

The Fong and Vasicek (1991) model in Section 13.5.1 (see Eqs. 13.48) is an affine model. Indeed, set $r_\tau = y_{1\tau}$ and $v_\tau = y_{2\tau}$ (i.e., $r^0 = 0$ and $r^1 = [0 \;\; 1]$), and note that under the assumptions in the main text, $y_{1\tau}$ and $y_{2\tau}$ are solutions to

$$\begin{cases} dy_{1\tau} = \left(\kappa_r\left(\bar{r} - y_{1\tau}\right) - \lambda_1 y_{2\tau}\right)d\tau + \sqrt{y_{2\tau}}\,d\tilde{W}_{1\tau} \\ dy_{2\tau} = \kappa_{Q,v}\left(\bar{v}_Q - y_{2\tau}\right)d\tau + \xi_v\sqrt{y_{2\tau}}\left(\rho\,d\tilde{W}_{1\tau} + \sqrt{1-\rho^2}\,d\tilde{W}_{2\tau}\right) \end{cases}$$

where $\kappa_{Q,v} \equiv \kappa_v + \xi_v(\lambda_1\rho + \sqrt{1-\rho^2}\lambda_2)$, $\bar{v}_Q \equiv \frac{\kappa_v}{\kappa_{Q,v}}\bar{v}$, $d\tilde{W}_{i\tau} = dW_{i\tau} + \lambda_i\sqrt{y_{2\tau}}d\tau$ are Brownian motions under Q, and λ_i are risk premium parameters defined in the main text. In terms of Eq. (13A.6), we have $\alpha_i = 0$ and $\beta_i^\top = \begin{bmatrix} 0 & \xi_v^2 \end{bmatrix}$ for $i = 1, 2$, and

$$\kappa_Q = \begin{bmatrix} \kappa_r & \lambda_1 \\ 0 & \kappa_{Q,v} \end{bmatrix}, \quad \mu_Q = \begin{bmatrix} \bar{r} - \bar{v}_Q\frac{\lambda_1}{\kappa_r} \\ \bar{v}_Q \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \frac{1}{\xi_v} & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}, \quad V(y_\tau) = \begin{bmatrix} \sqrt{\xi_v^2 y_{2\tau}} & 0 \\ 0 & \sqrt{\xi_v^2 y_{2\tau}} \end{bmatrix}.$$

Therefore, the model predicts that the price is as in Eq. (13.49), where the functions $A(\cdot)$, $B(\cdot)$ and $C(\cdot)$ satisfy Eqs. (13A.7), i.e.,

$$\begin{cases} \dot{A}(x) = -\kappa_r\bar{r}B(x) + \kappa_{Q,v}\bar{v}_Q C(x) \\ \dot{B}(x) = 1 - \kappa_r B(x) \\ \dot{C}(x) = \lambda_1 B(x) - \kappa_{Q,v}C(x) + \frac{1}{2}\left(B^2(x) - 2\rho\xi_v B(x)\,C(x) + \xi_v^2 C^2(x)\right) \end{cases} \tag{13A.8}$$

subject to the boundary conditions, $A(0) = B(0) = C(0) = 0$.

13.17 Appendix 7: A few analytical details regarding the Hull and White model

As in the Ho and Lee model, the instantaneous forward rate $f(\tau,T)$ predicted by the Hull and White model is as in Eq. (13.69), where functions A_2 and B_2 can be easily computed from Eqs. (13.74) and (13.75) as:

$$A_2(\tau,T)=\sigma^2\int_\tau^T B(s,T)B_2(s,T)ds-\int_\tau^T\theta_s B_2(s,T)ds,\quad B_2(\tau,T)=e^{-\kappa(T-\tau)}.$$

Therefore, the instantaneous forward rate $f(\tau,T)$ predicted by the Hull and White model is obtained by replacing the previous equations in Eq. (13.69). The result is then equated to the observed forward rate $f_{\$}(t,\tau)$ so as to obtain:

$$f_{\$}(t,\tau)=-\frac{\sigma^2}{2\kappa^2}\left(1-e^{-\kappa(\tau-t)}\right)^2+\int_t^\tau\theta_s e^{-\kappa(\tau-s)}ds+e^{-\kappa(\tau-t)}r_t.$$

By differentiating the previous equation with respect to τ, and rearranging terms,

$$\begin{aligned}\theta_\tau&=\frac{\partial}{\partial\tau}f_{\$}(t,\tau)+\frac{\sigma^2}{\kappa}\left(1-e^{-\kappa(\tau-t)}\right)e^{-\kappa(\tau-t)}+\kappa\left(\int_t^\tau\theta_s e^{-\kappa(\tau-s)}ds+e^{-\kappa(\tau-t)}r_t\right)\\&=\frac{\partial}{\partial\tau}f_{\$}(t,\tau)+\frac{\sigma^2}{\kappa}\left(1-e^{-\kappa(\tau-t)}\right)e^{-\kappa(\tau-t)}+\kappa\left(f_{\$}(t,\tau)+\frac{\sigma^2}{2\kappa^2}\left(1-e^{-\kappa(\tau-t)}\right)^2\right),\end{aligned}$$

which reduces to Eq. (13.76) after using simple algebra.

13.18 Appendix 8: Expectation theory and embedding in selected models

A. Expectation theory

Assume rhat

$$\sigma(\cdot,\cdot)=\sigma \quad \text{and} \quad \lambda(\cdot)=\lambda, \tag{13A.9}$$

where σ and λ are constants. We derive the dynamics of r, compare them with f, and formulate some basic claims regarding the expectation theory. We have:

$$r_\tau = f(t,\tau)+\int_t^\tau \alpha(s,\tau)ds+\sigma\left(W_\tau-W_t\right),$$

where

$$\alpha(\tau,T)=\sigma(\tau,T)\int_\tau^T \sigma(\tau,\ell)d\ell+\sigma(\tau,T)\lambda_\tau=\sigma^2(T-\tau)+\sigma\lambda.$$

Hence,

$$\int_t^\tau \alpha(s,\tau)ds=\frac{1}{2}\sigma^2\left(\tau-t\right)+\sigma\lambda(\tau-t).$$

Finally,

$$r_\tau = f(t,\tau)+\frac{1}{2}\sigma^2\left(\tau-t\right)+\sigma\lambda(\tau-t)+\sigma\left(W_\tau-W_t\right),$$

and since $E\left(W_\tau|\,\mathcal{F}_t\right)=W_t$,

$$E\left(r_\tau|\,\mathcal{F}_t\right)=f(t,\tau)+\frac{1}{2}\sigma^2\left(\tau-t\right)+\sigma\lambda(\tau-t).$$

Even with $\lambda<0$, this model does *not* guarantee that $E(r_\tau|\,\mathcal{F}_t)<f(t,\tau)$. As shown in the following example, this is due to the nonstationary nature of the volatility function. Indeed, suppose, next, that instead of Eq. (13A.9), we have that

$$\sigma(t,T)=\sigma\cdot\exp(-\gamma(T-t)) \quad \text{and} \quad \lambda(\cdot)=\lambda,$$

where σ,γ and λ are constants. In this case, we have:

$$r_\tau = f(t,\tau)+\int_t^\tau \alpha(s,\tau)ds+\sigma\int_t^\tau e^{-\gamma(\tau-s)}\cdot dW_s,$$

where

$$\alpha(s,\tau)=\sigma^2e^{-\gamma(\tau-s)}\int_s^\tau e^{-\gamma(\ell-s)}d\ell+\sigma\lambda e^{-\gamma(\tau-s)}=\frac{\sigma^2}{\gamma}\left(e^{-\gamma(\tau-s)}-e^{-2\gamma(\tau-s)}\right)+\sigma\lambda e^{-\gamma(\tau-s)}.$$

Finally,

$$E\left(r_\tau|\,\mathcal{F}_t\right)=f(t,\tau)+\int_t^\tau \alpha(s,\tau)ds=f(t,\tau)+\frac{\sigma}{\gamma}\left(1-e^{-\gamma(\tau-t)}\right)\left(\frac{\sigma}{2\gamma}\left(1-e^{-\gamma(\tau-t)}\right)+\lambda\right).$$

Therefore, it is sufficient to have a risk-premium such that $-\lambda>\frac{\sigma}{2\gamma}$, to generate the prediction that:

$$E\left(r_\tau|\,\mathcal{F}_t\right)<f(t,\tau) \quad \text{for any } \tau.$$

In other words, $\lambda<0$ is a necessary condition, not sufficient. Notice that when $\lambda=0$, it always holds that $E\left(r_\tau|\,\mathcal{F}_t\right)>f(t,\tau)$.

B. Embedding

We now embed the Ho and Lee model in Section 13.6.2 in the HJM format. In the Ho and Lee model,

$$dr_\tau = \theta_\tau d\tau + \sigma d\tilde{W}_\tau,$$

where $\tilde{W}$ is a Q-Brownian motion. By Eq. (13.69) in Section 13.6,

$$f(r,t,T) = -A_2(t,T) + B_2(t,T)r,$$

where $A_2(t,T) = \int_t^T \theta_s ds - \frac{1}{2}\sigma^2(T-t)$ and $B_2(t,T) = 1$. Therefore, by Eqs. (13.88),

$$\begin{aligned} \sigma(t,T) &= B_2(t,T)\cdot\sigma = \sigma, \\ \alpha(t,T) - \sigma(t,T)\lambda_t &= -A_{12}(t,T) + B_{12}(t,T)r + B_2(t,T)\theta_t = \sigma^2(T-t). \end{aligned}$$

Next, we embed the Vasicek model in Section 13.6 into the HJM format. The Vasicek model is:

$$dr_\tau = (\theta - \kappa r_\tau)d\tau + \sigma d\tilde{W}_\tau,$$

where $\tilde{W}$ is a Q-Brownian motion. By results in Section 13.4,

$$f(r,t,T) = -A_2(t,T) + B_2(t,T)r,$$

where $-A_2(t,T) = -\sigma^2\int_t^T B(s,T)B_2(s,T)ds + \theta\int_t^T B_2(s,T)ds$, $B_2(t,T) = e^{-\kappa(T-t)}$ and $B(t,T) = \frac{1}{\kappa}\left(1-e^{-\kappa(T-t)}\right)$. By Eqs. (13.88),

$$\sigma(t,T) = \sigma\cdot B_2(t,T) = \sigma\cdot e^{-\kappa(T-t)};$$

$$\alpha(t,T) - \sigma(t,T)\lambda_t = -A_{12}(t,T) + B_{12}(t,T)r + (\theta-\kappa r)B_2(t,T) = \frac{\sigma^2}{\kappa}\left(1-e^{-\kappa(T-t)}\right)e^{-\kappa(T-t)}.$$

Clearly this model cannot be embedded within HJM because it is not perfectly fitting. Condition (13.89) does not hold for Vasicek model. However, the model is embeddable once θ is turned into an infinite dimensional parameter *à la* Hull and White (see Section 13.4).

13.19 Appendix 9: Additional results on string models

We prove Eq. (13.91). We have, $\alpha^I(\tau,T)=\frac{1}{2}\int_\tau^T g(\tau,T,\ell_2)\,d\ell_2+cov(\frac{dP}{P},\frac{d\xi}{\xi})$, where

$$g(\tau,T,\ell_2)\equiv\int_\tau^T \sigma(\tau,\ell_1)\,\sigma(\tau,\ell_2)\,\psi(\ell_1,\ell_2)\,d\ell_1.$$

Differentiation of the *cov* term is straightforward. Moreover,

$$\begin{aligned}\frac{\partial}{\partial T}\int_\tau^T g(\tau,T,\ell_2)\,d\ell_2 &= g(\tau,T,T)+\int_\tau^T \frac{\partial g(\tau,T,\ell_2)}{\partial T}d\ell_2\\ &=\sigma(\tau,T)\cdot\int_\tau^T \sigma(\tau,x)\left(\psi(x,T)+\psi(T,x)\right)dx\\ &=2\sigma(\tau,T)\cdot\int_\tau^T \sigma(\tau,x)\,\psi(x,T)\,dx.\end{aligned}$$

13.20 Appendix 9: Changes of numéraire and Jamshidian's (1989) formula

Consider the following *change-of-numéraire* arithmetics. Let

$$\frac{dX_j}{X_j} = \mu_j d\tau + \sigma_j dW, \quad j \in \{A, B\}.$$

We have:

$$\frac{d(X_A/X_B)}{X_A/X_B} = \left(\mu_A - \mu_B + \sigma_B^2 - \sigma_A\sigma_B\right) d\tau + (\sigma_A - \sigma_B)\, dW. \tag{13A.10}$$

Next, we apply this result to the process $y(\tau, S) \equiv \frac{P(\tau,S)}{P(\tau,T)}$, under Q_F^S as well as under Q_F^T. We aim to derive the solution of $y(\tau, S)$ at T for Vasicek, viz

$$y(T, S) \equiv \frac{P(T, S)}{P(T, T)} = P(T, S) \text{ under } Q_F^S \text{ as well as under } Q_F^T.$$

This solution is needed to calculate the two probabilities in Eq. (13.102).

By Itô's lemma, and the fact that $P_r = -BP$, we have that under the risk-neutral probability Q,

$$\frac{dP(\tau, x)}{P(\tau, x)} = r d\tau - \sigma B(\tau, x) d\tilde{W}_\tau, \;\; x \geq T.$$

By applying Eq. (13A.10) to $y(\tau, S)$,

$$\frac{dy(\tau, S)}{y(\tau, S)} = \sigma^2 \left[B^2(\tau, T) - B(\tau, T)B(\tau, S)\right] d\tau - \sigma \left[B(\tau, S) - B(\tau, T)\right] d\tilde{W}_\tau. \tag{13A.11}$$

Next, we change probability by hinging upon the tools of Appendix 3. We have,

$$dW_\tau^{Q_F^x} = d\tilde{W}_\tau + \sigma B(\tau, x)\, d\tau$$

is a Brownian motion under the x-forward martingale probability. Replace then $W^{Q_F^x}$ into Eq. (13A.11), then integrate, and obtain:

$$\frac{y(T, S)}{y(t, S)} = P(T, S)\frac{P(t, T)}{P(t, S)} = e^{-\frac{1}{2}\sigma^2 \int_t^T [B(\tau,S) - B(\tau,T)]^2 d\tau - \sigma \int_t^T [B(\tau,S) - B(\tau,T)] dW_\tau^{Q_F^T}},$$
$$\frac{y(T, S)}{y(t, S)} = P(T, S)\frac{P(t, T)}{P(t, S)} = e^{\frac{1}{2}\sigma^2 \int_t^T [B(\tau,S) - B(\tau,T)]^2 d\tau - \sigma \int_t^T [B(\tau,S) - B(\tau,T)] dW_\tau^{Q_F^S}}.$$

Rearranging terms gives Eqs. (13.103) in the main text.

References

Ahn, D-H, R.F. Dittmar and A.R. Gallant (2002): "Quadratic Term Structure Models: Theory and Evidence." *Review of Financial Studies* 15, 243-288.

Aït-Sahalia, Y. (1996): "Testing Continuous-Time Models of the Spot Interest Rate." *Review of Financial Studies* 9, 385-426.

Ahn, C.-M. and H.E. Thompson (1988): "Jump-Diffusion Processes and the Term Structure of Interest Rates." *Journal of Finance* 43, 155-174.

Ang, A. and M. Piazzesi (2003): "A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables." *Journal of Monetary Economics* 50, 745-787.

Balduzzi, P., S.R. Das, S. Foresi and R. K. Sundaram (1996): "A Simple Approach to Three Factor Affine Term Structure Models." *Journal of Fixed Income* 6, 43-53.

Bartlett, B. (2006): "Hedging Under SABR Model." *Wilmott Magazine* July/August, 68-70.

Black, F. (1976): "The Pricing of Commodity Contracts." *Journal of Financial Economics* 3, 167-179.

Black, F. and M. Scholes (1973): "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81, 637-659.

Brace, A., D. Gatarek and M. Musiela (1997): "The Market Model of Interest Rate Dynamics." *Mathematical Finance* 7, 127-155.

Brigo, D. and F. Mercurio (2006): *Interest Rate Models—Theory and Practice, with Smile, Inflation and Credit*. Springer Verlag Finance (2nd Edition).

Brunnermeier, M. (2009): "Deciphering the Liquidity and Credit Crunch 2007-08." *Journal of Economic Perspectives* 23, 77-100.

Carverhill, A. (1994): "When is the Short-Rate Markovian?" *Mathematical Finance* 4, 305-312.

Cochrane, J.H. and M. Piazzesi (2005): "Bond Risk Premia." *American Economic Review* 95, 138-160.

Collin-Dufresne, P. and R. S. Goldstein (2002): "Do Bonds Span the Fixed-Income Markets? Theory and Evidence for Unspanned Stochastic Volatility." *Journal of Finance* 57, 1685-1729.

Conley, T.G., L.P. Hansen, E. G. J. Luttmer and J. A. Scheinkman (1997): "Short-Term Interest Rates as Subordinated Diffusions." *Review of Financial Studies* 10, 525-577.

Cox, J. C., J.E. Ingersoll and S. A. Ross (1979): "Duration and the Measurement of Basis Risk." *Journal of Business* 52, 51-61.

Cox, J. C., J.E. Ingersoll and S. A. Ross (1985): "A Theory of the Term Structure of Interest Rates." *Econometrica* 53, 385-407.

Dai, Q. and K.J. Singleton (2000): "Specification Analysis of Affine Term Structure Models." *Journal of Finance* 55, 1943-1978.

Diebold, F.X. and C. Li (2006): "Forecasting the Term Structure of Government Bond Yields." *Journal of Econometrics* 130, 337-364.

Duffie, D. and R. Kan (1996): "A Yield-Factor Model of Interest Rates." *Mathematical Finance* 6, 379-406.

Duffie, D. and K.J. Singleton (1999): "Modeling Term Structures of Defaultable Bonds." *Review of Financial Studies* 12, 687-720.

Estrella, A. and G. Hardouvelis (1991): "The Term Structure as a Predictor of Real Economic Activity." *Journal of Finance* 46, 555-76.

Fama, E.F. and R.R. Bliss (1987): "The Information in Long-Maturity Forward Rates." *American Economic Review* 77, 680-692.

Fong, H.G. and O.A. Vasicek (1991): "Fixed Income Volatility Management." *The Journal of Portfolio Management* (Summer), 41-46.

Geman, H. (1989): "The Importance of the Forward Neutral Probability in a Stochastic Approach to Interest Rates." Unpublished working paper, ESSEC.

Geman H., N. El Karoui and J.C. Rochet (1995): "Changes of Numéraire, Changes of Probability Measures and Pricing of Options." *Journal of Applied Probability* 32, 443-458.

Goldstein, R.S. (2000): "The Term Structure of Interest Rates as a Random Field." *Review of Financial Studies* 13, 365-384.

Hagan, P. S. and D.E. Woodward (1999): "Equivalent Black Volatilities." *Applied Mathematical Finance* 6, 147-157.

Hagan, P.S., D. Kumar, A.S. Lesniewski, and D. E. Woodward (2002): "Managing Smile Risk." *Wilmott Magazine*, September, 84-108.

Harvey, C.R. (1991): "The Term Structure and World Economic Growth." *Journal of Fixed Income* 1, 4-17.

Harvey, C.R. (1991): "The Term Structure Forecasts Economic Growth." *Financial Analysts Journal* May/June 6-8.

Heaney, W.J. and P.L. Cheng (1984): "Continuous Maturity Diversification of Default-Free Bond Portfolios and a Generalization of Efficient Diversification." *Journal of Finance* 39, 1101-1117.

Heath, D., R. Jarrow and A. Morton (1992): "Bond Pricing and the Term-Structure of Interest Rates: a New Methodology for Contingent Claim Valuation." *Econometrica* 60, 77-105.

Heston, S.L. (1993): "A Closed Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options." *Review of Financial Studies* 6, 327-344.

Ho, T.S.Y. and S.-B. Lee (1986): "Term Structure Movements and the Pricing of Interest Rate Contingent Claims." *Journal of Finance* 41, 1011-1029.

Hördahl, P., O. Tristani and D. Vestin (2006): "A Joint Econometric Model of Macroeconomic and Term Structure Dynamics." *Journal of Econometrics* 131, 405-444.

Hull, J. (2003): *Options, Futures, and Other Derivatives*. Prentice Hall. 5th edition (International Edition).

Hull, J. and A. White (1990): "Pricing Interest Rate Derivative Securities." *Review of Financial Studies* 3, 573-592.

Jamshidian, F. (1989): "An Exact Bond Option Pricing Formula." *Journal of Finance* 44, 205-209.

Jamshidian, F. (1997): "Libor and Swap Market Models and Measures." *Finance and Stochastics* 1, 293-330.

Jöreskog, K.G. (1967): "Some Contributions to Maximum Likelihood Factor Analysis." *Psychometrica* 32, 443-482.

Karlin, S. and H.M. Taylor (1981): *A Second Course in Stochastic Processes*. San Diego: Academic Press.

Kennedy, D.P. (1994): "The Term Structure of Interest Rates as a Gaussian Random Field." *Mathematical Finance* 4, 247-258.

Kennedy, D.P. (1997): "Characterizing Gaussian Models of the Term Structure of Interest Rates." *Mathematical Finance* 7, 107-118.

Kessel, R.A. (1965): "The Cyclical Behavior of the Term Structure of Interest Rates." National Bureau of Economic Research Occasional Paper No. 91.

Kloeden, P. and E. Platen (1992): *Numeric Solutions of Stochastic Differential Equations*. Berlin: Springer Verlag.

Knez, P. J., R. Litterman and J. Scheinkman (1994): "Explorations into Factors Explaining Money Market Returns." *Journal of Finance* 49, 1861-1882.

Lamberton, D. and B. Lapeyre (1997): *Introduction au Calcul Stochastique Appliqué à la Finance*. Paris: Ellipses.

Langetieg, T. (1980): "A Multivariate Model of the Term Structure of Interest Rates." *Journal of Finance* 35, 71-97.

Laurent, R.D. (1988): "An Interest Rate-Based Indicator of Monetary Policy." *Federal Reserve Bank of Chicago Economic Perspectives* 12, 3-14.

Laurent, R.D. (1989): "Testing the Spread." *Federal Reserve Bank of Chicago Economic Perspectives* 13, 22-34.

Litterman, R. and J. Scheinkman (1991): "Common Factors Affecting Bond Returns." *Journal of Fixed Income* 1, 54-61.

Litterman, R., J. Scheinkman, and L. Weiss (1991): "Volatility and the Yield Curve." *Journal of Fixed Income* 1, 49-53.

Longstaff, F.A. and E. S. Schwartz (1992): "Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model." *Journal of Finance* 47, 1259-1282.

Mele, A. (2003): "Fundamental Properties of Bond Prices in Models of the Short-Term Rate." *Review of Financial Studies* 16, 679-716.

Mele, A. and F. Fornari (2000): *Stochastic Volatility in Financial Markets: Crossing the Bridge to Continuous Time*. Boston: Kluwer Academic Publishers.

Mele, A. and O. Obayashi (2015): *The Price of Fixed Income Market Volatility*. Springer Verlag Finance (forthcoming).

Merton, R.C. (1973): "Theory of Rational Option Pricing." *Bell Journal of Economics and Management Science* 4, 141-183.

Miltersen, K., K. Sandmann and D. Sondermann (1997): "Closed Form Solutions for Term Structure Derivatives with Lognormal Interest Rate." *Journal of Finance* 52, 409-430.

Nelson, C.R. and A.F. Siegel (1987): "Parsimonious Modeling of Yield Curves." *Journal of Business* 60, 473-489.

Piazzesi, M. (2010): "Affine Term Structure Models." Chapter 12 in: Aït-Sahalia Yacine and Lars Peter Hansen (eds.): *Handbook of Financial Econometrics, Volume 1–Tools and Techniques*. North Holland; pp. 691-766.

Rebonato, R. (1998): *Interest Rate Option Models*. Wiley.

Rebonato, R. (1999): *Volatility and Correlation*. Wiley.

Ritchken, P. and L. Sankarasubramanian (1995): "Volatility Structure of Forward Rates and the Dynamics of the Term Structure." *Mathematical Finance* 5, 55-72.

Sandmann, K. and D. Sondermann (1997): "A Note on the Stability of Lognormal Interest Rate Models and the Pricing of Eurodollar Futures." *Mathematical Finance* 7, 119-125.

Santa-Clara, P. and D. Sornette (2001): "The Dynamics of the Forward Interest Rate Curve with Stochastic String Shocks." *Review of Financial Studies* 14, 149-185.

Stanton, R. (1997): "A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk." *Journal of Finance* 52, 1973-2002.

Stock, J.H. and M. W. Watson (1989): "New Indexes of Coincident and Leading Economic Indicators." In: Blanchard, O. J. and S. Fischer (Eds.): *NBER Macroeconomics Annual 1989*, MIT Press, 352-394.

Stock, J. H. and M. W. Watson (2003): "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature* 41, 788-829.

Vasicek, O. (1977): "An Equilibrium Characterization of the Term Structure." *Journal of Financial Economics* 5, 177-188.

Veronesi, P. (2010): *Fixed Income Securities: Valuation, Risk and Risk Management*. John Wiley and Sons.

14
Risky debt and credit derivatives

14.1 Introduction

This chapter deals with the pricing of securities that carry credit risk. It examines the main conceptual approaches to deal with credit risk as well as how this risk can be transferred through dedicated *credit derivatives*. It is instructive to review the historical reasons leading up to the creation and trading of these derivatives. The next subsection contains such a succinct account; Section 14.1.2 provides a roadmap to this chapter.

[In progress]

14.1.1 A brief history of credit risk and financial innovation

During the mid 1980s, a market begins to develop regarding the first interest rates derivatives reviewed in the previous chapter. This market would grow to an extent that during the late 1980s already, the appetite for these derivatives would proliferate and also lead to additional and fairly complex products, arising through innovation and competition. It is natural: financial innovation is relatively easy to imitate, which leads banks to increasing creativeness—increasing creativeness is needed to keep the innovators' initial competitive advantage as long as possible.

The early 1990s were extraordinary years. On the one hand, interest rates were low amid concerns the U.S. economy had still not recovered from the 1991 recession. On the other hand, capital market volatility was quite muted. Low interest rates and low capital market volatility are natural drivers for motivating the introduction of new derivatives aimed to boosting investors' returns.[1] But the financial turmoil in 1994 brought the interest rate climate to suddenly change while some of these products were producing large losses. These losses would trigger a call for regulation by public opinion and certain policy makers even while the ISDA (International Swaps and Derivatives Association) would debate that more regulation would destroy market creativity.

Regulatory pressures would vanish by the mid 1990s, when the market started to innovate again amid a general consensus that derivative risks could be controlled through market discipline, not regulation. Swap markets recovered. They did so slowly though, as these derivatives

[1] Examples of products introduced at the time are LIBOR squared, inverse floaters, or power options, sponsored by JPMorgan.

were already in the end of the innovation cycle: the process of imitation had led swap-related derivatives to become a "mass product," with profit margins having been eroded in the meantime. The market was ready for a new major innovation wave.

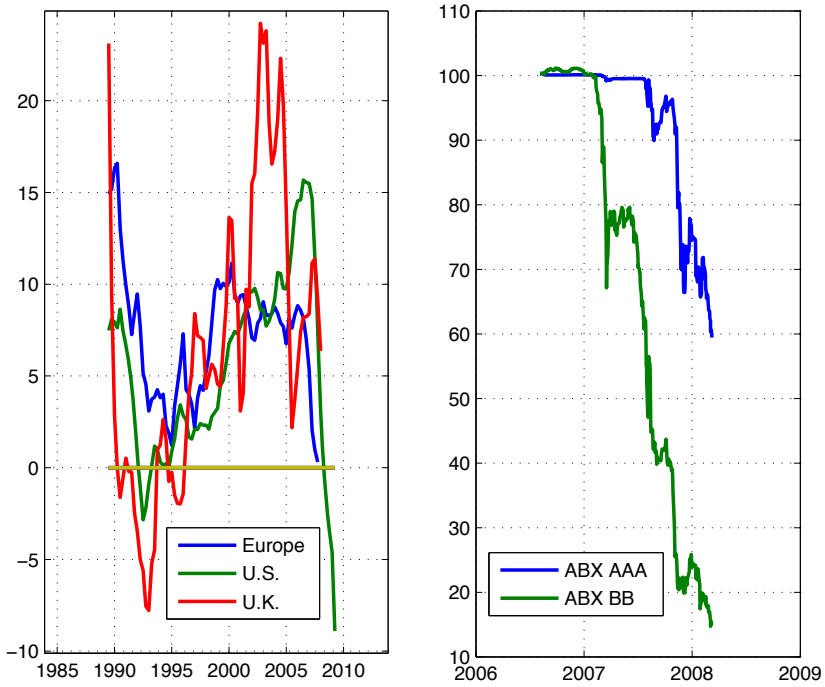
Credit risk was the next innovation stream. Global institutions (e.g., JPMorgan, Credit Suisse, Bankers Trust) soon realized that borrower defaulting was a source of substantial risk that could be so conveniently re-allocated through dedicated derivatives. Similarly as classic derivatives transfer market risk, credit derivatives could transfer market risk. Institutions such as JPMorgan had additional motivations to innovate in this space, given the vast pools of loans contained in its books: importantly, these loans required too many reserves and were therefore expensive.

One solution was to proceed with *securitization*. Securitization is a process by which some illiquid assets (say some loans) are gathered ("packaged") into a common pool that backs the issuance of new securities aimed to display an enhanced liquidity obtained through packaging, credit and liquidity enhancements. These new securities are, in fact, derivatives written on the initial illiquid assets. Two leading examples of this process include the securitization of mortgages and receivables. Financial institutions find the securitization process attractive, as they can carve out certain items in their balance sheet, thus boosting their return on investments; moreover, by securitizing assets, less capital is needed to meet capital requirements standards. For example, the accounts receivables of a corporation may be used to back the issue of commercial paper known as *asset-backed commercial paper*. A well-functioning securitization system is a way (not the only way) to transfer and trade credit risk.

Global institutions would then repackage loans into derivatives, in a way that default risk and/or part of the securitized loans could be transferred to outside investors. Note that credit derivatives were also a regulatory mitigation device, partly useful as a response to regulation. The underlying ideas were (i) to turn loans into derivatives that could be sold, and (ii) to create new insurance products such as *credit default swaps*. At the very beginning, derivatives were just designed to have single loans as the underlying. Afterwards, the idea emerged to create structures organized in derivatives bundles, with cash flow indexed to baskets of loans—the ancestors to *collateralized debt obligations* ("CDOs"). For example, JPMorgan created "Bistro" (Broad Index Secured Trust Offering), a structure relying on a variety of assets, ranging from corporate debt to student loans; ABN-Amro created similar structures ("Heineken" and "Amstel"). During this innovation process, competition increased and profit margins fell again, leading to renewed motivation for additional innovation.

Year	US ABS (Outstanding)	Global CDO (Issuance)	US Agency MBS (Issuance)	US Agency CMO (Issuance)
2000	1084	68	474	586
2001	1230	78	1086	1455
2002	1381	83	1447	2019
2003	1507	87	2131	2762
2004	1814	157	1015	1379
2005	2111	251	983	1345
2006	2700	455	923	1240
2007	2945	430	1189	1471
2008	2599	62	1169	1339
2009	2326	4	1734	2022
2010	2034	7	1420	1885

In billions US dollars. Source: Sifma.



Left hand side panel: U.S. and European House Price Change (year-to-year, in percentage). Right hand side panel: Indexex of CDS on U.S. Mortgage-Related Securities Prices. Source: IMF, Global Financial Stability Report, April 2008.

The response to increased competition was the creation of structured products referenced to riskier assets. Importantly, during the mid 1990s, derivatives teams begun to interact with teams managing loans extended to borrowers with poor credit history—"subprime mortgages." Subprime loans would begin to be securitized and structured into CDOs with global financial institutions involved (e.g., Merrill Lynch or UBS). The mortgage banking model had shifted from one of "buy to hold" to one of "originate and distribute." The subprime crisis erupted in 2007, and sent the global financial system into nearly two years of turmoil, in a situation where the payments and settlement system was jeopardized.

The mechanics of the 2007-2008 crisis are well-understood. While low interest rates helped sustain the boom in the housing market, the originate-and-distribute model had operated in a way that lending standards were not in line with market expectations (see Section 14.6.2). The subprime mortgage market would sink amid increasing interest rates and rapidly decelerating housing prices. The "shadow banking system," which helped sustain the originate-and-distribute model, was actually an important piece of the crisis: the uncertainties related to the entities financing this system triggered a sharp liquidity dry-up and, then, a credit crunch, followed by a drop in the real economic activity, which magnified the credit crunch, over a spiral.

[In progress]

14.1.2 Plan of the chapter

This chapter reviews conceptual approaches to the pricing of defaultable securities as well as the basic mechanics of derivatives that transfer credit risk. Some of these approaches rely on models that predict classical irrelevance results (Modigliani and Miller, 1958; see Chapter 2): the firm capital structure does not matter for the value of a firm. This result should sound like a remainder for some of the subsequent developments in this chapter. For example, Section 14.2 deals with "structural approaches" to debt evaluation, relying on assumptions that are at times consistent and at other times inconsistent with those underlying classical irrelevance results. Section 14.2 also deals with "reduced-form approaches" by which default risk is modeled as an exogenously given event. Section 14.4 reviews the main credit derivatives that aim to re-allocate credit risk, such as credit default swaps or securitized obligations.

These lectures have never really dealt with issues regarding risk-management. Section 14.4 contains both introductory discussions regarding risk-management in general and some details of credit risk management. It also discusses regulatory developments over the relatively recent history. Section 14.5 discusses a few more details regarding the 2007-08 financial crisis, which originally erupted when losses mounted in credit markets, and then spread to the overal economy. It is an exemplary "case study" (very well utilized in the literature) that illustrates how a shock in capital markets can affect global developments over a vicious circle: an important instance of endogenous risk discussed in quite a few junctures of Part II (see, e.g., Chapters 9 and 11).

14.2 Conceptual approaches to valuation of defaultable securities

14.2.1 Firm value, or structural, approaches

Relies on the *structure* of the firm. Shares and bonds as derivatives written upon the firm asset value. We begin reviewing the Merton's model, which stems from a Modigliani-Miller world, where, as we know, the value of the firm is not affected by the leverage (see Chapter 2). Note

since the beginning the limitation of such a model: it's a model with which we evaluate debt, assuming a world where leverage does not even affect the value of the firm that is issuing the debt!

Consider the following stylized balance sheet.

Assets (A)	Equity (E) (Shares)
	Debt (D) (Bonds)

Therefore, we have the accounting identity: Assets = Equity + Debt, or

$$A = E + D.$$

When debt expires, debt-holders receive the minimum between the nominal value of debt and the value of the assets the firm can liquidate to honour debt. Debt-holders are senior claimants. Equity holders are juniors, i.e., they are residual claimants to the firm's assets.

We use these basic insights and illustrate the first approach to the modeling of the risk-structure of interest rates—the Merton-KMV approach. In this approach, equity is the same as a European call option written on the firm's assets, with expiration equal to the debt expiration, and strike equal to the nominal value of debt. The current value of debt equals the value of the assets minus the value of equity, i.e. the value of a risk-free discount bond minus the value of a put option on the firm with strike price equal to the nominal value of debt, as shown by Eq. (14.3) below.

Merton (1974) uses the Black and Scholes (1973) formula to derive the price of debt. The main assumption underlying this model is that the assets of the firm can be traded, and that their value A_t satisfies[2]

$$\frac{dA_t}{A_t} = rdt + \sigma d\tilde{W}_t, \tag{14.1}$$

where $\tilde{W}_t$ is a Brownian motion under the risk-neutral probability, σ is the instantaneous standard deviation, and r is the short-term rate on riskless bonds.

Let N be the nominal value of debt, T be time of expiration of debt; D_t the debt value as of at time $t \leq T$. As argued earlier, shareholders are long a European call option, and the bond-holders are residual claimants. Mathematically,

$$D_T = \begin{cases} A_T, & \text{if the firm defaults, i.e. } A_T < N \\ N, & \text{if the firm is solvent, i.e. } A_T \geq N \end{cases}$$

We can decompose the firm asset value at time T, into the sum of the value of equity and the value of debt at T:

$$D_T = \min\{A_T, N\} = A_T \; - \underbrace{\max\{A_T - N, 0\}}_{\equiv \text{ Equity at } T}. \tag{14.2}$$

Note, also, that,

$$D_T = \min\{A_T, N\} = N - \underbrace{\max\{N - A_T, 0\}}_{\equiv \text{ Put on the firm}}. \tag{14.3}$$

[2]Eq. (14.1) could be generalized to one in which $dA_t = (rA_t - \delta_t)\,dt + \sigma A_t d\tilde{W}_t$, where δ_t is the instantaneous cash flow to the firm. This would make the firm value equal to $A_0 = \mathbb{E}\left(\int_0^\infty e^{-rt}\delta_t dt\right)$. For example, one could take δ_t to be a geometric Brownian motion with parameters g and σ, in which case $A_t = (r-g)^{-1}\,\delta_t$, forever, but we're just ignoring this complication.

That is, credit risk raises the cost of capital.

Note that convexity gives equityholders incentives to undertake a risk-taking behavior: the terminal payoff to them is increasing in the asset volatility, σ. This property has been emphasized to explain adverse selection and credit rationing in Chapter 9. Section 14.5.5 also relies on this basic property and explains how availability of credit and limited liability may lead to bubbles. Instead, the terminal payoff to debtholders is concave, such that debt value is, then, decreasing in assets volatility, a property discussed further in the next section.

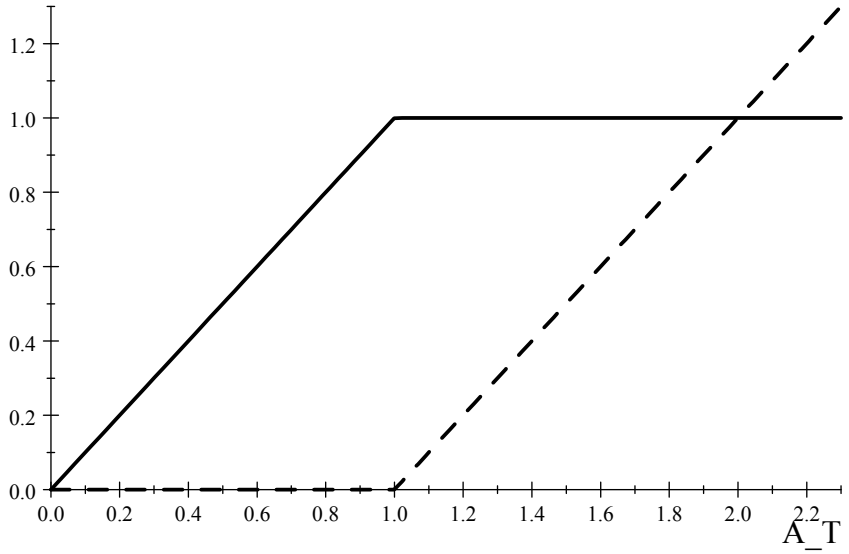


FIGURE 14.1. Dashed line: the value of equity at the debt maturity, T, $\max\{A_T - N, 0\}$, plotted as a function of the firm asset value, A_T. Solid line: the value of debt at maturity, $\min\{A_T, N\}$ as a function of A_T. Nominal value of debt is fixed to $N = 1$.

14.2.1.1 Merton

The *current* value of the bonds equals the current value of the assets, A_0, minus the current value of equity. The *current* value of equity can obtained through the Black & Scholes formula, as equity is a European call option on the firm, struck at N. By Eq. (14.2), and standard risk-neutral evaluation, the *current* value of debt, D_0, is,

$$D_0 = A_0\Phi\left(-d_1\right) + Ne^{-rT}\Phi(d_1 - \sigma\sqrt{T}), \quad d_1 = \frac{\ln\left(A_0/N\right) + \left(r + \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}, \tag{14.4}$$

where $\Phi\left(\cdot\right)$ denotes the distribution function of a standard normal variable.[3]

[3]For the details, note that $D_0 = e^{-rT}\mathbb{E}\left[D_T|\,A_0\right]$ and, then, by Eq. (14.2),

$$D_0 = e^{-rT}\mathbb{E}\left(A_T|\,A_0\right) - e^{-rT}\mathbb{E}\left[\max\{A_T - N, 0\}|\,A_0\right] = A_0 - \left[A_0\Phi\left(d_1\right) - Ne^{-r\tau}\Phi(d_1 - \sigma\sqrt{T})\right],$$

where the last equality follows by the Black & Scholes formula. Eq. (14.4) follows after rearranging terms in the previous equation.

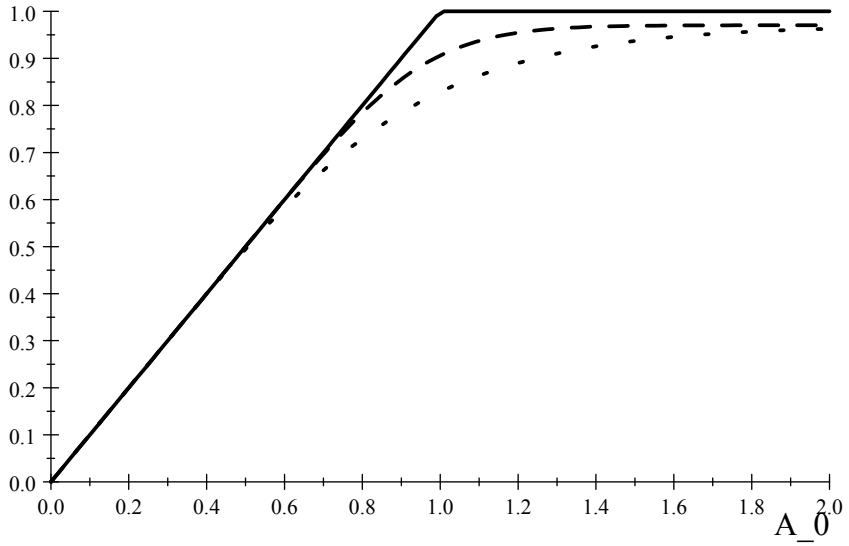


FIGURE 14.2. Solid line: the no-arbitrage bound, $\min\{A_0, N\}$, depicted as a function of A_0, when the nominal value of debt is fixed to $N = 1$. Dashed line: the bond value predicted by the Merton's model when $T = 1$, $r = 3\%$ and $\sigma = 20\%$, annualized. Dotted line: same as the dashed line, but with a higher asset volatility, $\sigma = 40\%$.

Bond prices are decreasing in the asset volatility as bad outcomes are exaggerated on the downside, due to the concavity properties depicted in Figure 14.1. Note that the property that the term structure of interest rates increases with the volatility of the fundamentals is quite sharp here. Chapter 13 (see Section 13.3.4) explains that the relation between the yield curve and the volatility of the fundamentals (i.e., the short-term rate) is complex, relying as it does on two opposing effects—a "convexity" and a "risk-premium" effect. In bad times, it should be the risk-premium effect to dominate, thereby leading to a positive link between the volatility of the fundamentals and the yield curve. In good times, a convexity effect would lead the yield curve to be negatively related to the volatility of the fundamentals. Instead, the prediction in this section is quite neat: the term structure of interest rates always increases with the volatility of the fundamentals. Naturally, this prediction relies on a channel that is completely distinct from the risk-premium channel discussed in Chapter 13.

The term structure of interest rates is defined as usual as:

$$R \equiv -\frac{1}{T}\ln\left(\frac{D_0}{N}\right) = r + \mathrm{s}_0,$$

where

$$\mathrm{s}_0 = -\frac{1}{T}\ln\left(\frac{A_0}{N}e^{rT}\Phi\left(-d_1\right) + \Phi\left(d_1 - \sigma\sqrt{T}\right)\right). \tag{14.5}$$

We usually refer to s_0 as the "term-spread" for a given fixed maturity, and to the mapping maturity-spreads as the "risk-structure of interest rates."

Figure 14.3 depicts the spread predicted by this model. Credit spreads shrink to zero as time-to-maturity becomes smaller and smaller. This property of the model stands in sharp contrast with the empirical behavior of credit spreads, which are high even for short-maturity bonds. This property arises because the model is driven by Brownian motions, which have

have continuous sample paths, such that given a firm asset value $A > N$, the probability of bankruptcy, arising when A hits N from above, approaches zero very fast as time-to-maturity goes to zero. Because credit spreads reflect default probabilities, as explained in detail below (see Eq. (14.9)), credit spreads shrink to zero quickly as time-to-maturity approaches zero.

Naturally, one might end up with credit spreads sufficiently high at short maturities, by assuming the firm asset value is sufficiently small. For example, in Figure 14.3.1, credit spreads are "high" at short maturities, when $A = 1.1$. However, even with $A = 1.1$, credit spreads are still zero at very short maturities. More fundamentally, requiring such a small value for A is problematic. Firms with such a low asset value would command a much higher spread than that in Figure 14.3.1. All in all, the Brownian motion model in this section lacks some source of risk driving the behavior of short-term spreads. In Section 14.2.2, we will show that this issue can be addressed assuming that firm's default can be triggered by "jumps."

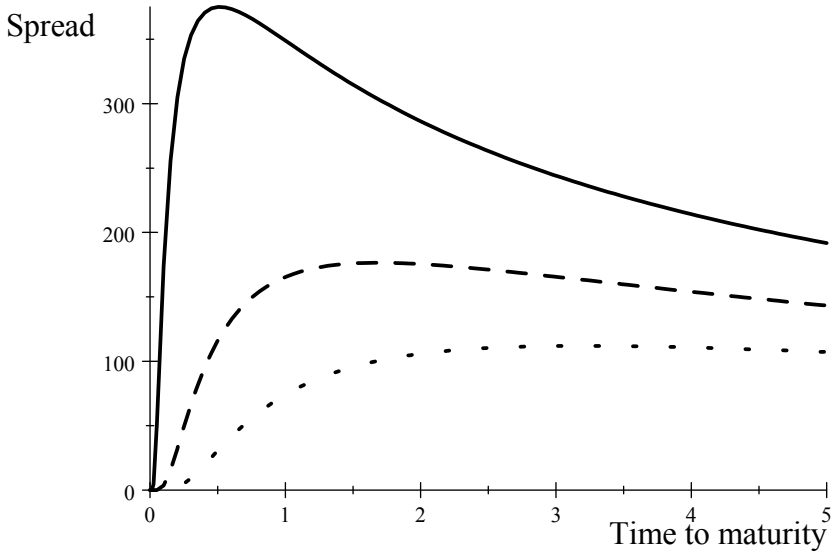


FIGURE 14.3.1. The term structure of spreads, s_0, in basis points, predicted by Merton's model, obtained with initial asset values $A_0 = 1.1$ (solid line), $A_0 = 1.2$ (dashed line), and $A_0 = 1.3$ (dotted line). The short-term rate, $r = 3\%$, and asset volatility is $\sigma = 0.20$. Nominal debt $N = 1$.

Naturally, the term-structure of credit spreads has a rather different shape, when the current firm asset value is below N, as depicted in Figure 14.3.2. In this case, the probability the firm defaults is close to one when time to maturity is close to zero, such that the spreads would then be arbitrarily large as we get closer and closer to maturity. For visualization purposes, Figure 14.3.2 is truncated to only include values of the spreads for maturities larger than one year.

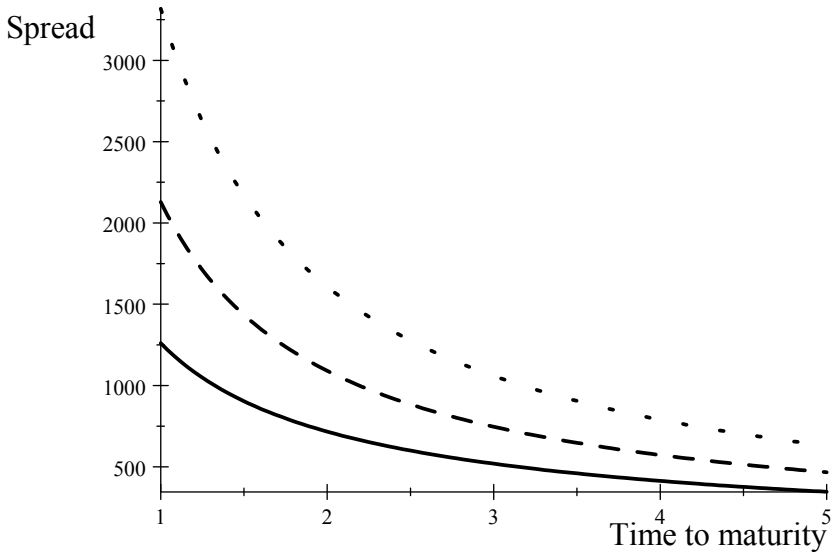


FIGURE 14.3.2. The term structure of spreads, s_0, in basis points, predicted by Merton's model, obtained with initial asset values $A_0 = 0.9$ (solid line), $A_0 = 0.8$ (dashed line), and $A_0 = 0.7$ (dotted line). The short-term rate, $r = 3\%$, and asset volatility is $\sigma = 0.20$. Nominal debt $N = 1$.

What is the asymptotic behavior of the spread predicted by this model? If $r > \frac{1}{2}\sigma^2$, then, as $T \to \infty$, the probability of survival for the firm, which we shall show to be $\Phi\left(d_1 - \sigma\sqrt{T}\right)$ below (see Eq. (14.7)), approaches one, $\Phi\left(d_1 - \sigma\sqrt{T}\right) \to 1$. That is, the firm asset value is expected to grow so large that default will never occur, such that the bond becomes riskless and $s_0 \to -\frac{1}{T}\ln\Phi\left(d_1 - \sigma\sqrt{T}\right) \to 0$. Intuitively, when $r > \frac{1}{2}\sigma^2$, the asset volatility is so small, that the exponential trend for A_t will make it unlikely that the firm asset value will fall below the constant value N. In other words, the Merton's model predicts that in the long-run, things can only go well for the firm, a view quite opposite to that leading to positive spreads for long maturities.

To summarize, the Merton's model predict that *short-term* spreads are zero and *long-term* spreads are likely so. Intensity models, such as those analyzed in Section 14.2.3, help mitigate these counterfactual features. Intuitively, an intensity model is one where a firm can default at any time, thereby leading to positive short-term spreads. If the probability of default is time-varying and mean-reverting, long-term spreads might the be positive, especially when the current default probability is so low to be expected to increase to high levels, due to mean-reversion.

14.2.1.2 Assessing credit quality

We introduce a useful summary statistics to measure credity quality, called *distance-to-default* (under $\mathbb{Q}$). We can use the previous model to estimate the likelihood of default for a given firm. First, we develop Eq. (14.2),

$$D_T = \min\{A_T, N\} = A_T \cdot \mathbb{I}_{\{A_T < N\}} + N \cdot \mathbb{I}_{\{A_T \geq N\}},$$

where $\mathbb{I}_{\{\mathcal{E}\}}$ is the indicator function, i.e. $\mathbb{I}_{\{\mathcal{E}\}} = 1$ if the event $\mathcal{E}$ is true and $\mathbb{I}_{\{\mathcal{E}\}} = 0$ if the event $\mathcal{E}$ is false. Second, we have,

$$\begin{aligned} D_0 &= e^{-rT}\mathbb{E}\left(D_T\right) \\ &= e^{-rT}\left[\mathbb{E}\left(A_T \cdot \mathbb{I}_{\{A_T<N\}}\right) + N \cdot \mathbb{E}\left(\mathbb{I}_{\{A_T\geq N\}}\right)\right] \\ &= e^{-rT}\left[\mathbb{E}\left(A_T|\,\text{Default}\right)\mathbb{Q}\left(\text{Default}\right) + N \cdot \mathbb{Q}\left(\text{Survival}\right)\right], \end{aligned} \tag{14.6}$$

where $\mathbb{E}\left(A_T|\,\text{Default}\right)$ is the expected firm asset value given the event of default, $\mathbb{Q}\left(\text{Default}\right)$ is the probability of default, and $\mathbb{Q}\left(\text{Survival}\right) = 1 - \mathbb{Q}\left(\text{Default}\right)$ is the probability the firms does not default.[4]

Comparing Eq. (14.6) and Eq. (14.4) reveals that for the Merton's model,

$$\mathbb{Q}\left(\text{Survival}\right) = \Phi\left(d_2\right), \tag{14.7}$$

where for obvious reasons, the quantity,

$$d_2 = \frac{\ln\left(A_0/\,N\right) + \left(r - \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}, \tag{14.8}$$

is called *distance-to-default*. Distance to default is a useful summary statistics, providing an intutive gauge of how far the firm is from defaulting: the higher d_2, the higher the (risk-adjusted) probability of surviving, $\Phi\left(d_2\right)$. The higher the current firm asset value A_0 is, the less likely it is the firm will default at T.

By Eq. (14.1), we have that $\mathbb{E}\left(\ln A_T|\,A_0\right) = \ln A_0 + (r - \frac{1}{2}\sigma^2)T$, so Eq. (14.8) tells us that distance-to-default is simply the difference $\mathbb{E}\left(\ln A_T|\,A_0\right) - \ln N$, normalized by the standard deviation of the assets over the life of debt. Some, then, might prefer to use the slightly different formula,

$$\text{Distance-to-default} = \frac{\text{asset mkt value } - \text{default value}}{\text{asset mkt value} \times \text{asset volatility}}.$$

How does the probability of survival for a given firm relate to debt maturity or asset volatility? In Figure 14.4, the probability of survival decreases with, (i) debt maturity and (ii) asset volatility.

[4]The last equality in Eq. (14.6) follows because $\mathbb{E}\left(A_T\mathbb{I}_{\{A_T<N\}}\right) = \mathbb{E}\left(A_T\mathbb{I}_{\{A_T<N\}}|\,\mathbb{I}_{\{A_T<N\}} = 1\right)\mathbb{Q}\left(\mathbb{I}_{\{A_T<N\}} = 1\right) + \mathbb{E}\left(A_T\mathbb{I}_{\{A_T<N\}}|\,\mathbb{I}_{\{A_T<N\}} = 0\right)\mathbb{Q}\left(\mathbb{I}_{\{A_T<N\}} = 0\right) = \mathbb{E}\left(A_T|\,A_T < N\right)\mathbb{Q}\left(A_T < N\right)$.

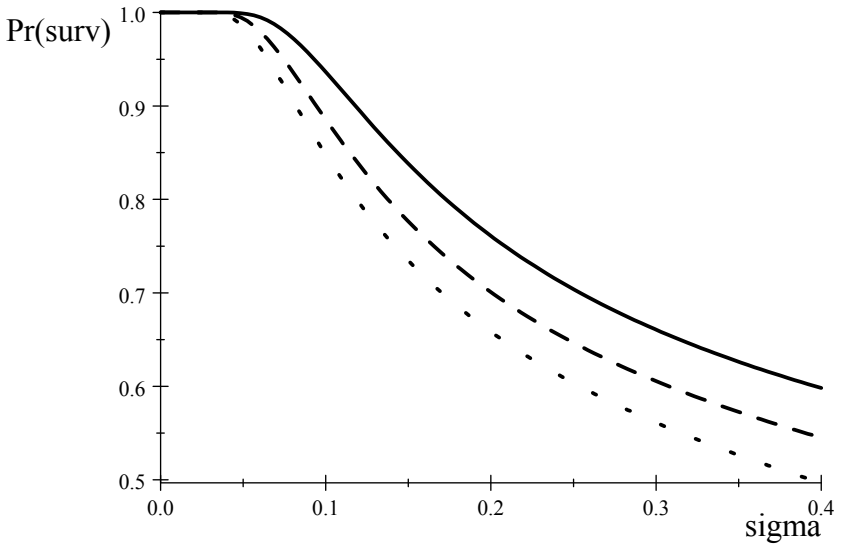
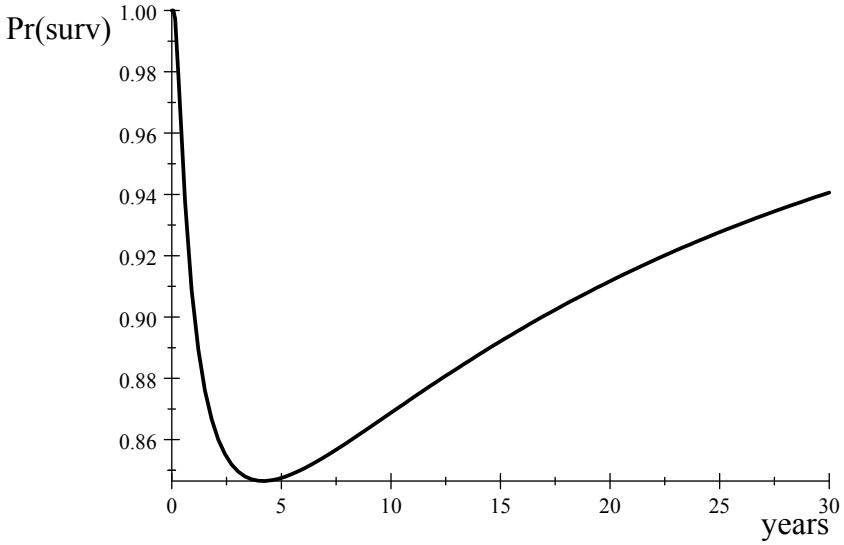


FIGURE 14.4. Probability of survival for a given firm predicted by the Merton's model, $\Phi(d_2)$, depicted as a function of the asset volatility, σ. Firm asset value is fixed at $A_0 = 1.1$, and plotted are survival probabilities for bonds maturing at $T = 0.5$ years (solid line), $T = 1$ year (dashed line) and $T = 2$ years (dotted line). The short-term rate, $r = 3\%$. Nominal debt $N = 1$.

Property (i) is not a general property, though. For example, we already pointed out that for large T, the probability of survival is close to one as soon $r > \frac{1}{2}\sigma^2$, a condition ensuring the firm asset value grows so large to ensure default becomes unlikely, eventually. The next picture shows that for a fixed σ, such that $r > \frac{1}{2}\sigma^2$, the probability of survival is non-monotonic in T.



Probability of survival for a given firm predicted by the Merton's model, $\Phi(d_2)$, depicted as a function of time-to-maturity, when the firm asset value is fixed at $A_0 = 1.1$, and asset volatility is $\sigma = 0.10$. The short-term rate, $r = 3\%$. Nominal debt $N = 1$.

This property arises for the following reason. Assuming that $A_0 > N$ and $r > \frac{1}{2}\sigma^2$, the first term of d_2 in Eq. (14.8) is decreasing in T, whereas the second is increasing. When T is small, the first term (and its sensitivity to T dominates, such that distance-to-default decreases with maturity. But for T large, the second term of d_2 dominates, and distance-to-default becomes eventually large. Non-monotonicities arise even at finite maturities, once we consider low values of A_0, in which case the relation between maturity and probability of survival can be increasing or decreasing, according to the values of σ, as shown in Figure 14.5. Intuitively, when $A_0 \approx N$, the probability of survival is:

$$\mathbb{Q}\left(\text{Survival}\right) = \Phi\left(d_2\right), \text{ with } d_2 = \frac{\ln\left(\frac{A_0}{N}\right) + \left(r - \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}} \approx \frac{\left(r - \frac{1}{2}\sigma^2\right)}{\sigma}\sqrt{T},$$

such that the survival probability decreases in T for large σ although then it increases in T for small σ. The intuition underlying this property is that for large σ, the probability the firm asset value will end up below N from $A_0 \approx N$ can only increase with time to maturity, T. Analytically, $\mathbb{E}\left(\ln A_T | A_0\right) = \ln A_0 + \left(r - \frac{1}{2}\sigma^2\right)T \approx \ln N + \left(r - \frac{1}{2}\sigma^2\right)T$, such that the probability the firm asset value will be below N does indeed increase with σ.

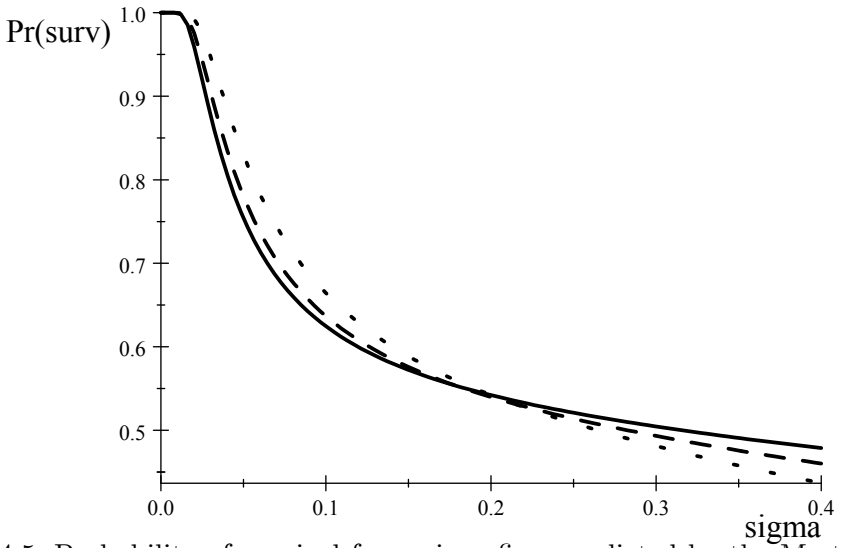


FIGURE 14.5. Probability of survival for a given firm predicted by the Merton's model, $\Phi\left(d_2\right)$, depicted as a function of the asset volatility, σ. The firm asset value is fixed at $A_0 = 1.01$, and plotted are survival probabilities for bonds maturing at $T = 0.5$ years (solid line), $T = 1$ year (dashed line) and $T = 2$ years (dotted line). The short-term rate, $r = 3\%$. Nominal debt $N = 1$.

A final useful concept is that of loss-given-default (under $\mathbb{Q}$), denoted as LGD in the sequel. Comparing Eq. (14.6) with Eq. (14.4) reveals another property of the Merton's model,

$$\mathbb{E}\left(A_T | \text{Default}\right) = \frac{A_0 e^{rT}\Phi\left(-d_1\right)}{\mathbb{Q}\left(\text{Default}\right)} = A_0 e^{rT}\frac{\Phi\left(-d_1\right)}{\Phi\left(-d_2\right)} = \mathbb{E}\left(A_T\right)\frac{\Phi\left(-d_1\right)}{\Phi\left(-d_2\right)} \leq \mathbb{E}\left(A_T\right).$$

Recovery rates are defined as the fraction of the bond value the bond-holders expect to obtain at maturity and in the event of default:

$$\text{Rec} \equiv \frac{\mathbb{E}\left(A_T | \text{Default}\right)}{N} = \frac{A_0}{N}e^{rT}\frac{\Phi\left(-d_1\right)}{\Phi\left(-d_2\right)}.$$

Loss-given-default is defined as the fraction of the bond value the bond-holders expect to lose at maturity and in the event of default, i.e., LGD = 1 − Rec. Finally, by Eq. (14.5), we can write,

$$\begin{aligned} \mathrm{s}_0 &= -\frac{1}{T}\ln\left(\frac{A_0}{N}e^{rT}\Phi\left(-d_1\right)+\Phi\left(d_2\right)\right) \\ &= -\frac{1}{T}\ln\left(\mathrm{Rec}\cdot\mathbb{Q}\left(\mathrm{Default}\right)+\mathbb{Q}\left(\mathrm{Survival}\right)\right) \\ &\approx \frac{1}{T}\left[\mathrm{LGD}\cdot\mathbb{Q}\left(\mathrm{Default}\right)\right]. \end{aligned} \tag{14.9}$$

This is actually a general formula, which goes beyond the Merton's model. It can easily be obtained through Eq. (14.6).

An important note. Previously, we defined survival probabilities, distance-to-default, and loss-given-default, under the risk-adjusted probability $\mathbb{Q}$. To calculate the same objects under the true probability $\mathbb{P}$, we replace r with the asset growth rate under the physycal probability, μ, in the formulae for the survival probabilities, $\Phi\left(d_2\right)$, distance-to-default, d_2, and loss-given-default.

However, it is hard to estimate μ for many single names. Moody's KMV EDF$^{\mathrm{TM}}$ are based on dynamic structural models like these, although the details are not publicly known. Finally, we could use historical data about *default frequencies* to estimate the probability that a given single name within a certain industry will default. These frequencies are based on samples of firms that have defaulted in the past, with similar characteristics to those of the firm under evaluation (in terms, for instance, of distance-to-default).

How to estimate A_t and σ? One algorithm is to start with some σ equal to the volatility of equity returns, say $\sigma^{(0)}$, and use Merton's formula for equity, to extract $A_t^{(0)}$ for each date $t\in\{1,\cdots,\mathcal{T}\}$, where $\mathcal{T}$ is the sample size. Then, use $A_t^{(0)}$ to compute the standard deviation of $\ln(A_t^{(0)}/A_{t-1}^{(0)})$. This gives say $\sigma^{(1)}$, which can be used as the new input into the Merton's formula to extract say $A_t^{(1)}$. We obtain a sequence of $(A_t^{(i)})_{t=1}^{\mathcal{T}}$ and $\sigma^{(i)}$, and we stop for i sufficiently large, according to some criterion.

14.2.1.3 One example

Assume a firm has asset value $A_0 = 110$, and that the asset value volatility is $\sigma = 30\%$, annualized. The safe interest rate is $r = 2\%$, annualized, and the expected growth rate of the asset value is $\mu = 5\%$, annualized. The firm has outstanding debt with nominal value $N = 100$, which expires in two years.

First, we compute the distance-to-default implied by the Merton's model, which is,

$$\text{D-t-D} = \frac{\ln\left(\frac{A_0}{N}\right)+\left(r-\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}} = \frac{\ln\left(1.1\right)+\left(0.02-\frac{1}{2}0.3^2\right)2}{0.3\sqrt{2}} = 0.10680.$$

Accordingly, the probability of default is

$$1-\Phi\left(0.10680\right) = 1-0.54253 = 0.45747.$$

We compute the same probability under the physical probability, by simply replacing $r = 2\%$ with $\mu = 5\%$, in the formula for D-t-D, obtaining

$$\text{D-t-D}_{\text{physical}} = \frac{\ln\left(\frac{A_0}{N}\right)+\left(\mu-\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}} = \frac{\ln\left(1.1\right)+\left(0.05-\frac{1}{2}0.3^2\right)2}{0.3\sqrt{2}} = 0.24822.$$

Therefore, the probability of default under the physical distribution is,

$$1 - \Phi_{\text{physical}}\left(0.24822\right) = 1 - 0.59802 = 0.40198.$$

It is, of course, lower under the physical probability than under the risk-neutral probability, due to the larger asset growth rate, $\mu > r$.

Finally, we can compute the spread on this bond, which is given by:

$$\text{Spread} = -\frac{1}{T}\ln\left(\frac{A_0}{N}e^{rT}\Phi\left(-d_1\right) + \Phi\left(d_2\right)\right),$$

where $d_2 =$ D-t-D, and $d_1 = d_2 + \sigma\sqrt{T}$. So we have,

$$\begin{aligned}\text{Spread} &= -\frac{1}{2}\ln\left(1.1e^{0.02\times 2}\Phi\left(-\left(0.10680 + 0.30\times\sqrt{2}\right)\right) + \Phi\left(0.10680\right)\right)\\ &= -\frac{1}{2}\ln\left(1.1e^{0.02\times 2}\times 0.29769 + 0.54253\right)\\ &= 6.20\%.\end{aligned}$$

14.2.1.4 Stocks and bonds

On the 5th of August 2011, the rating agency Standard & Poor's downgraded the US debt from AAA to AA+ for the first time in history. The US and global equity markets sunk (the DJIA lost nearly 6%) on the first trading day (Monday the 8th) following the announcement. Somehow paradoxically, US Treasuries rallied on the very same day, a phenomenon many commentators described as a "flight-to-quality" response to a quite unique event. The reason this rally seems paradoxical is that the downgrade regarded, obviously, US debt! Moreover, at the time, the US debt/GDP ratio was hovering at about 100%, a fact that mitigates the case of US Treasuries as "safe-heaven" assets.

But additional arguments made US Treasuries safe-heaven. First, some of the signals leading to the downgrade regarded the political gaming about an increase in the debt ceiling, a gaming that could have led to delinquencies.[5] However, this gaming and its effects were arguably transitory. Moreover, AA+ debt is accepted as a collateral in many transactions, and considered to be high grade debt in most investment mandates. Furthermore, an issue at the time was to speculate whether other rating agencies (Moody's and Fitch) would proceed with similar downgrades of US debt that could have made the 05-08 decision more solidly grounded, so to speak. Finally, the Standard & Poor's decision was not totally unexpected, as rumours about it would start circulate months earlier indeed.

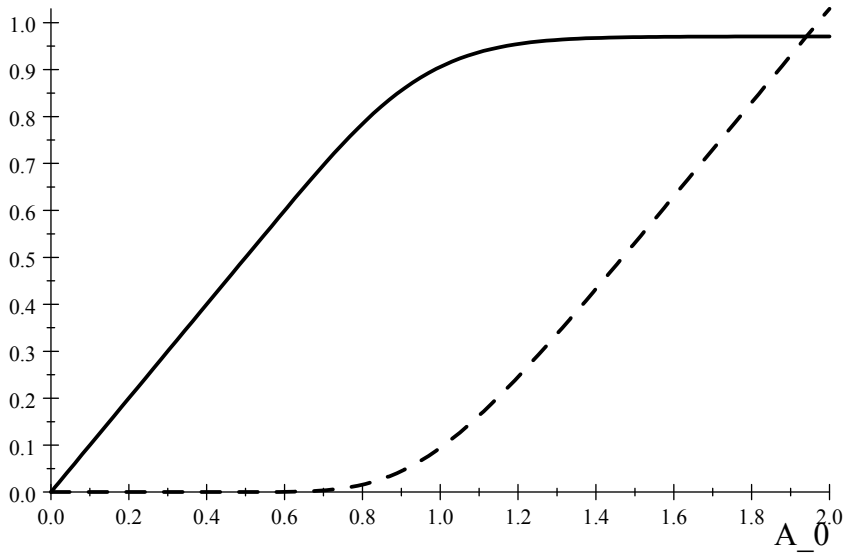
However, flight-to-quality does not seem to be an exhaustive explanation for the US Treasury rally during these events. First, during the period around the 5th of August, the US were flooded with bad news regarding the economic fundamentals, with many leading indicators reaching levels historically consistent with recessions. These news would make it likely (at the time) that the US economy would spiral towards a second recession in less than four years, i.e., right after that related to the subprime crisis (discussed in Section 14.5). The unfolding of these events is entirely consistent with we learnt in Chapter 13: recession fears translate into an expectation future short rates will lower as a result of the FED attempt to stimulate growth, whence the Treasury rally.

[5] In the US, the Congress has to approve an increase in debt capacity.

Therefore, the rally of US Treasuries at the time might merely reflect the expectation hypothesis that future rates would lower. In this period of bad news (which also included adverse developments regarding the debt crisis in Europe), the Standard & Poor's downgrade might have come as a yet-another negative signal about the general health of the US system, which would further deteriorate the general investment climate. Therefore, a rally in US Treasuries could be explained by both flight-to-quality effects (i.e., an uncertainty premium about the occurrence of possible disorderly tail events) *and* the market expectations about the FED response to particularly severe economic developments.

The model in this section suggests one additional potential channel conducive to the equity crash and the rally in US debt after the Standard & Poor's downgrade decision. What happens to the price of a stock and a bond, after an adverse shock hits the fundamentals? According to Merton's model, they both fall after a decrease in A. But which of the two prices will drop more? After all, debt is less risky than stocks (due to subordination), and bad news about the fundamentals should affect stocks more than bonds.

The next picture depicts the price of bonds and stocks predicted by the Merton's model. Naturally, the Merton's model is a very raw approximation to the events we are discussing in this section. These events relate to sovereign debt, not firm's debt! At the same time, this model can shed some light into these events. It predicts that bond prices do not move too much when the probability of default is small (i.e. when A_0 is large enough), which might roughly correspond to the situation where an agency announces a name to be downgraded from AAA to AA+. Instead, stocks prices fall, and substantially, due to convexity, following an increase in the probability of default (which occurs when A_0 falls in the Merton's model).



The solid line depicts the value of the bond and the dashed line depicts that of the stock, as predicted by the Merton's model, when the nominal value of debt is $N = 1$, and $T = 1$, $r = 3\%$ and $\sigma = 20\%$, annualized.

While the Merton's model does not obviously regard the joint behavior of stock and sovereign bond prices, it makes a sharp prediction regarding stock and bond prices for a firm. It is an open issue as to whether these predictions would also apply to market developments related to sovereigns. Naturally, the informal arguments of this section do not aim to rule out additional

explanations around the 05-08 events, such as flight-to-quality effects and the expectation hypothesis. Rather, they suggest one additional hypothesis: even absent flight-to-quality effects or the expectation hypothesis, bond prices should not substantially fall, once the probability of default for a "name" still remains very small.

14.2.1.5 First passage

The timing of default can be triggered by some exogeneously specified events. For example, default occurs if the value of the assets hits some exogenously lower bound even before debt expiration. These models are known as "first passage" models, because they rely on mathematical techniques that solve for the probability the first time the asset value hits some exogenous "barrier," as in Black and Cox (1976). It is an interesting case of how asset evaluation theories can develop. The evaluation methods in this section mostly rely on the "option-pricing toolkit" available since Black & Scholes and Merton. However, the first passage approach of Black and Cox would then inspire further work on design and pricing derivatives with barriers (i.e. barrier options), such as down (or up-) and-out and down (or up-) and-in options.

14.2.1.6 Strategic defaulting

The timing of default can actually be endogenous. We now analyze a simple model in which equity holders choose a defaulting barrier (i.e. the firm asset value that triggers bankruptcy) so as to maximize equity value. Naturally, this strategic defaulting infringes upon the assumptions underlying the Modigliani-Miller theorem, such that the value of the firm will now depend on its capital structure. The mechanism in this section is the following. The firm issues debt and needs to ensure coupon payments could flow to the benefit of bondholders. If the value of the assets becomes low enough compared to the nominal value of debt, it would be beneficial for equity holders to default. In fact, the model can be made even richer: issuing debt adds value to the firm because debt acts as a tax-shielding device, so now the time-to-default will be affected by this tax benefit. The higher the tax benefit, the higher the expected time-to-default. Finally, issuing debt exposes the firm to default such that, naturally, the overall value of the firm is also affected by bankruptcy costs (e.g., legal expenses).

More in detail, the model assumes that equity holders finance debt coupons by continuously raising equity capital at each point in time the firm cash flows are not sufficient to honour the coupon payments. The firm cash flows include both asset performance and tax benefits associated with the issuance of debt. Then, a possibility of bankruptcy is opened to the firm, which occurs (endogenously) as soon as the equity holders consider that the value of the assets is too small to warrant them a lifetime positive expected return (in excess of debt services). Specifically, in bad times, when the assets perform poorly, the equity holders do not necessarily liquidate the firm, because there might be chances that the assets could perform better in the future. However, should the assets value become small enough, the equity holders will stop paying the coupons and will liquidate the firm. Naturally, equity holders choose the value of the asset that triggers bankruptcy to maximize the value of equity.

The model we analyze is developed by Leland (1994, Section VI.B), and extended to one with finite maturity debt by Leland and Toft (1996). Anderson and Sundaresan (1996) consider cases of debt re-negotiation. [In progress: Literature review far from being completed] Leland's model considers liquidation of the firm as a strategic choice of the equity holders, as explained. In fact, the US bankruptcy code includes both a liquidation process (Chapter 7) and a reorganization process (Chapter 11), but Leland's model only considers firm's liquidation at bankruptcy. Broadie, Chernov and Sundaresan (2007) generalize this setting to one where the

firm may choose to default through a reorganization process, in which case no equity is issued to honour debt services as in the model analyzed in this section.

Formally, the terms leading to strategic defaulting are as follows. First, to model the firm cash flows, we relax the assumption that the value of the assets, A_t, is solution to Eq. (14.1). Instead, we assume that the firm cash instantaneous flows are equal to a constant fraction δ of the assets A_t, such that, and generalizing Eq. (14.1),[6]

$$\frac{dA_t}{A_t} = (r - \delta)\, dt + \sigma d\tilde{W}_t.$$

Second, debt is infinitely lived, in that it pays off an instantaneous coupon equal to C, forever, conditionally upon survival; in the absence of default risk, the value of debt would simply equal C/r. Third, tax benefits are assumed to be proportional to the coupon, θC. Fourth, there are bankruptcy costs: if the firm defaults at $A = A^B$, recovery is $(1 - \alpha)\, A^B$. Equity holders choose A^B. Naturally, $A^B < A_0$.

The value of debt is a function of the firm asset value, A, say $D(A)$. Moreover, the firm finances the net cost of the coupon C by issuing additional equity, as explained above, and until the equity value is zero, i.e., until $A = A^B$, as seen below. That is, bankruptcy occurs when the firm cannot meet the instantaneous coupon payments. Under the risk-neutral probability, the value of debt satisfies:

$$\underbrace{\frac{d}{dT}\mathbb{E}\left[D\left(A_T\right)\middle|\, A_t\right]\Big|_{T=t}}_{\text{=Expected capital gains}} + \underbrace{C}_{\text{=coupon}} = rD\left(A_t\right). \tag{14.10}$$

Due to Itô's lemma, Eq. (14.10) is an ordinary differential equation, subject to the following boundary conditions. First, at bankruptcy, $D\left(A^B\right) = (1 - \alpha)\, A^B$. Second, for large A, debt is substantially riskless, i.e., $\lim_{A\to\infty} D(A) = \frac{C}{r}$. The solution to (14.10), which also satisfies these boundary conditions, is:

$$D(A) = \left(1 - p_B(A)\right)\frac{C}{r} + p_B(A)\left[(1 - \alpha)\, A^B\right], \tag{14.11}$$

where

$$p_B(A) \equiv \left(\frac{A^B}{A}\right)^{\gamma}, \tag{14.12}$$

and γ is a positive constant.[7] We interpret $p_B(A)$ as the present value of £1, contingent on future bankruptcy, as further explained in Appendix 1. Accordingly, $\left(1 - p_B(A)\right)/r$ is the present value of £1 up to bankruptcy.

The total benefits arising from tax shielding are

$$TB(A) = \left(1 - p_B(A)\right)\theta\frac{C}{r}.$$

and the present value of bankruptcy costs is,

$$BC(A) = p_B(A)\,\alpha A^B.$$

[6] In particular, it is straightforward to check that the evaluation formula in Footnote 2 of this chapter collapses to $A_0 = \mathbb{E}\left(\int_0^\infty e^{-rt}\delta A_t dt\right)$, once the instantaneous cash flows $\delta_t = \delta A_t$.

[7] Notably, $\gamma = \sigma^{-2}(r - \delta - \frac{1}{2}\sigma^2 + \sqrt{\left(r - \delta - \frac{1}{2}\sigma^2\right)^2 + 2\sigma^2 r})$.

Denoting equity with E, we have that

$$E + D = \text{Firm value} = A + TB(A) - BC(A).$$

Summing up,

$$E(A) = A - (1 - p_B(A))(1 - \theta)\frac{C}{r} - p_B(A) A^B. \tag{14.13}$$

Equity equals (i) the firm asset value, A; minus (ii) the present value of debt contingent on no-bankruptcy, net of tax benefits, $(1 - p_B(A))(1 - \theta)\frac{C}{r}$; minus (iii) the present value of debt contingent on bankruptcy, and net of bankruptcy costs, $p_B(A) A^B$. The second term decreases with the default boundary, A^B or, equivalently, $p_B(A)$. The third term, instead, increases with A^B. So the time equity-holders wait before declaring bankruptcy, which is inversely related to A^B, affects these two terms in opposite ways.

Equity-holders choose A^B to maximize the value of equity. The solution is a default boundary, A^B, such that the value of equity does not change for small changes in the value of the assets A around A^B, or $A^B : E'(A)|_{A=A^B} = 0$, a smooth pasting condition, as argued below. The result is:

$$A^B = \frac{\gamma}{1+\gamma}\frac{(1-\theta)C}{r}. \tag{14.14}$$

The expression for A^B is intuitive. Consider, for example, the case in which no tax benefits accrue to the firm while issuing debt, i.e., $\theta = 0$. In this case, defaults occurs when the total assets value falls below a level "a bit lower" than the nominal value of debt, i.e., at $A^B = \frac{\gamma}{1+\gamma}\frac{C}{r}$. As θ increases, the value of A^B in Eq. (14.14) lowers and collapses to zero when the tax benefits exactly compensate for the "debt burden," $\theta = 1$. In this case, $E(A) = A$. Finally, and similarly to the American option case, the value of the option to wait can be shown to be increasing in uncertainty, σ^2.

It is easy to check that A^B does maximize the value $E\left(A; A^B\right) \equiv E(A)$ in that $A^B : 0 = \partial E\left(A; A^B\right)/\partial A^B = 0$.[8] In other words, the equity holders access to an instantaneous dividend given by $\delta A_t - C + \theta C$, such that their optimization program is

$$E(A_t) = \sup_{\tau} E_t\left(\int_t^{\tau} e^{-r(s-t)}\left(\delta A_s - (1-\theta)C\right) ds\right). \tag{14.15}$$

Appendix 1 shows that the solution to this real option problem is that described by Eq. (14.14). That is, the shareholders are willing to accept some temporary negative values of the dividends (and in this case, they would inject new equity capital to make sure debt is honoured), althought then they would declare bankruptcy as soon as the assets value reaches the threshold level A^B in (14.14), such that the expected net present value of the dividends, and, hence, equity, E, is zero.

This model can now be used to analyze how the capital structure affects the value of the firm. When $\theta > 0$, there are indeed levels of leverage that maximize the value of the firm, according to simulations reported in Leland (1994), and this "optimal leverage" can be less than 100% even in the absence of bankruptcy costs, $\alpha = 0$ (Leland, 1994, p. 1230).

[In progress, show some pictures of leverage and the value of the firm]

[8] More formally, it is easy to see that the firm value is maximized by setting A^B as small as possible. This maximization process would imply that $E(A) < 0$ for all $A \in \left(A^B, (1-\theta)\frac{C}{r}\right)$, $A^B \equiv 0$. But limited liability requires that $E(A) \geq 0$ for all $A \geq A^B$. It is easy to see that if $A^B < (1-\theta)\frac{C}{r}$, then, $E''(A) > 0$. Thus, and because $E(A)|_{A=A^B} = 0$, the lowest possible value of A^B consistent with positive equity value is $A^B : E'(A)|_{A=A^B} = 0$.

14.2.1.7 Pros and cons of structural approaches to risky debt assessment

Pros. First, they allow to think about more complicated structures or instruments easily (e.g., convertibles, as we see in the next section). Second, they lead to simple yet consistent relations between different securities issued by the same name. Structural approaches were very useful for theoretical research during the 1990s.

Cons. The firm asset value and asset volatility are not observed. Must rely on calibration/estimation methods. Bond prices generated by the model $\neq$ market prices. These models are a bit difficult to use in practice, for trading or hedging purposes, as we know that in this case we need theoretical prices that exactly match market prices. Finally, how do we go for sovereign issuers?

Most important. Structural models predict unrealistically low short-term spreads: see, e.g., Figure 14.3. The intuition is that diffusion processes are smooth: the probability of default tends to zero as time to maturity approaches zero, because default cannot just jump in an unexpected way. This is not what we exactly observe. Jumps seem to be a more realistic device to modeling spreads, and will be introduced in Section 14.2.3.

14.2.2 The structural approach in practice: the pricing of convertible bonds

Convertible bonds offer bond-holders the option to convert their bonds into shares of the firm.[9] Chapter 12 provides an introductory discussion of these assets (see Section 12.8.2), along with a numerical example of how to price them through a binomial tree. This section analyzes convertible bonds in a continuous time model of the firm's capital structure. We assume the option to convert can be exercized at any time up to maturity. By definition, the face value of the convertible is

$$\text{Face value} = £1 \equiv \text{CR} \times \text{CP}, \tag{14.16}$$

where CR is the *conversion ratio*, i.e. the number of shares this face value converts into, and CP is the *conversion price*, i.e. the stock price implicitly defined by Eq. (14.16).

Typically, the bond is any like other fixed income instrument, with coupon payments, callable features, and credit risk. Callable features are almost invariably embedded into this type of contracts. The parity, or *conversion value*, is the value of the bond if the bond-holders decide to convert. It is defined as

$$\text{CV} = \text{CR} \times S,$$

where S is the price of the common share. Not only is the convertible bond price affected by interest rates, credit risk, or timing risk; it is also affected by the movements of the underlying stock price, due to a positive probability the bond might "become" a share in the future—convertible bonds are *hybrid* instruments. The embedded option offers the bond-holders the possibility to obtain equity returns (not just bond returns) in good times, while offering protection against the downside. As mentioned in Chapter 12 (see Section 12.8), convertible bonds are usually callable as well. In this case, bond-holders are usually given the right to convert the bonds, once they are called. The rationale behind callability is to induce the bond-holder to convert the bond earlier.

Useful to trade volatility. Simplest example of convertible arbitrage is going long a convertible and shorting a Treasury, which is the same as going long an option on the firm. Useful when there are no available options on the firm to trade, and/or when these are very illiquid.

[9] Strictly speaking, the "option" embedded into this kind of asset is a *warrant*, not an option. A warrant gives the holder a right to purchase new shares, i.e. shares issued by the firm.

Pricing convertible bonds is a topic that has been intensively studied, theoretically. Ingersoll (1977) provides the first theoretical insights into the pricing of convertible and callable bonds. Let us define the *dilution factor*, denoted as γ, as the fraction of common equity that would be held by the convertible bondowners if the entire issue was converted. If there are n^{out} shares outstanding, and the convertible bond can be exchanged for n shares, then, in aggregate,

$$\gamma = \frac{n}{n + n^{\text{out}}}.$$

We assume that the market value of the firm is equal to the value of its assets, A, which is a Geometric Brownian motion, as in Eq. (14.1). Let $B^{\text{conv}}(A, \tau; N)$ be the aggregate value of the convertible bond with time to maturity τ and face value N. To simplify the presentation, we do not consider callability issues. However, we shall provide some intuition about this issue later. Let us assume that the stocks and the convertible bonds are the only two claims in the capital structure of the firm. Since, after conversion, only the stocks will remain, then, the post-conversion value of the convertible bonds is simply the conversion value of the convertible, i.e. γA. Moreover, we have that for any $\tau \geq 0$,

$$\gamma A \leq B^{\text{conv}}(A, \tau; N) \leq A. \tag{14.17}$$

The first inequality in (14.17) is simple to understand. Indeed, suppose that $B^{\text{conv}}(A, \tau; N) < \gamma A$. Then, we can purchase the convertibles, convert them into shares and, finally, sell the shares for γA. The second inequality follows by the limited liability of the equity holders, and the Modigliani-Miller theorem.

At maturity,

$$B^{\text{conv}}(A, 0; N) = \min\{A, \max\{N, \gamma A\}\}. \tag{14.18}$$

Indeed, $\bar{B} \equiv \max\{N, \gamma A\}$ is the value of the convertible, in case of no-default. Then, $\min\{A, \bar{B}\}$ is what the firm will pay to the bond-holders: A in case of default, and $\bar{B}$ in case of no-default.

We can re-express the terminal payoff in Eq. (14.18) in a manner that allows for a better understanding of the issues underlying the exercise of the convertibles. In particular, we have that,

$$B^{\text{conv}}(A, 0; N) = \min\{A, \max\{N, \gamma A\}\} = \max\{\gamma A, \min\{A, N\}\}. \tag{14.19}$$

Indeed, let $\hat{B} \equiv \min\{A, N\}$ the payment the firm is ready to supply if the the bond-holders do not convert. Then, $\max\{\gamma A, \hat{B}\}$ is obviously the payoff profile to the bond-holders.

The terminal payoff in Eq. (14.19) illustrates very clearly that convertible bonds embed an option to convert, on top of the plain vanilla non-convertible bond. Intuitively, at maturity, a non-convertible bond is worth $\min\{A, N\}$, and the option to convert is either worthless (in case of non conversion) or worth $\gamma A - N$ (in case of conversion), i.e. it is $\max\{\gamma A - N, 0\}$. This intuition is confirmed, mathematically, as we have that:

$$\max\{\gamma A, \min\{A, N\}\} = \min\{A, N\} + \max\{\gamma A - N, 0\}.$$

Therefore, the value of the terminal payoff is, by Eq. (14.19),

$$B^{\text{conv}}(A, 0; N) = \min\{A, N\} + \gamma \max\{A - N/\gamma, 0\}. \tag{14.20}$$

One can show that it is not optimal to exercize the option to convert before maturity. Therefore, to price the convertible bond, we only need to deal with the terminal payoff in Eq. (14.20).

Eq. (14.20) shows that the current value of the convertible bond is the sum of the value of a "straight" bond plus the value of γ options on the firm with strike price equal to N/γ. Accordingly, let $B(A,\tau;N)$ and $W(A,\tau;N/\gamma)$ be the prices of the straight bond and the option on the firm. We have,

$$B^{\text{conv}}(A,\tau;N) = B(A,\tau;N) + \gamma W(A,\tau;N/\gamma). \tag{14.21}$$

We may use the Merton's (1974) model to find the price of the straight bond, $B(A,\tau;N)$. By results in Section 14.2, it is

$$B(A,\tau;N) = A\Phi(-d_1) + Ne^{-r\tau}\Phi\left(d_1 - \sigma\sqrt{\tau}\right), \quad d_1 = \frac{\ln(A/N) + \left(r + \frac{1}{2}\sigma^2\right)\tau}{\sigma\sqrt{\tau}}, \tag{14.22}$$

where σ is the instantaneous volatility of the assets, r is the (constant) instantaneous short-term rate, and Φ is the cumulative distribution of a standard normal. Similarly, we may use the Black-Scholes formula to compute the function W:

$$W(A,\tau;N/\gamma) = A\Phi(d_1) - \frac{N}{\gamma}e^{-r\tau}\Phi\left(d_1 - \sigma\sqrt{\tau}\right). \tag{14.23}$$

Eq. (14.22) reveals the intuitive property that as A gets large, $B(A,\tau;N) \approx Ne^{-r\tau}$: the probability of default gets extremely tiny as the value of the assets gets large. Moreover, the Black-Scholes formula, Eq. (14.23), suggests that $W(A,\tau;N/\gamma) \approx A - e^{-r\tau}N/\gamma$ as A gets large. Therefore, by Eq. (14.21), we have that, for large A, $B^{\text{conv}}(A,\tau;N) \approx Ne^{-r\tau} + \gamma(A - e^{-r\tau}N/\gamma) = \gamma A$. Eq. (14.22) also shows that for small values of A, $B^{\text{conv}}(A,\tau;N) \approx 0$. To sum-up, the value of the convertible bond is less than the value of the firm, A, and larger than the conversion value, γA. Moreover, it approaches γA, as the value of the firm gets large. Figure 14.6 depicts the price of the convertible bond as function of the value of the firm, as predicted by Eq. (14.21), for a particular example. It is possible to show that the value of a callable convertible bond is between the value of the straight and that of the convertible.

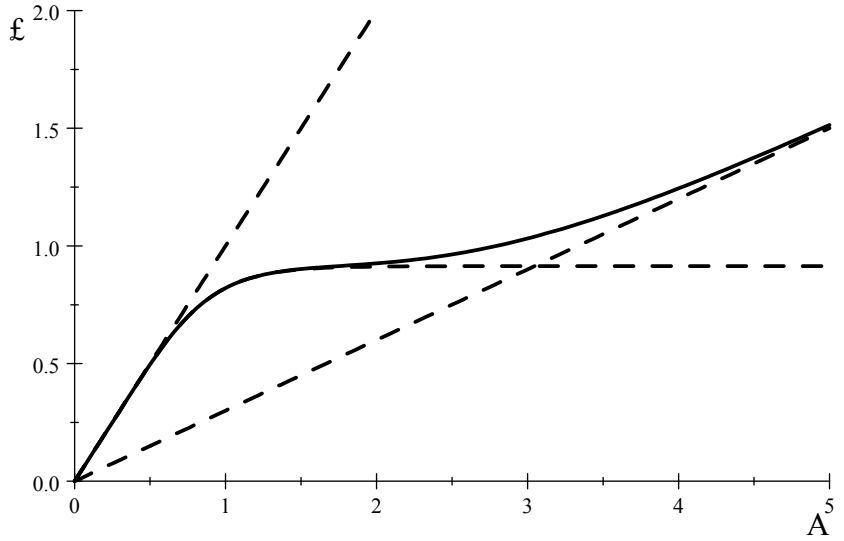


FIGURE 14.6. The value of convertible and straight bonds as a function of the current asset value, A, when the short-term rate $r = 3\%$, the asset volatility $\sigma = 0.20$, time to maturity $\tau = 3$ years, the dilution factor $\gamma = 30\%$, and nominal debt $N = 1$. The solid

line depicts the value of the convertible bond. The dashed straight line starting from the origins, and flattening out to the constant $Ne^{-r\tau} = 0.91393$, is the value of the straight bond. The two dashed straight lines starting from the origins are the no-arbitrage bounds γA and A in Eq. (14.17).

14.2.3 Reduced form approaches: rare events, or intensity, models

Default often displays a few striking features. It arrives unexpectedly, it is rare, and causes discontinuous price changes. The structural models in the previous section do not accommodate for these features because diffusion processes are continuous, as explained in Chapter 5. As a result, passage times are known, "locally," so to speak. This feature is responsible of the low short-term spreads these models predict.

14.2.3.1 Poisson-driven defaults

As an alternative to Brownian motions, we can model defaults, by assuming their arrival is a Poisson process, of the kind introduced in Chapter 5. Suppose to "count" the number of times some event happens. Denote with N_t the corresponding "counting process," as in Figure 14.7.

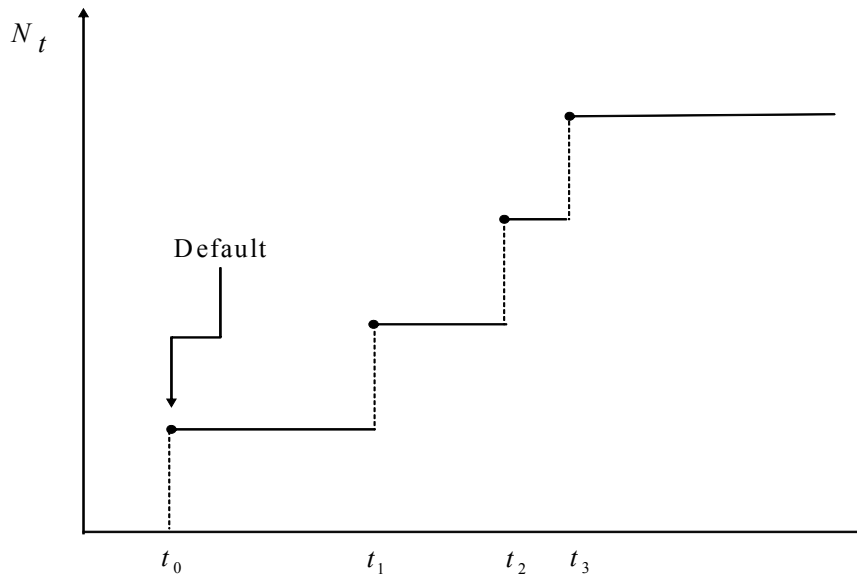


FIGURE 14.7.

The time of default is simply the first time N_t "jumps," e.g. t_0 in Figure 14.7. More in detail, assume we chop a given interval $[0, T]$ in n pieces, and consider each resulting interval $\Delta t = \frac{T}{n}$. Assume that the jump probability over each of these small intervals of time Δt is proportional to Δt, with proportionality factor equal to λ,

$$p \equiv \Pr\{\text{One jump over } \Delta t\} = \lambda \Delta t. \tag{14.24}$$

Assume the number of jumps over the n intervals follows a binomial distribution:

$$\Pr\{k \text{ jumps over } [0, T]\} = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{where } p = \lambda \frac{T}{n}.$$

For n large, or, equivalently, for small intervals Δt,

$$\Pr\{k \text{ jumps over } [0,T]\} \approx \frac{(\lambda T)^k}{k!}\left(1-\frac{\lambda T}{n}\right)^{n-k} \approx \frac{(\lambda T)^k}{k!}e^{-\lambda T}.$$

We rely on these heuristic calculations and derive a few basic properties of default. We have,

$$\begin{aligned}\Pr\{\text{Survival by } T\} &= \Pr\{0 \text{ jumps over } [0,T]\} = e^{-\lambda T}\\ \Pr\{\text{Default by } T\} &= \Pr\{\text{at least one jump over } [0,T]\}\\ &= 1-\Pr\{\text{Survival by } T\} = 1-e^{-\lambda T}\\ \Pr\{\text{Default occurs at some } t\} &= \lambda e^{-\lambda t}dt\end{aligned}$$

Note that the expected time to default equals λ^{-1}.

We can use these probabilities to value debt subject to default risk. Consider a simple case, namely when default can occur over any $t \in (t,T]$, *but default implications do only occur at* T, in that at T, the bond is equal to either the recovery value (in case of default over any $t \in (t,T]$) or the nominal value (in case of survival). In this case we have, by Eq. (14.6), that:

$$D_0 = e^{-rT}\underbrace{\left[\text{Rec}_N\mathbb{Q}\left(\text{Default}\right) + N\mathbb{Q}\left(\text{Survival}\right)\right]}_{\equiv B_0}, \tag{14.25}$$

where $\text{Rec}_N \equiv \text{Rec}\cdot N$ denotes the expected recovery face value of the asset. Using the probabilities predicted by the Poisson model, we obtain

$$B_0 = \text{Rec}_N\left(1-e^{-\lambda T}\right) + Ne^{-\lambda T} = N\cdot\left(1-\text{LGD}\left(1-e^{-\lambda T}\right)\right). \tag{14.26}$$

Appendix 2 supplies an alternative derivation of Eq. (14.26). The unattractive feature of the model is that there is no time variation in the term structure of credit spreads. But the model illustrates the main economic mechanisms underlying asset evaluation in the presence of default risk. Note how intuitive the expression for the bond price is. The second equality of (14.26) says that the (undiscounted) price of defaultable bond is a "discount" that takes into account the expected loss, defined as LGD times the probability of default, $\left(1-e^{-\lambda T}\right)$. Models with random intensity rates (λ) predict time variation in the term structure of credit spreads (see Section 14.3.2.5), although the intuition underlying them follows that for the simple model in this section (see Eq. (14.41)).

14.2.3.2 Predicted spreads

The implications for the spreads corresponding to small maturities T can be easily seen after some approximations,

$$\text{Spread} = -\frac{1}{T}\ln\left(\frac{B_0}{N}\right) \approx -\frac{1}{T}\left(\frac{B_0}{N}-1\right) = \frac{1}{T}\left(1-e^{-\lambda T}\right)\cdot\text{LGD}.$$

Note that in contrast to the structural models reviewed in Section 14.2.1, the spread is not zero when T is small. Rather, it is given by the expected default loss per period, defined as the instantaneous probability of default times LGD,

$$\text{Short-Term Spread} = \lambda\cdot\text{LGD}.$$

Therefore, models with jumps have the potential to explain the empirical behavior of credit spreads at short maturities discussed in Section 14.2.1. As explained, structural models, being typically driven by Brownian motions, cannot lead to positive spreads at very short maturities, as they imply that the probability of default decays quickly as time-to-maturity goes to zero. Instead, models with jumps predict a possibility the firm can experience a "sudden death": default can occur with positive probability at any time, even when the debt is about to expire.

A theoretical model of Duffie and Lando (2001) shows how a structural model of the firm can lead to positive short-term spreads, once we assume incomplete information and learning about the asset value. In their model, learning takes place with some delay, which leaves investors concerned about what they really know about the firm asset value. It is this concern to lead to positive credit spreads in their model, to an extent comparable to that generated by a jump process.

Figure 14.8.1 depicts the behavior of the spread predicted by the model at all maturities, given by[10]

$$\text{Spread} = -\frac{1}{T}\ln\left(\frac{B_0}{N}\right) = -\frac{1}{T}\ln\left(\frac{\text{Rec}}{N}\cdot\left(1-e^{-\lambda T}\right)+e^{-\lambda T}\right).$$

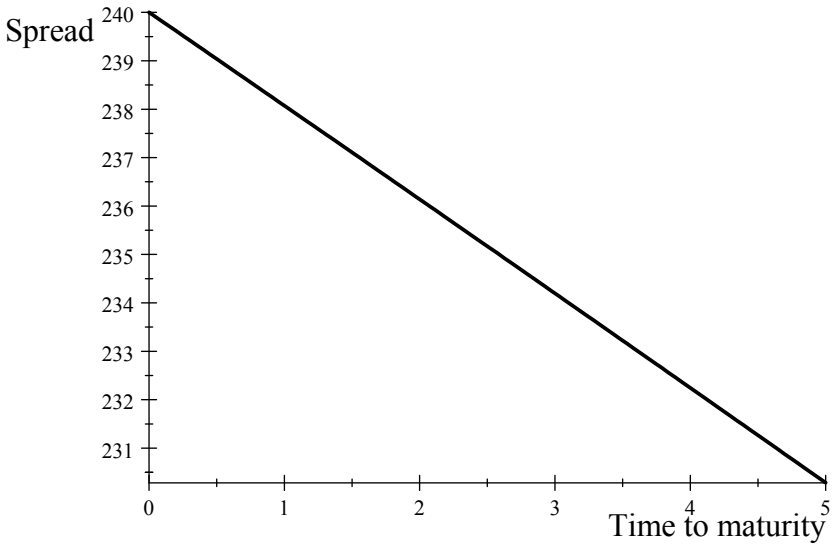


FIGURE 14.8.1. The term structure of bond spreads (in basis points) implied by an intensity model, with recovery rate equal to 40% and intensity equal to $\lambda = 0.04$, implying an expected time-to-default equal to $\lambda^{-1} = 25$ years.

In this example, spreads are decreasing in time-to-maturity. Eventually, as time to maturity gets large, the bond becomes somehow certain to default, with the unusual feature to deliver, for sure, some recovery rate at some point—the bond is certain to deliver the recovery rate. Indeed, in Appendix 2, we show that if the recovery value of the bond is not constant, but shrinks exponentially to zero as $\text{Rec}_T \equiv R\cdot e^{-\kappa T}$, for two constants R and κ, then, asymptotically, the spread is:

$$\lim_{T\to\infty} s(T) = \min\{\kappa,\lambda\}. \tag{14.27}$$

[10]We are taking the year as the reference period. Thus, a value of $\lambda = 0.04$ means that we expect a credit event (i.e., default) to occur every 25 years: from Chapter 5 (see Section 5.10), the expected number of events on a time interval $T-t$ is equal to $\lambda(T-t)$.

The interpretation of κ does not link to discounting. Rather, κ might be referred to as a "recovery dissipation rate" due to unfolding of time, in the following sense. As time unfolds, the likelihood of occurrence of bad events increases, which leads the expected recovery to worsen. Eq. (14.27) shows that if the dissipation rate is sufficiently large, term spreads can be increasing, as we discuss more comprehensively in a moment.

An instance leading to such an expected recovery rate is one where the recovery value of the bond equals R, if the firm defaults at any time T, and provided an hidden risk does not materialize, namely the risk that the firm will not distribute any recovery value at all, in case of bankruptcy. If this risk is independent of bankruptcy, and Poisson, with instantaneous risk-neutral probability κ, the expected recovery is precisely $\text{Rec}_T = R \cdot e^{-\kappa T}$. This is indeed a quite simple way to model stochastic recovery rates.

Figure 14.8.2 plots the term-structure of spreads predicted by this model, obtained with the same parameter values used to produce the spreads in Figure 14.8.1, and utilizing three values for the dissipation rate: $\kappa = 0.05$, 0.03 and 0.013. Naturally, instantaneous spreads (i.e. those corresponding to time to maturity equal to zero) are $(1 - R) \cdot \lambda = 240$ (in basis points) in all cases.

When $\kappa > \lambda$, long-term maturity spreads are always higher than short, by Eq. (14.27). When $\kappa < \lambda$, instead, spreads for large maturities can be either higher or lower than short, according to whether κ is higher or lower than $(1 - R) \cdot \lambda$. In this particular example, long-term spreads equal 400 basis points, i.e. the default intensity, λ. When $\kappa = 0.03$, which is lower than $\lambda = 0.04$, they are higher and when $\kappa = 0.013$, they are lower. In fact, when $\kappa = 0.013$, the term structure of the spreads is even hump-shaped, a feature not visible from the picture. As is clear, this very simple model predicts features of both short-term and long-term spreads that the Merton's model in Section 14.2.1.1 cannot, realistically.

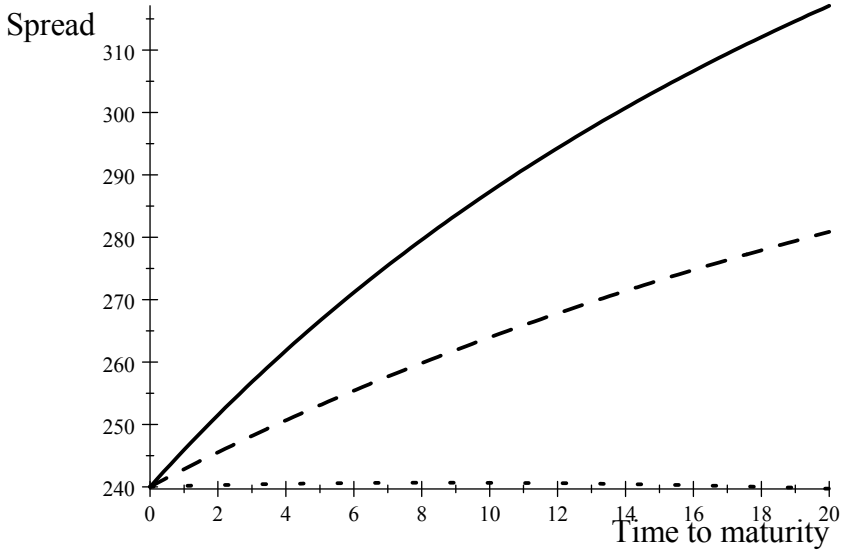


FIGURE 14.8.2. The term structure of bond spreads (in basis points) implied by an intensity model with recovery rate equal to $0.40e^{-\kappa\tau}$, where τ is time to maturity and κ is the recovery dissipation rate, taken to equal $\kappa = 0.05$ (solid line), $\kappa = 0.03$ (dashed line), and $\kappa = 0.013$ (dotted line). The instantaneous probability of default is taken to equal to $\lambda = 0.04$, implying an expected time-to-default equal to $\lambda^{-1} = 25$ years.

We can interpret this behavior of term spreads as follows. Suppose we are in good times, when λ is small relative to κ. We are in good times precisely because we expect things would change adversely in the future, captured by a large value of κ. In this case, the term structure of spreads is increasing. Instead, in bad times, when λ is large compared to κ, we might expect future times to improve, which we might model by assuming κ is small.

Figure 14.8.2 shows that long maturity spreads are smaller than in good times. Naturally, we would expect that spreads should increase for any maturity in bad times, although this property is not captured by the numerical examples in Figure 14.8.2, where we fix $\lambda = 0.04$. The point of this exercise is to show that the *slope* of the term structure of the spreads lowers as we enter bad times, when we only consider changes in the dissipation rate, κ. Allowing for a countercyclical λ would reinforce the conclusions of this exercise. While these conclusions rely on comparative statics, Section 14.4.5.5 shows that they still hold in a dynamic context, where the intensity λ follows a mean-reverting continuous-time model.

14.2.3.3 One example

Naturally, the intensity, λ, is the risk-neutral instantaneous probability of default, not the physical, λ^* say. The ratio λ/λ^* is generally larger than one. Its inverse, λ^*/λ, is an indicator of the risk-appetite in the credit market. Similarly, LGD is an expectation under the risk-neutral probability, and should contain useful indications about market participants risk appetite.

Assume that under the risk-neutral probability, the instantaneous intensity of default for a given firm is $\lambda = 4\%$, annualized, and that under the physical, the instantaneous probability of default for the same firm is $\lambda^* = 2\%$, annualized. From here, we can compute the probability of survival of the firm within 5 years, under both probabilities. They are:

$$e^{-5\lambda} = e^{-5\times 0.04} = 0.81873, \quad e^{-5\lambda^*} = e^{-5\times 0.02} = 0.90484.$$

Naturally, the probability of survival is lower under the risk-neutral probability.

Next, assume that the spread on a 5 year bond with face value $N = 1$, equals 3%. What is the implied expected recovery rate from this spread? We have,

$$D_0 = e^{-rT}\left[\text{Rec}\cdot\mathbb{Q}\left(\text{Default}\right) + N\cdot\mathbb{Q}\left(\text{Survival}\right)\right] = e^{-rT}\left[\text{Rec}\times(1-0.81873) + 1\times 0.81873\right].$$

The spread is,

$$\text{s}_0 = 3\% = -\frac{1}{5}\ln\left(\frac{\text{Rec}\times(1-0.81873)+1\times 0.81873}{1}\right).$$

Solving for Rec, gives, Rec = 23.16%.

14.2.4 Ratings

In practice, corporate debt is rated by rating agencies, such as Moodys and Standard and Poor's. Depending on the rating, corporate debt may be either investment grade or non-investment grade ("junk"). Moodys ratings range from Aaa to C. Standard and Poor's range from Aaa to D. One can compute the probability of "migrations" based on past experience ⟶ Transition

probabilities. Consider, for example, the following table:

One year rating transition probabilities (%), S&P's 1981-1991

		To							
		AAA	AA	A	BBB	BB	B	CCC	D
From	AAA	89.1	9.63	0.78	0.19	0.3	0	0	0
	AA	0.86	90.1	7.47	0.99	0.29	0.29	0	0
	A	0.09	2.91	88.94	6.49	1.01	0.45	0	0.09
	BBB	0.06	0.43	6.56	84.27	6.44	1.6	0.18	0.45
	BB	0.04	0.22	0.79	7.19	77.64	10.43	1.27	2.41
	B	0	0.19	0.31	0.66	5.17	82.46	4.35	6.85
	CCC	0	0	1.16	1.16	2.03	7.54	64.93	23.19
	D	0	0	0	0	0	0	0	100

TABLE 14.1

14.2.4.1 Foundations

A natural approach is to assess credit risk by making reference to probabilities of default built up on transition probabilities like those in Table 14.1.

Such an approch, also known as a migration approach, is somewhat less drastic than that based on rare events, and hopefully more realistic. However, it is also technically more complex than the intensity approach of the previous section. We provide the most foundational issues of this approach, leaving some details in the Appendix.

At time t, there exists several rating classes, Z say, denoted as Rat_t,

$$\text{Rat}_t \in \{1, 2, \cdots, Z\}.$$

Transition probabilities of rating from time t to time T are,

$$P(T-t)_{ij} \equiv \Pr(\text{Rat}_T = j | \text{Rat}_t = i), \quad i, j \leq Z.$$

We can build a Markov chain from here, by assuming that $P(T-t)_{ij}$ only depends on $T-t$. Finally, we must have that,

$$P(T-t)_{ij} \geq 0 \text{ and } \sum_{j=1}^{Z} P(T-t)_{ij} = 1.$$

For example, the probability of transition from rating $\text{Rat}_t = i$ to rating $\text{Rat}_{t+1} = j$ in one year is, $P(1)_{ij}$. Table 14.1 contains one possible example of $P(1)_{ij}$. The probability of transition from rating $\text{Rat}_t = i$ to rating $\text{Rat}_{t+2} = j$ in two years is $P(2)_{ij}$, and is obtained as follows,

$$P(2)_{ij} = \sum_{k=1}^{Z} \underbrace{P(1)_{ik}}_{\Pr(\text{transition from } i \text{ to } k \text{ in one year})} \cdot \underbrace{P(1)_{kj}}_{\Pr(\text{transition from } k \text{ to } j \text{ in one further year})}$$

More generally, we have, $P(T) = P(1)^T$, where $P(T)$ is the matrix with elements $\{P(T)_{ij}\}$.

For example, the probability transition matrix P in Table 14.1 is,

$$P = \begin{bmatrix} 89.1 & 9.63 & 0.78 & 0.19 & 0.3 & 0 & 0 & 0 \\ 0.86 & 90.1 & 7.47 & 0.99 & 0.29 & 0.29 & 0 & 0 \\ 0.09 & 2.91 & 88.94 & 6.49 & 1.01 & 0.45 & 0 & 0.09 \\ 0.06 & 0.43 & 6.56 & 84.27 & 6.44 & 1.6 & 0.18 & 0.45 \\ 0.04 & 0.22 & 0.79 & 7.19 & 77.64 & 10.43 & 1.27 & 2.41 \\ 0 & 0.19 & 0.31 & 0.66 & 5.17 & 82.46 & 4.35 & 6.85 \\ 0 & 0 & 1.16 & 1.16 & 2.03 & 7.54 & 64.93 & 23.19 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 100 \end{bmatrix}$$

The 15 year transition matrix is:

$$P(15) \approx \begin{bmatrix} 20.01 & 35.82 & 23.91 & 9.92 & 4.05 & 3.06 & 0.43 & 2.66 \\ 3.38 & 30.28 & 32.71 & 15.91 & 6.38 & 5.11 & 0.77 & 5.34 \\ 1.17 & 13.12 & 34.21 & 21.93 & 9.69 & 8.01 & 1.29 & 10.33 \\ 0.64 & 6.76 & 22.21 & 22.40 & 12.42 & 11.93 & 2.09 & 21.39 \\ 0.33 & 3.22 & 10.71 & 13.616 & 11.36 & 14.68 & 2.78 & 43.16 \\ 0.14 & 1.65 & 5.01 & 6.75 & 7.48 & 13.17 & 2.64 & 63.04 \\ 0 & 1.08 & 3.54 & 3.90 & 3.51 & 5.60 & 1.22 & 81.02 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 100 \end{bmatrix}$$

14.2.4.2 Evaluation

The previous probabilities, $\{P(T)_{ij}\}$, are meant to be taken under the *physical* world, not the *risk-neutral*. They can be used for risk-management purposes, but certainly not for pricing. Indeed, historical default rates are too low to explain the price of defaultable securities. A natural explanation relies on the presence of risk-premia. To use migration data for pricing, it is vital to implement a number of steps.

The first step relates to clean up the data. For example, it might be that downgrades from class i to class $i+2$ are more frequent than downgrades from class i to class $i+1$, an occurrence which we wish to smooth. Moreover, we would need to remove zero entries: although some rating events did not happen in the past, they might well occur in the future. Finally, we need to add positive risk-premia to the previous smoothed data, to recover realistic asset prices.

As for the pricing details, the migration model relies on the assumption that there are Z classes of assets. Each single asset may migrate from one class to another. Because evaluation is a dynamic business, we cannot evaluate defaultable securities within a given class without simultaneously evaluate the defaultable securities in the remaining classes. For example, there could be a chance that a given asset will "mutate" into a different one over the next year (i.e. one belonging to another rating class). Therefore, the price of this asset, today, needs to reflect the price of the asset in the other classes where it can possibly migrate. As a result, we must simultaneously solve for all the asset prices in all the rating classes. This approach, developed by Jarrow, Lando and Turnbull (1997), is quite complex and is given a succinct account in the Appendix.

Consider a simple case, arising when default can occur at any time before maturity T, but default implications arise only at maturity, in that the bond pays a recovery value only at T, should the firm default at any time prior to T. Let $\mathbb{Q}_i(T-t)$ denote the *risk-neutral* probability the firm defaults, by time T, given it belongs to rating i at time T. By Eq. (14.6),

$$\frac{D_{0,i}}{N} = e^{-rT}\left[\frac{\text{Rec}}{N}\mathbb{Q}_i(T-t) + (1-\mathbb{Q}_i(T-t))\right].$$

The risk neutral probabilities, $\mathbb{Q}_i(T-t)$, must be found using migration frequencies such as those in Table 14.1, which we must "clean up" and corrct with appropriate risk-premia as discussed.

14.2.4.3 One example

Consider the following transition matrix:

		To		
		A	B	Def
From	A	0.9	0.07	0.03
	B	0.15	0.75	0.10
	Def	0	0	1

where Def denotes the state of default. What is the probability that a name A will remain name A in two years? What is the probability that a name A will default in two years?

Consider the following two year transition matrix:

$$\mathbb{Q}(2) = \underbrace{\begin{bmatrix} 0.90 & 0.07 & 0.03 \\ 0.15 & 0.75 & 0.10 \\ 0 & 0 & 1 \end{bmatrix}}_{\equiv \mathbb{Q}(1)} \cdot \underbrace{\begin{bmatrix} 0.90 & 0.07 & 0.03 \\ 0.15 & 0.75 & 0.10 \\ 0 & 0 & 1 \end{bmatrix}}_{\equiv \mathbb{Q}(1)},$$

such that:

$$\begin{aligned} \Pr\{\text{A is A in 2 years}\} &= \underbrace{0.90 \times 0.90}_{A \to A \to A} + \underbrace{(0.07) \times (0.15)}_{A \to B \to A} + \underbrace{0.03 \times 0}_{A \to Def \to A} \\ &= 0.8205, \end{aligned}$$

and

$$\begin{aligned} \Pr\{\text{A defaults in 2 years}\} &= \underbrace{0.90 \times 0.03}_{A \to A \to Def} + \underbrace{(0.07) \times (0.10)}_{A \to B \to Def} + \underbrace{0.03 \times 1}_{A \to Def \to Def} \\ &= 0.064. \end{aligned}$$

In general, we have that:

$$\mathbb{Q}(2)_{ij} = \sum_{k=1}^{3} \mathbb{Q}(1)_{ik} \cdot \mathbb{Q}(1)_{kj},$$

and for any T,

$$\mathbb{Q}(T) = \mathbb{Q}(1)^T = \begin{bmatrix} 0.90 & 0.07 & 0.03 \\ 0.15 & 0.75 & 0.10 \\ 0 & 0 & 1 \end{bmatrix}^T.$$

Next, consider the following transition matrix, under the risk-neutral probability:

		To		
		A	B	Def
From	A	0.80	0.20	0
	B	0.15	0.75	0.10
	Def	0	0	1

From here, we may easily compute, again, the (risk-neutral) probability A will default in two years, and the probability B will default in two years. We have,

$$\mathbb{Q}(2) = \underbrace{\begin{bmatrix} 0.80 & 0.20 & 0 \\ 0.15 & 0.75 & 0.10 \\ 0 & 0 & 1 \end{bmatrix}}_{\equiv \mathbb{Q}(1)} \cdot \underbrace{\begin{bmatrix} 0.80 & 0.20 & 0 \\ 0.15 & 0.75 & 0.10 \\ 0 & 0 & 1 \end{bmatrix}}_{\equiv \mathbb{Q}(1)},$$

such that:

$$\begin{aligned} \Pr\{\text{A defaults in 2 years}\} &= \mathbb{Q}(2)_{13} \\ &= \underbrace{0.80 \times 0}_{A \to A \to Def} + \underbrace{(0.20) \times (0.10)}_{A \to B \to Def} + \underbrace{0 \times 1}_{A \to Def \to Def} \\ &= 0.02. \end{aligned}$$

(multiply first row by the third column), and,

$$\begin{aligned} \Pr\{\text{B defaults in 2 years}\} &= \mathbb{Q}(2)_{23} \\ &= \underbrace{0.15 \times 0}_{B \to A \to Def} + \underbrace{(0.75) \times (0.10)}_{B \to B \to Def} + \underbrace{0.10 \times 1}_{B \to Def \to Def} \\ &= 0.175. \end{aligned}$$

(multiply second row by the third column).

Finally, suppose that the bonds issued by both A and B mature in two years. Furthermore, assume that if these two bonds default, they pay off the same recovery rate, equal to 30%, and only at the end of the second period. From here, we can compute the credit spreads for the two bonds. We have,

$$\begin{aligned} e^{rT} \times \text{Price_A} &= (0.30) \times (0.02) + (1 - 0.02) = 0.986 \\ &\Rightarrow \text{Spread_A} = -\frac{1}{2} \ln(0.986) = 7.0495 \times 10^{-3}. \end{aligned}$$

and,

$$\begin{aligned} e^{rT} \times \text{Price_B} &= (0.30) \times (0.175) + (1 - 0.175) = 0.8775 \\ &\Rightarrow \text{Spread_B} = -\frac{1}{2} \ln(0.8775) = 6.5339 \times 10^{-2}. \end{aligned}$$

14.3 Credit derivatives and structured products based thereon

14.3.1 Options and spreads

14.3.1.1 Total Return Swaps

In a total return swap (TRS, henceforth), one party, who owns some asset, that underlying the TRS, receives from the counterparty payments based on a mutually agreed rate, either fixed or variable, and makes payments to the counterparty based on the *return* of the underlying asset, which includes both the income it generates and any capital gains. The underlying asset can be a loan, a bond, an equity index, or a basket of assets. The interest payments are typically

based on the LIBOR plus a spread. Consider the following example. Party A receives LIBOR + fixed spread equal to 3%. Party B receives the total return of the S&P 500 on a principal amount of £1 million. If the LIBOR is 7% and the S&P 500 is up by 12%, A pays B 12% and B pays A 7% + 3%. By netting, A pays B £20,000, i.e. £1 million × (12% − 10%).

While TRS are usually categorized as credit derivatives, they combine both market risk and credit risk. The main benefit from going long a TRS is that the party with the asset on the balance sheet buys protection against loss in value. The main benefit from shorting a TRS is that it allows the counterparty to receive the payoffs of the underlying without necessarily having to put this underlying in the balance sheet. Hedge funds find it quite convenient to short a TRS, as this allows them to have views with limited collateral upfront. The market for TRS is over-the-counter and market participants include institutions only.

14.3.1.2 Spread Options

Spread Options (SO, henceforth) are options written on the difference between two indexes. For example, let $S_1(T)$ and $S_2(T)$ be the prices of two assets at time T. The payoff promised by a SO entered at some time $t < T$, might be $\max\{S_1(T) - S_2(T) - K, 0\}$, where K is the strike of the SO. A SO can be written on the spread between two rates of returns too. Importantly, a SO can be written on the spread between the yield of a corporate bond and the yield of a Treasury bond. Examples include: (i) NOB spread (notes - bonds), which are spreads between maturities; (ii) Spreads between quality levels, such as the TED spread (treasury bills − Eurodollars); (iii) MOB spreads, i.e. the difference between municipal bonds and treasury bonds. More generally, the definition of a SO has now been extended to include payoffs written as a *linear combination* of indexes, interest rates and yields.

14.3.1.3 Credit spread options

Credit spread options (CSO, henceforth) are options where the payoff is the difference between (i) the spread between two reference securities (say Italian Government bonds and US Government bonds having the same maturity, or the spread between some stock return in excess of the LIBOR, or two credit instruments), and (ii) a given strike spread, for a certain maturity date. It may be an American or European option. So CSOs allow to hedge against, or take specific views about, changes in credit spreads. For example, an investor, while bullish on Italian bonds, might hedge against the uncertain outcome of a political election, which could trigger a widening of short-term spreads of Italians versus US. The investor, then, might go long a CSO, with time to maturity around the days of the political election, where the underlying are the Italian and Government bonds expiring in ten years, say. A possible payoff to the CSO holder might be proportional to, (ITA/US − K), where ITA/US is the ten year Italian-US spread in three months, and K is the strike spread.

14.3.2 Credit Default Swaps

14.3.2.1 Single name swaps

TRS provide protection against a general loss in asset value, which could be triggered by both market or credit risk, although it is obviously more often market risk than credit to kick in. Credit Default Swaps (CDS, henceforth) differ from TRS insofar as they provide protection against a credit event.

The premium, assumed to be paid quarterly, on a CDS contract agreed at time t, is obtained by equating the expected discounted value of the protection (the *floating protection leg*), to

the expected discounted value of all the premiums paid over the life of the contract (the *fixed premium leg*), i.e. at dates $t_i : t < t_1 < t_2 < \cdots < t_{4M}$, where $t_i = t + \frac{i}{4}$, and M is the number of years of the CDS contract. The discounted expected floating protection leg is:

$$\text{Protection}_t = \sum_{i=1}^{4M} e^{-r(t_i - t)} \text{LGD}\left(t_i\right) \Pr\left\{\text{Def.} \in \left(t_{i-1}, t_i\right) \middle| \, \text{Surv. at } t_{i-1}\right\},$$

and the discounted expected fixed leg is

$$\text{Premium}_t = \sum_{i=1}^{4M} e^{-r(t_i - t)} \text{CDS}_t\left(M\right) \Pr\left\{\text{Surv. at } t_i\right\},$$

where r is the (constant) risk free rate, $\text{CDS}_t\left(M\right)$ is the premium paid every quarter, prevailing at time t, and $\text{LGD}\left(t_i\right)$ is the LGD at time t_i, which for simplicity is assumed to be constant, i.e. known at time t.

Equating Premium_t and Protection_t, and solving for $\text{CDS}_t\left(M\right)$, leaves:

$$\text{CDS}_t\left(M\right) = \frac{\sum_{i=1}^{4M} e^{-r(t_i - t)} \text{LGD}\left(t_i\right) \Pr\left\{\text{Def.} \in \left(t_{i-1}, t_i\right) \middle| \, \text{Surv. at } t_{i-1}\right\}}{\sum_{i=1}^{4M} e^{-r(t_i - t)} \Pr\left\{\text{Surv. at } t_i\right\}}. \tag{14.28}$$

It is a forward premium, as long as $t < t_0$. We assume that if the obligor defaults prior to the start date, t_0, the contract is terminated.

At first glance, the previous derivation might look like "actuarial," although it is not, actually. The reason is that the probabilities in Eq. (14.28) are *risk-neutral* probabilities. As such, they are, obviously, the same as those we use to price the bonds underlying the CDS contract. Therefore, there are are no-arb relations that link bond prices to CDS premiums, which shall be emphasized later on (see Section 14.3.5.4). This point illustrates in a remarkable way one key difference between finance and insurance. Even if in insurance, one may end up pricing some products through risk-adjusted probabilities, finance is where we typically end up having many more traded risks than in insurance, and these risks are tightly related through no-arb restrictions.

Eq. (14.28) is a general formula we can use, once we have a model determining the risk-neutral probability of default. In this chapter, we implement Eq. (14.28) through a reduced-form approach, which will allow us to find the quarterly premium (or spread) $\text{CDS}_t\left(M\right)$ quite easily, as follows.

We have, denoting again with λ the instantaneous probability of default, that $\Pr\{\text{Surv. at } t_i\} = e^{-\lambda(t_i - t)}$, and that $\Pr\{\text{Def.} \in (t_{i-1}, t_i)\} = e^{-\lambda(t_{i-1} - t)} - e^{-\lambda(t_i - t)}$. Intuitively, if the name survives at t_i (event E_i), it must necessarily have survived at t_{i-1} (event E_{i-1}), but the converse is not true: $E_i \subset E_{i-1}$, and the complement of E_i to E_{i-1} is simply the event of default between t_{i-1} and t_i.[11] Substituting the previous probabilities into Eq. (14.28) leaves

$$\text{CDS}_t\left(M\right) = \frac{\sum_{i=1}^{4M} e^{-r(t_i - t)} \text{LGD}\left(t_i\right) \left(e^{-\lambda(t_{i-1} - t)} - e^{-\lambda(t_i - t)}\right)}{\sum_{i=1}^{4M} e^{-(r+\lambda)(t_i - t)}}. \tag{14.29}$$

The denominator in the RHS of Eq. (14.29) is the *defaultable*-PVBP (Present Value of a Basis Point), in perfect analogy with the expressions for the forward swap rate given in Chapter

[11] Mathematically, we have that $\Pr\{\text{Default at any } z \in (t_{i-1}, t_i)\} = \int_{t_{i-1}}^{t_i} \Pr\{\text{Default at } z\} dz$, where $\Pr\{\text{Default at } z\} = \lambda e^{-\lambda(z-t)} dz$.

13. Assuming that $\text{LGD}(t_i)$ is constant and equal to LGD for each t_i, then, for a generic $\Delta t = t_i - t_{i-1}$, Eq. (14.29) can be simplified to,

$$\text{CDS}_t(M) = \text{LGD} \cdot \left(e^{\lambda \Delta t} - 1\right). \tag{14.30}$$

That is approximately, for small Δt,

$$\text{CDS}_t(M) \approx \lambda \cdot \text{LGD} \cdot \Delta t \equiv (\text{expected loss per unit of time}) \cdot \Delta t. \tag{14.31}$$

Naturally, λ is the *risk-neutral* instantaneous probability of default for the security.

Note that in this simple model, the CDS premiums for a fixed maturity are constant over time, as a result of the assumption that the intensity of default, λ and the short-term rate, r, are constant. Moreover, note that Eq. (14.31) shows that the CDS premium is approximately the same as the instantaneous spread of a defaultable bonds, as explained in Section 14.2. This property is to be expected, because a purchase of a defaultable bond and protection on it amounts to a synthetic default-free bond. Therefore, there must be a no-arbitrage relation between CDS spreads and defaultable bond spreads, as anticipated earlier.

In pratice, the approximation in (14.31) does not hold for longer maturities, due to the unrealistic assumptions underlying it (λ is constant, LGD is constant, r is constant, etc.). On the contrary, we often observe CDS spreads increasing with maturity, as further explained in Section 14.2.5.4. Indeed, we may take interesting views. For example, buying CDS for two years and selling CDS for three is a view that default will not occur between the second and the third year from now.

14.3.2.2 Marking to market

Suppose you go long a CDS, meaning that at time t, you commit to a swap agreement—you pay $\text{CDS}_t(M)$ at time t_i if the name survives by time t_i, and receive $\text{LGD}(t_i)$, if default occurs in the time interval $[t_{i-1}, t_i]$, for $4M$ time intervals. Each swap payoff—the "CDS-let" so to speak—is

$$\text{cds}_t(t_i) \equiv \text{LGD}(t_i) \cdot \mathbb{I}_{\{\text{Def.} \in (t_{i-1}, t_i) | \text{Surv. at } t_i\}} - \text{CDS}_t(M) \cdot \mathbb{I}_{\{\text{Surv. at } t_i\}}, \tag{14.32}$$

where

$$\text{CDS}_t(M) : 0 = \sum_{i=1}^{4M} e^{-r(t_i - t)} \mathbb{E}_t\left(\text{cds}_t(t_i)\right),$$

and $\mathbb{E}_t$ denotes the expectation conditional upon the information set at time t, taken under the risk-neutral probability. The solution to this equation is just that in Eq. (14.29).

What happens to the value of this contract at any subsequent time $\tau \in (t, t_0)$? The marking to market value of the CDS is the present value of the risk-neutral expectation of the single swaps payments $\text{cds}_t(t_i)$ in Eq. (14.32), consistently with explanations in Chapter 11 (see Section 11.4.6). So the marking to market value of the CDS at τ is,

$$\begin{aligned}
\text{MtM}_\tau(M) &\equiv \sum_{i=1}^{4M} e^{-r(t_i - \tau)} \mathbb{E}_\tau\left(\text{cds}(t_i)\right) \\
&= \sum_{i=1}^{4M} \left[e^{-r(t_i - \tau)} \text{LGD}(t_i) \left(e^{-\lambda(t_{i-1} - \tau)} - e^{-\lambda(t_i - \tau)}\right) - \text{CDS}_t(M)\, e^{-(r+\lambda)(t_i - \tau)}\right] \\
&= \left[\text{CDS}_\tau(M) - \text{CDS}_t(M)\right] \sum_{i=1}^{4M} e^{-(r+\lambda)(t_i - \tau)},
\end{aligned}$$

where the last line follows by the definition of $\text{CDS}_\tau(M)$, i.e. by setting $t \equiv \tau$ in Eq. (14.29). These results perfectly match those in Chapter 13 regarding forward swap rates.

Note that in this model, marking-to-market is deterministic, because CDS premiums for a fixed maturity are constant over time, due to the fact that both λ and r are constant. Marking-to-market is actually zero once we assume that loss-given-default is constant, as Eq. (14.30) shows that CDS spreads do not change in this case.

14.3.2.3 CDS on indexes, and options based thereon

A CDS index is a basket of credit entities in which the protection buyer pays the same premium on all the names in the index, until a fixed expiration date. Credit events are typically bound to bankruptcy or delinquencies. After a credit event, the entity is removed from the index and the contract goes through with a reduced notional amount, until expiration, as explained in more detail below.

While CDS on single names are over-the-counter, CDS indexes are standardized and at the time of writing give rise to relatively more liquid markets, as historical data on bid-ask spreads show. In fact, it can be cheaper to hedge a portfolio of CDS or bonds with a CDS index than it would be to buy many CDS to achieve a similar effect. There exist two main indices: (i) CDX index, which contains North American and Emerging Market companies; and (ii) iTraxx index, which contains companies from the rest of the world.

Credit default swaptions are options to enter a CDS—typically a CDS index. Consider, first, swaptions on single names. A payer swaption gives the right to buy protection at some future date at some CDS fixed strike, and a receiver swaption gives the right to sell protection. If default of the name occurs prior to the swaption maturity, the contract is terminated. Note that evaluating credit default swaptions is trivial in the pricing context of Section 14.4.3.1, because CDS premiums move deterministically over time when the intensity of default and the short-term rate are both constant. Section 14.4.3.9 hinges upon a continuous-time model of stochastic intensity rates, and supplies an evaluation framework for these products.

Credit default index swaptions work differently. Firstly, as noted, at inception, a credit default index swap (CDIS) is referenced to a number of fixed companies chosen by a market maker, each carrying a given weight. Secondly, buyers of CDIS are typically those who provide protection to market makers: they stand ready to pay a predetermined loss-given-default for any default that occurs before maturity, which is constant and identical for all reference entities in the index. In exchange, the market makers pay the CDIS buyer a periodic fixed premium—the credit default index spread. After a default takes place, the nominal value of the CDIS is reduced by one, and no replacement of the defaulted firm would take place, as further explained in Section 14.4.3.9.

14.3.2.4 Disentangling default probability from risk-aversion

The following picture, taken from Fender and Hördahl (2007), illustrates the behavior of the credit market risk appetite before the 2007 credit market turmoil.

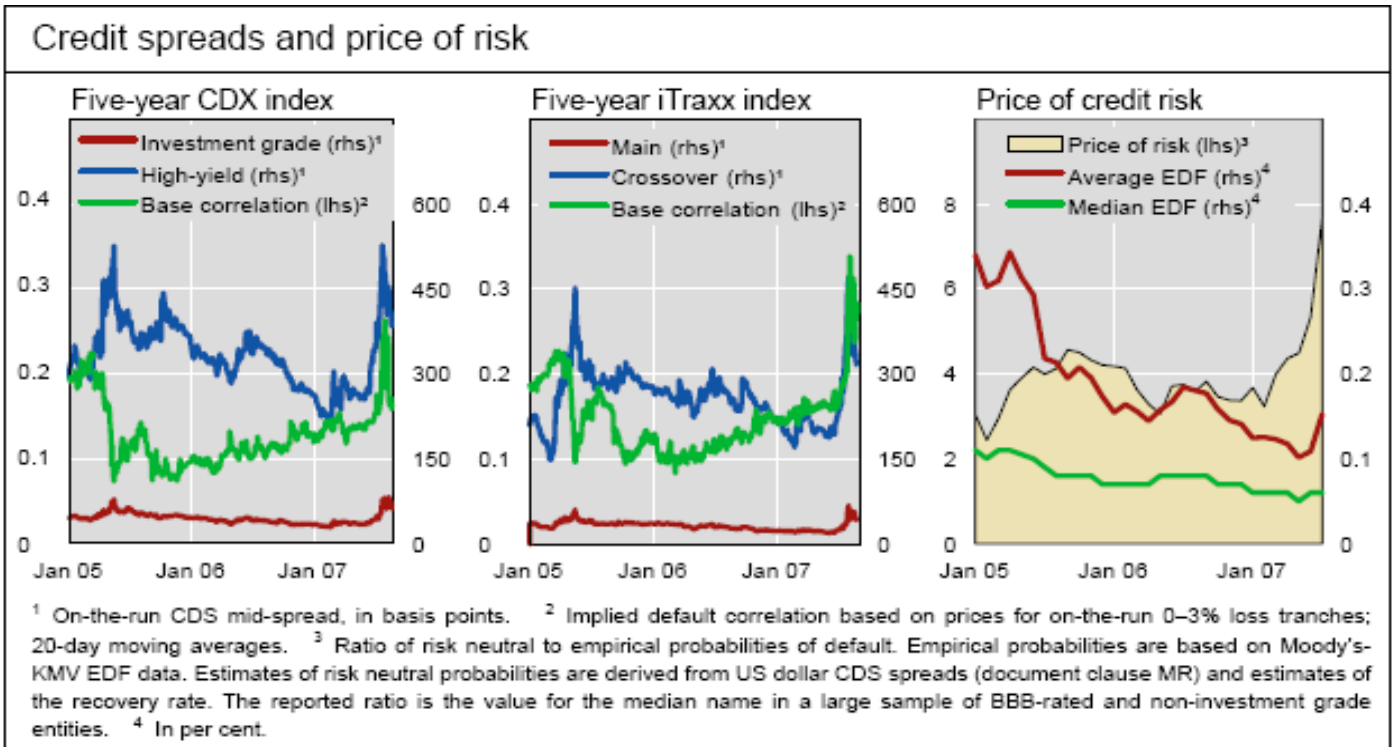


FIGURE 14.9. Antonio Mele does not claim any copyright on this picture, which is taken from Fender and Hördahl (2007). The picture has been put here for illustrative purposes only, and permission to the authors shall be duly asked before the book will be published.

How did the authors estimate the price of risk? Consider the expected losses under the actuarial, or physical probability for a given security. The counterpart to Eq. (14.31), under the physical probability, is:

$$\text{Expected Losses}^P \equiv \lambda^P \cdot \text{LGD} \cdot \Delta t,$$

where λ^P is the *physical* instantaneous probability of default for a given security. Assume that LGD is constant, to simplify. If investors require compensation for default events, the actuarial losses should be less than the CDS spread, i.e. $\text{Expected Losses}^P < \text{CDS}$, or,

$$\lambda > \lambda^P.$$

The risk-premium is defined as the difference between the actuarial losses, Expected Losses^P, and the CDS premium,

$$\text{Risk-Premium} = \left(\lambda - \lambda^P\right) \cdot \text{LGD} \cdot \Delta t.$$

The price of risk in Figure 14.9 is defined as the ratio of the CDS spread over Expected Losses^P,

$$\text{Price-of-Risk} = \frac{\lambda}{\lambda^P}.$$

Early references to estimation methods are Duffie et al. (2005) and Amato (2005). Typically, Expected Losses^P are proxied by Moody's KMV's Expected Default Frequencies (EDFs$^{\text{TM}}$), obtained through fully specified structural models for credit risk. The next pictures are taken

from Amato (2005). As we can see, during the 2003-2005 period, credit spreads were so low, and this in turn gave incentives to CDO issuers to look for illiquid and relatively more complex assets to put as collateral, which led to the issuance of CDO relying on ABS such as MBS, or CDO^2, explained below.

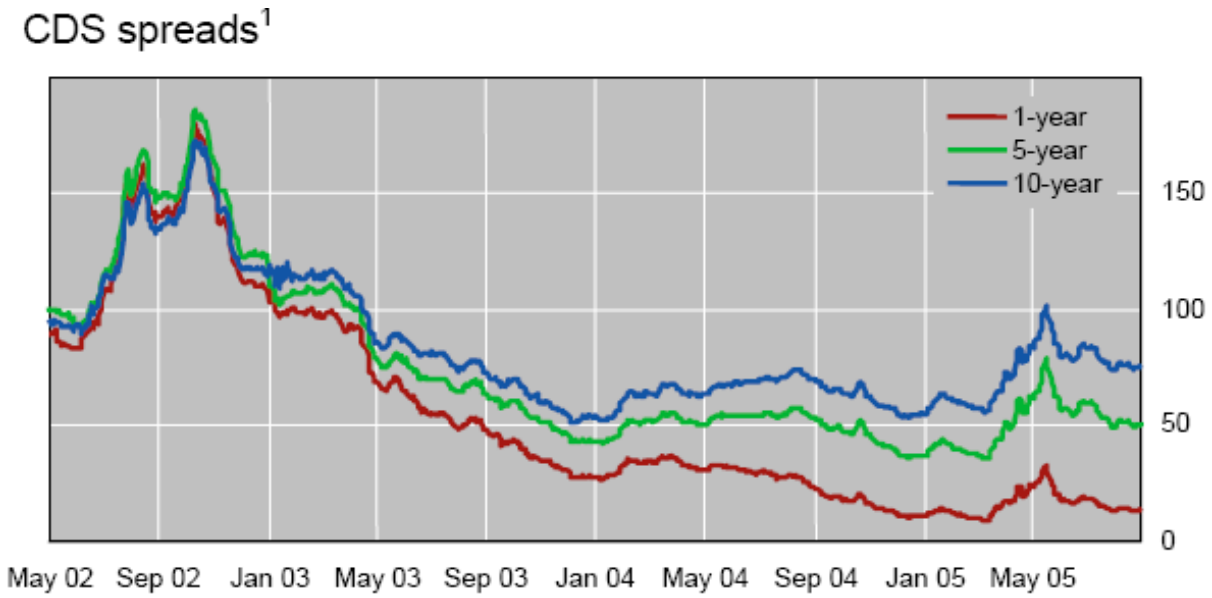


FIGURE 14.10. Antonio Mele does not claim any copyright on this picture, which is taken from Amato (2005). The picture has been put here for illustrative purposes only, and permission to the author shall be duly asked before the book will be published.

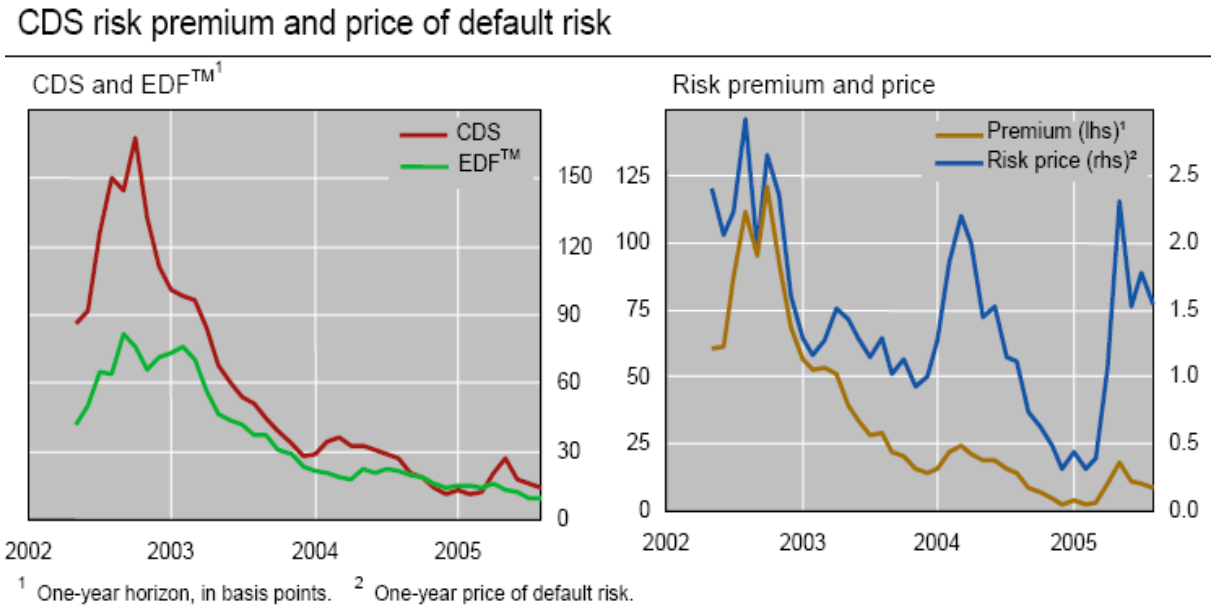


FIGURE 14.11. Antonio Mele does not claim any copyright on this picture, which is taken from Amato (2005). The picture has been put here for illustrative purposes only, and permission to the author shall be duly asked before the book will be published.

The following picture illustrates the behavior of CDS indexes during approximately 20 years before the 2007-2009 credit market turmoil.

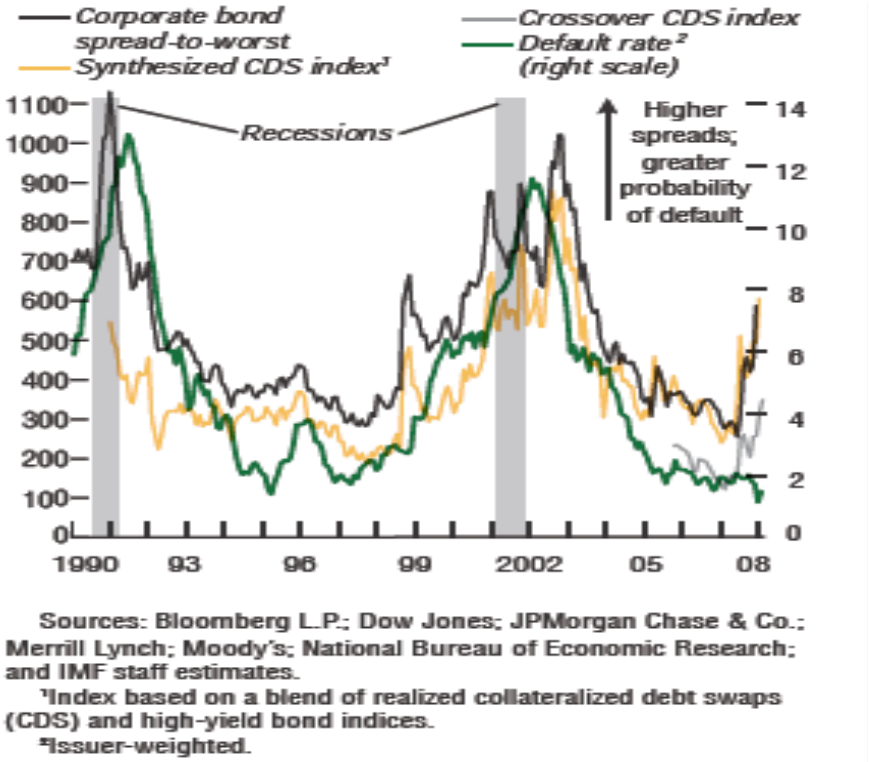


FIGURE 14.12. Valuation of Financial Instruments Based on Implied Probability of Default. Antonio Mele does not claim any copyright on this picture, which is taken from IMF (2008). The picture has been put here for illustrative purposes only, and permission to the authors shall be duly asked before the book will be published.

14.3.2.5 Continuous time

We may relax the assumption the instantaneous intensity of default, λ, is constant. This intensity is defined under the risk-neutral probability and can change either because the intensity of default under the physical probability changes or because risk-appetite changes, or both. We examine the asset pricing implications of time-varying intensities, by exploring how probabilities of survival change in a simple setting, where we do not single out the reasons leading to variations in λ.

First, we assume the instantaneous probability of default can only change discretely, giving rise to random intensities λ_t, meaning that λ_t is the intensity of default in the time interval $[t-1,t]$. Let $\mathcal{F}_t$ be the information set as of time t. We assume that λ_t is $\mathcal{F}_t$-measurable. What is the probability of survival of any given name in this case? We have, by Bayes' theorem,

$$\Pr\left\{\text{Surv at } t|\,\text{Surv at } t-1\right\}=\frac{\Pr\left\{\text{Surv at } t\right\}}{\Pr\left\{\text{Surv at } t-1\right\}}. \tag{14.33}$$

By a repeated use of Eq. (14.33),

$$\begin{aligned}\Pr\{\text{Surv at } t\} &= \Pr\{\text{Surv at } t|\,\text{Surv at } t-1\}\Pr\{\text{Surv at } t-1\}\\ &= \cdots\\ &= \prod_{n=1}^{t}\Pr\{\text{Surv at } n|\,\text{Surv at } n-1\}.\end{aligned} \tag{14.34}$$

So we are left with finding $\Pr\{\text{Surv at } n|\,\text{Surv at } n-1\}$. Consider the following arguments. If λ_n was not random and fixed at some $\bar{\lambda}_n$, then, $\Pr\{\text{Surv at } n|\,\text{Surv at } n-1\} = e^{-\bar{\lambda}_n}$. When λ_n is random, $e^{-\lambda_n}$ is the probability of survival, conditioned upon some particular value the intensity could possibly take. Heuristically, then, $\Pr\{\text{Surv at } n|\,\text{Surv at } n-1\} = \sum_{s\in\mathcal{S}} e^{-\lambda_n(s)}\Pr\{s\}$, where $\lambda_n(s)$ is, so to speak, the value λ_n would take in state s, $\Pr\{s\}$ is the likelihood that state s occurs and, finally, $\mathcal{S}$ is the set of all possible states, as illustrated by Figure 14.3.

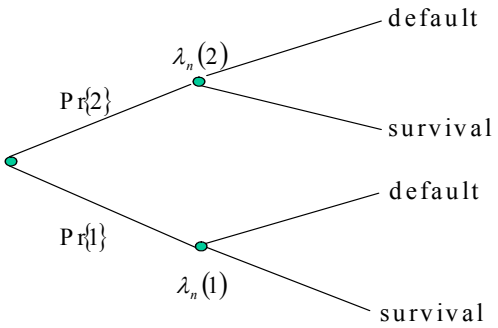


FIGURE 14.13. This picture illustrates the determination of the probability of survival in the case of random default intensities going over one period and two states. At the beginning of period n, nature draws the event defining the intensity of default, which is either $\lambda_n(1)$ with probability $\Pr\{1\}$, or $\lambda_n(2)$ with probability $\Pr\{2\} = 1-\Pr\{1\}$. Then, the two paths leading to survival have probability of occurrence equal to $\Pr\{1\}e^{-\lambda_n(1)}$ and $\Pr\{2\}e^{-\lambda_n(2)}$, such that the total probability of survival equals $\Pr\{1\}e^{-\lambda_n(1)} + \Pr\{2\}e^{-\lambda_n(2)}$.

Therefore, $\Pr\{\text{Surv at } n|\,\text{Surv at } n-1\} = \mathbb{E}\left(e^{-\lambda_n}\middle|\,\mathcal{F}_{n-1}\right)$, where $\mathbb{E}$ denotes the expectation taken under the risk-neutral probability. Inserting this result into Eq. (14.34), and using the Law of Iterated Expectations, leaves:

$$\Pr\{\text{Surv at } t\} = \mathbb{E}\left(e^{-\sum_{n=1}^{t}\lambda_n}\right).$$

Under regularity conditions, we can easily extend the previous result to a continuous time setting. For example, we may assume that the risk-neutral default intensity, λ_t, is solution to:

$$d\lambda_t = \phi\left(\bar{\lambda}-\lambda_t\right)dt + \sigma\sqrt{\lambda_t}dW_t, \quad \lambda_0 = \lambda. \tag{14.35}$$

where W is a standard Brownian motion under the risk-neutral probability, and ϕ, $\bar{\lambda}$ and σ are three positive constants. This is the same as the Cox, Ingersoll and Ross (1985) (CIR) model

of the short-term rate reviewed in Chapter 13. Therefore, under the parameter restrictions in Chapter 13, λ_t is always positive, and

$$P_{\text{surv}}(\lambda, t) \equiv \Pr\{\text{Surv at } t\} = \mathbb{E}\left(e^{-\int_0^t \lambda_s ds}\right). \tag{14.36}$$

Eq. (14.36) is, formally, the same as the Feynman-Kac representation of a solution to a partial differential equation, solved by a bond price in the Cox, Ingersoll and Ross (CIR) (1985) model of the previous chapter (see Section 13.4.3.3). In other words, the survival probability in Eqs. (14.35)-(14.36) is mathematically the same as the price of a zero coupon bond in the CIR model. Therefore, the closed-form solution for $P_{\text{surv}}(\cdot,\cdot)$ is:

$$\begin{aligned} P_{\text{surv}}(\lambda, M) &= \Phi(M)\, e^{-B(M)\lambda}, \\ \Phi(M) = \left(\frac{2\gamma e^{\frac{1}{2}(\phi+\gamma)M}}{(\phi+\gamma)(e^{\gamma M}-1)+2\gamma}\right)^{\frac{2\phi\bar{\lambda}}{\sigma^2}}, &\quad B(M) = \frac{2\left(e^{\gamma M}-1\right)}{(\phi+\gamma)(e^{\gamma M}-1)+2\gamma}, \quad \gamma = \sqrt{\phi^2+2\sigma^2}. \end{aligned} \tag{14.37}$$

More generally, we can build up a whole family of models with a closed-form solution, the affine class reviewed in Chapter 13, by assuming that:

$$\lambda_t = \lambda_0 + \lambda_1 \cdot y_t, \tag{14.38}$$

where λ_0 is a constant, λ_1 is a vector of constants, and y is a multivariate jump-diffusion process, with drift and diffusion terms as in Chapter 13 (see Section 13.4.6). This model is interesting, as we can judiciously choose the components of y_t which we suppose may affect the default intensity. For example, some of them could be unobservable, and others could be observable, and relate, say, to the business cycle or even the structure of the firm.

Given any solution for the survival probability predicted by any of these affine models when $y_0 = y$, $P_{\text{surv}}(y, t)$ say, we can easily compute

$$\Pr\{\text{Default} \in (t_{i-1}, t_i)\} = P_{\text{surv}}(y, t_{i-1}) - P_{\text{surv}}(y, t_i). \tag{14.39}$$

We can look at the bond spreads and the CDS spreads implied by this modeling choice. In Appendix 3, we show the price of a defaultable pure discount bond expiring in M years is:

$$P(y, M) = e^{-rM} P_{\text{surv}}(y, M) + \int_0^M e^{-rt} \Pr\{\text{Default} \in dt\} \text{Rec}(t)\, dt, \tag{14.40}$$

where $\text{Rec}(t)$ denotes the recovery value in case of default, supposed to be known. Naturally, these derivations are perfectly consistent with similar derivations in Chapter 13 (see Section 13.4.7), although we are now emphasizing the aspects regardind survival probabilities.

As for the forward CDS spreads, we have, by Eq. (14.28),

$$\text{CDS}_t(M) = \frac{\sum_{i=1}^{4M} e^{-r(t_i-t)} \text{LGD}(t_i)\left[P_{\text{surv}}(y, t_{i-1}) - P_{\text{surv}}(y, t_i)\right]}{\sum_{i=1}^{4M} e^{-r(t_i-t)} P_{\text{surv}}(y, t_i)},$$

where M is, again, the number of years the CDS refers to, and $t_i = t + \frac{i}{4}$.

Assume the short-term rate, r, is zero, and that loss-given-default is constant and equal to LGD. Then, as shown in Appendix 3, the price of a defaultable pure discount bond, $P(\lambda, M)$, and the CDS premium, obtained from the forward once we set $t = 0$, $\text{CDS}_0(M)$, are given by:

$$P(\lambda, M) = 1 - \text{LGD} \cdot (1 - P_{\text{surv}}(\lambda, M)), \quad \text{CDS}_0(M) = \text{LGD} \cdot \frac{1 - P_{\text{surv}}(\lambda, M)}{\sum_{i=1}^{4M} P_{\text{surv}}(\lambda, t_i)}. \tag{14.41}$$

The expression for the price is a natural extension of the basic evaluation model in Section 14.2.3.1 (see Eq. (14.26)). The model, now, predicts time variation in credit spreads for both bonds and CDS premiums. Figure 14.14 depicts the bonds spread, $\frac{-1}{M}\ln P(\lambda, M)$, and the annualized credit default spreads, $4 \times \text{CDS}_0(M)$, with parameters in Eq. (14.35) fixed at $\phi = 0.30$, $\bar{\lambda} = 0.04$ and $\sigma = \sqrt{\frac{1}{2}\bar{\lambda}}$, and LGD $= 0.60$, and two values of the current intensity: $\lambda = \bar{\lambda} = 0.04$, and $\lambda = 0.02$. Assuming LGD is constant is not plausible, empirically. Instead, we know LGD moves procyclically, although it does not exhibit strong business cycle features, for sovereigns. Regarding sovereigns, the size of the country and debt distribution seem to be by far more important.

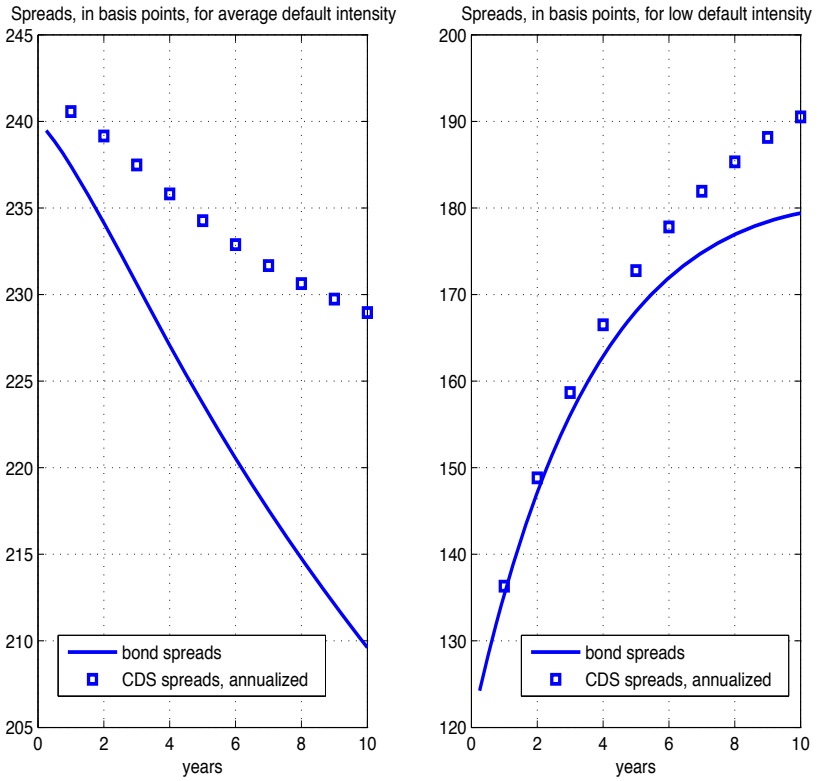


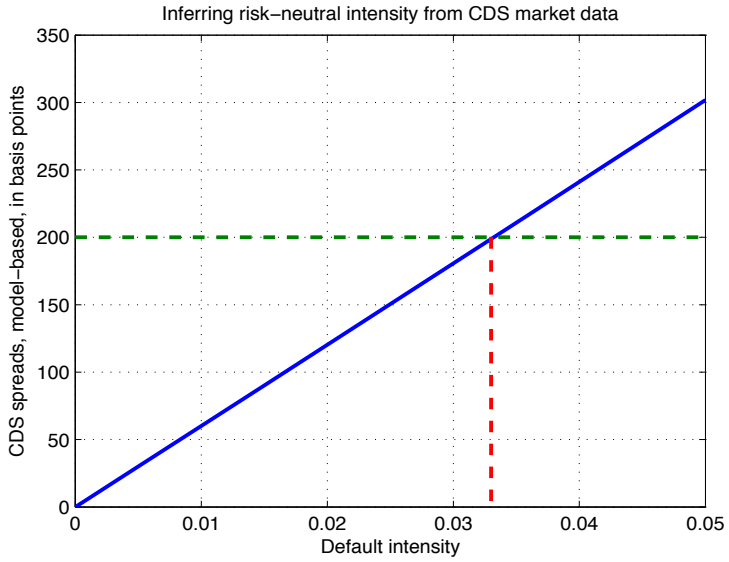
FIGURE 14.14. Spreads on bonds and CDS predicted by the affine model in Eq. (14.35). The left panel depicts the spreads when the current default intensity equals the long-run mean, $\lambda = \bar{\lambda} = 0.04$. The right panel depicts the spreads in good times, i.e., when the current intensity of default takes a low value, $\lambda = 0.02$. In each case the recovery rate equals 40%.

The mechanism is that given the mean-reverting behavior of λ, good times are likely followed by bad, such that when $\lambda = 0.02$, we expect default rates to rise in the future. Spreads increase with maturity as a result. Moreover, bond spreads are approximately equal to CDS spreads at short maturities. At longer maturities, the two spreads diverge, with CDS spreads, $4 \times \text{CDS}_0(M)$, dominating bonds spreads, $\frac{-1}{M}\ln P(\lambda, M)$. Moreover, the two curves are decreasing in time to maturity even when the current value of the intensity equals the long-run one, $\bar{\lambda}$. This property is due to the assumption that recovery rates are constant, as explained in the

constant intensity case dealt with in Section 14.2.3.2. In Appendix 5, we provide additional details and explanations regarding these properties.

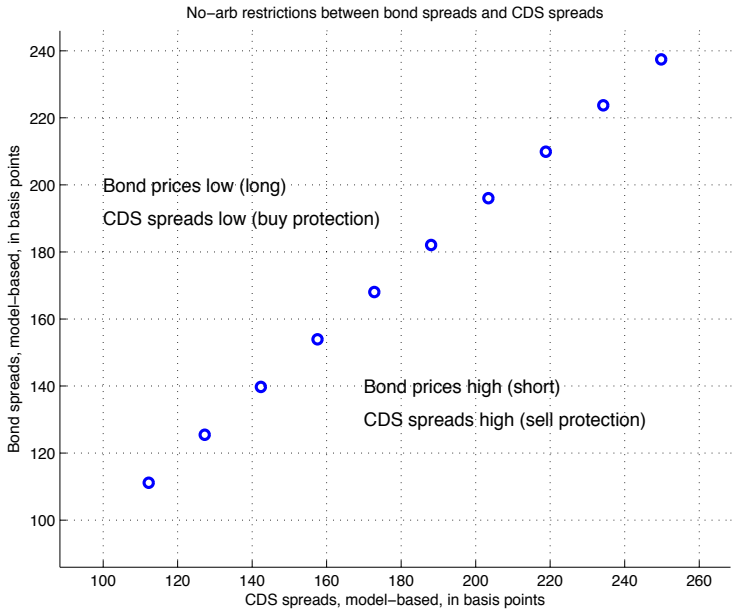
14.3.2.6 A trading strategy

Bond prices and CDS spreads are driven by the same state variable, the default intensity, and so they are restricted to lie on some space, to be consistent with no-arbitrage. To illustrate, consider, first, the simple case where the default intensity is constant, such that CDS spreads are given by Eq. (14.29). Given this model, we can look at the market data for CDS spreads, and infer the risk-neutral intensity, as in the picture below.



In this picture, the CDS spreads predicted by Eq. (14.29) are depicted as a function of the risk-neutral intensity, λ, assuming $M = 5$ years, LGD $= 0.60$ and the short-term rate r is zero. For example, if we had to observe a CDS premium equal to 200 basis points, we would infer a risk-neutral intensity approximately equal to $\hat{\lambda} = 0.033$. The key point is this same $\hat{\lambda}$ should also be pricing the zero, such that for $M = 5$, $P(M) = 1 - \text{LGD} \cdot (1 - e^{-\hat{\lambda} M}) = 0.90874$, and so we go long both the bond and the CDS if the bond market price is less than 0.90874. The rationale behind this strategy is that the intensity implied by the current CDS premium is too low to justify the bond price, so bond prices and/or intensity (and, hence, CDS premiums) should increase. The exactly opposite reasoning applies when the bond price is higher than 0.90874.

The previous example relies on a constant default intensity; however, the same strategy applies when default intensities are stochastic. The picture below shows the no-arbitrage restrictions between bond spreads and CDS spreads, obtained with the same parameter values as those in Figure 14.14, and current intensity values ranging from 0.0050 to 0.05. It also provides indications of a strategy aiming to exploit deviations from theoretical parity.



14.3.2.7 Hazard rates

In a pricing context, the relevant probabilities of survival are obviously conditioned upon the time of evaluation, time 0 say. For example, the probability of default in Eq. (14.39) is only conditioned to the information we have at time zero. More generally, the probability of defaulting in the time interval (t_{i-1}, t_i), conditional upon survival at time $t < t_{i-1}$, is:

$$\Pr\{\text{Def.} \in (t_{i-1}, t_i) | \text{Surv. at } t\} = \frac{P_{\text{surv}}(y, t_{i-1}) - P_{\text{surv}}(y, t_i)}{P_{\text{surv}}(y, t)}. \tag{14.42}$$

For example, for $t = t_{i-1}$, and (t_{i-1}, t_i) small, and λ deterministic, consider the following approximation to this conditional probability:

$$\begin{aligned}\Pr\{\text{Def.} \in (t_{i-1}, t_i) | \text{Surv. at } t\} &\approx \frac{\frac{\partial}{\partial t} P_{\text{surv}}(y, t)}{P_{\text{surv}}(y, t)} (t_i - t_{i-1}) \\ &\equiv \frac{p_{\text{def}}(y, t)}{1 - P_{\text{def}}(y, t)} (t_i - t_{i-1}) \\ &= \lambda_t (t_i - t_{i-1}),\end{aligned}$$

with straight forward notation. The previous expressions are known as *hazard rates*. They coincide with $\lambda_t dt$, when λ_t is deterministic. If λ_t is not deterministic, simple computations lead to:

$$\Pr\{\text{Def.} \in (t, t + dt) | \text{Surv. at } t\} = \mathbb{E}_{Q_\lambda}(\lambda_t)\, dt, \tag{14.43}$$

where Q_λ is a new probability, with Radon-Nikodym derivative given by:

$$\left.\frac{dQ_\lambda}{dQ}\right|_{\mathcal{F}_t} = \frac{e^{-\int_0^t \lambda_s ds}}{P_{\text{surv}}(\lambda, t)}. \tag{14.44}$$

Accordingly, under Q_λ, the state variables in Eq. (14.38) follow a diffusion process, with a drift process tilted, due to this change of probability. For example, in the simple setting of Eq. (14.35), we have that, for a fixed t,

$$\begin{aligned} &d\lambda_s = (\mathcal{B}_0 - \mathcal{B}_1^t(s)\,\lambda_s)\,ds + \sigma\sqrt{\lambda_s}dW_{\lambda,s}, \quad s \in (0,t], \quad \lambda_0 = \lambda, \\ &\mathcal{B}_0 = \phi\bar{\lambda}, \quad \mathcal{B}_1^t(s) = \phi + B(t-s)\,\sigma^2, \quad B(\cdot) \text{ as in Eq. (14.37)}, \end{aligned} \tag{14.45}$$

where W_λ is a Brownian motion under Q_λ. Therefore, by Eq. (14.43), and computations,

$$\Pr\{\text{Def.} \in (t, t+dt) | \text{Surv. at } t\} = \left(\frac{\lambda}{G(t)} + \mathcal{B}_0 \int_0^t \frac{G(s)}{G(t)} ds\right) dt, \; G(x) \equiv e^{\int_0^x \mathcal{B}_1^t(u)du}.$$

Appendix 5 provides a proof of these results, which to the best of our knowledge, are developed here for the first time.

14.3.2.8 Extracting probabilities of default from market data

Markets obviously convey information about default probabilities, which could be extracted under a number of assumptions. To illustrate, assume zero recovery and that both the short-term rate and the default intensity are continuous-time Markov and independent of each other. Then, the price of a defaultable zero is $P_{\text{def}}(\lambda, M) = P(M) \cdot P_{\text{surv}}(\lambda, M)$, where $P_{\text{def}}(\lambda, M)$ and $P(M)$ are the prices of a defaultable and a non-defaultable zero. Therefore, we can read the *risk-neutral* probability of survival from the defaultable/non-defaultable price ratio:

$$P_{\text{surv}}(\lambda, M) = \frac{P_{\text{def}}(\lambda, M)}{P(M)}. \tag{14.46}$$

Naturally, surviving until some M_2 implies having survived until any $M_1 : M_1 < M_2$ and, from M_1 to M_2. Therefore, $P_{\text{surv}}(\lambda, M_2) = P_{\text{surv}}(\lambda, M_1) \cdot P_{\text{surv}}(\lambda, M_1, M_2)$, where $P_{\text{surv}}(\lambda, M_1, M_2)$ is the risk-neutral probability of survival between M_1 and M_2. Using Eq. (14.46), then, we can extract this probability, as follows:

$$P_{\text{surv}}(\lambda, M_1, M_2) = \frac{P_{\text{def}}(\lambda, M_2)}{P_{\text{def}}(\lambda, M_1)} \frac{P(M_1)}{P(M_2)}.$$

The previous example relies on the simplifying assumption of a zero recovery rate, but it can be generalized to the case where the recovery rate is nonzero. However, in this case, an identification issue arises, as prices would convey information about both default probabilities and recovery rates.

14.3.2.9 Pricing credit default swaptions

Swaptions on single names

With stochastic intensity rates, we can think about the pricing of the credit default swaptions that we briefly mentioned in Section 14.3.2.3. We, now, actually assume that both default intensities and the short-term rate are stochastic: we assume that the short-term rate is a a diffusion process, and that default arrives as a Cox process with intensity λ adapted to r.[12]

[12]Section 14.4.2.5 shows how to generalize Eq. (14.28) to allow for stochastic intensities. It is easy to generalize further while allowing for stochastic interest rates.

Consider the following definition of a *default swap* in Section 14.3.2.1—a contract whereby a party stands ready to pay his counterparty a loss determined by a credit event for a given flows of premiums, referred to as CDS premiums. At this level, we are not assuming that this swap is worthless at origination. We assume that loss-given-default is constant and equal to LGD. Denote the CDS premium, or "coupon," agreed at time t with C_t, $t \leq t_0$. Note that like in Section 14.3.2.3, the contract we are dealing with is a *forward* default swap. We assume that the contract is terminated should the underlying obligor default prior to the start date, t_0.

In this contract, the protection buyer commits to a swap agreement whereby it pays the coupon C_t at time t_i, if the name survives by time t_i, and receives LGD if default occurs at some point in time $\tau^* \in [t_0, t_{4M}]$. Therefore, the value of the default swap originated at time t is

$$\mathrm{DS}_t = \sum_{i=1}^{4M} \mathbb{E}_t \left(e^{-\int_t^{t_i} r_\tau d\tau} \mathrm{cds}_t \left(t_i \right) \right) = \mathrm{LGD} \cdot V_{0t} - C_t \cdot V_{1t}, \tag{14.47}$$

where $\mathbb{E}_t$ denotes the risk-neutral expectation, taken conditional upon the information set at time t, and

$$V_{0t} \equiv \mathbb{E}_t \left(e^{-\int_t^{\tau_*} r_\tau d\tau} \cdot \mathbb{I}_{\{\tau^* \in [t_0, t_{4M}]\}} \right), \quad V_{1t} \equiv \sum_{i=1}^{4M} \mathbb{E}_t \left(e^{-\int_t^{t_i} r_\tau d\tau} \cdot \mathbb{I}_{\{\text{Surv. at } t_i\}} \right). \tag{14.48}$$

The interpretation of V_{0t} is that of the value of one dollar paid off the first time after default, provided default occurs prior to the maturity of the default swap, $4M$.[13] Instead, V_{1t} is the value of an annuity of one dollar paid at the dates t_1, $t_2, \cdots, t_{4M}$, until default or maturity of the default swap, whichever occurs first. In other words, V_{1t} is the value of a basket of defaultable bonds with zero recovery value—a *defaultable present value of the basis point*.

The forward default spread is the value of C_τ such that $\mathrm{DS}_\tau = 0$. It is:

$$\mathrm{CDS}_\tau \left(M \right) = \mathrm{LGD} \cdot \frac{V_{0\tau}}{V_{1\tau}}, \tag{14.49}$$

such that the value of the default swap at time τ, agreed at time $t \leq \tau$, and denoted as $\mathrm{DS}_\tau \left(t, T \right)$, can be expressed as:

$$\mathrm{DS}_\tau \left(t, T \right) = \mathrm{LGD} \cdot V_{0\tau} - C_t \cdot V_{1\tau} = V_{1\tau} \cdot \left(\mathrm{CDS}_\tau \left(M \right) - C_t \right). \tag{14.50}$$

Note that the derivation leading to Eq. (14.50) generalizes that underlying the marks-to-market updates in Section 14.3.2.2.

Next, note that for any $\mathbb{F}_T$-measurable random variable X_T adapted to $(r_\tau)_{\tau \in [t,T]}$ and satisfying enough regularity conditions, we have that for fixed T, and by the Law of Iterated

[13] Eq. (14.48) does not generalize Eq. (14.28) in Section 14.4.2, where it was assumed that premiums and protections are both paid at the end of the period. In the present section, instead, we are assuming that premiums are paid as soon as default occurs, as revealed by the expression for V_{0t} in Eq. (14.48).

Expectations,

$$\begin{aligned}
&\mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau}\cdot X_T\cdot \mathbb{I}_{\{\text{Surv. at } T\}}\right)\\
&=\mathbb{E}_t\left[\mathbb{E}\left(e^{-\int_t^T r_\tau d\tau}\cdot X_T\cdot \mathbb{I}_{\{\text{Surv. at } T\}}\middle|\,\mathbb{F}_T^r\right)\right]\\
&=\mathbb{E}_t\left[e^{-\int_t^T r_\tau d\tau}\cdot X_T\cdot \mathbb{E}\left(\mathbb{I}_{\{\text{Surv. at } T\}}\middle|\,\mathbb{F}_T^r\right)\right]\\
&=\mathbb{E}_t\left(e^{-\int_t^T (r_\tau+\lambda_\tau) d\tau}\cdot X_T\right),
\end{aligned}\tag{14.51}$$

where $\mathbb{F}_T^r$ is the information set at time T, which includes the path of the short-term rate only.

Define the probability Q_{sc} through the Radon-Nikodym derivative

$$\left.\frac{dQ_{\text{sc}}}{dQ}\right|_{\mathbb{F}_T}=e^{-\int_t^T (r_\tau+\lambda_\tau)d\tau}\frac{V_{1T}}{V_{1t}},\tag{14.52}$$

where Q is the risk-neutral probability. We shall argue in a moment that Q_{sc} does integrate to one.[14] Following Schönbucher (2003, p. 180), we refer to Q_{sc} as the *survival contingent probability*. Chapter 5 provides foundations on changes of probability, with this probability being a special case of a general framework.

We claim that for any $\tau\leq t_0$,

$$V_{it}=\mathbb{E}_t\left(e^{-\int_t^\tau (r_s+\lambda_s)ds}V_{i\tau}\right),\quad i=1,2,\tag{14.53}$$

implying that Q_{sc} integrates to one. Indeed, ote that V_{it} is the value of a basket of securities paying off £1 contingent upon the event that the firm has not defaulted prior to time τ, $\tau\leq t_0$, in which case the value drops to zero. Following derivations in Chapter 13, we have that $LV_i-rV_i+\lambda(0-V_i)=0$, where L is the infinitesimal generator for diffusions, whence Eq. (14.53). Alternatively, Eq. (14.53) follows by Eq. (14.51).

Therefore, given the definition of Q_{sc} in Eq. (14.52) and the martingale property of V_{0t} in Eq. (14.53), we have that the forward default spread in Eq. (14.49) is a martingale under the survival contingent probability:

$$\begin{aligned}
&\mathbb{E}_t^{\text{sc}}\left(\text{CDS}_\tau(M)\right)\\
&=\mathbb{E}_t\left(e^{-\int_t^\tau (r_s+\lambda_s)ds}\frac{V_{1\tau}}{V_{1t}}\text{CDS}_\tau(M)\right)=\mathbb{E}_t\left(e^{-\int_t^\tau (r_s+\lambda_s)ds}\frac{V_{0\tau}}{V_{1t}}\text{LGD}\right)=\text{CDS}_t(M),
\end{aligned}$$

where $\mathbb{E}_t^{\text{sc}}$ denotes the time-t conditional expectation taken under the survival contingent probability, and where we have used the pricing equation (14.53) and the definition of $\text{CDS}_t(M)$ in Eq. (14.49).

We evaluate default swaptions by relying on the survival contingent probability. By Eq. (14.50), the payoff of a swaption payer is,

$$\text{DS}_T(t,T)=\mathbb{I}_{\{\text{Surv. at } T\}}\cdot V_{1T}\cdot\left(\text{CDS}_T(M)-K\right)^+,\quad T=t_0,$$

[14] We could complete the arguments while relying on Schönbucher (2003, Chapter 7), Lando (2004, Chapter 5) and Chapter 13 of these lectures (see below).

for a strike K. Relying on the property in Eq. (14.51) leaves:

$$\begin{aligned}&\mathbb{E}_t\left(e^{-\int_t^T r_\tau d\tau}\mathrm{DS}_T\left(t,T\right)\right)\\&=\mathbb{E}_t\left(e^{-\int_t^T (r_\tau+\lambda_\tau)d\tau}V_{1T}\cdot\left(\mathrm{CDS}_T\left(M\right)-K\right)^+\right)=V_{1t}\cdot\mathbb{E}_t^{\mathrm{sc}}\left(\mathrm{CDS}_T\left(M\right)-K\right)^+.\end{aligned}$$

We know that $\mathrm{CDS}_t\left(M\right)$ is a martingale under the survival contingent probability. Let W_τ^{sc} be a Brownian motion under Q^{sc}. Assume that

$$\frac{d\mathrm{CDS}_\tau\left(M\right)}{\mathrm{CDS}_\tau\left(M\right)}=\sigma dW_\tau^{\mathrm{sc}},\quad \tau\in\left[t,T\right],$$

where σ is the volatility parameter, a constant. We can apply Black (1976) to obtain evaluation formulae in this environment.

Swaptions on CDS indexes

Consider, first, a *CDS index*, as succinctly described in Section 14.3.2.3. Let n be the number of names in the index decided at time t. Each name has notional value equal to $\frac{1}{n}$, the same loss-given-default LGD and the same default intensity λ.

When any name j defaults at time τ_j, the index experiences a loss at τ_j equal to $\mathrm{LGD}\frac{1}{n}$, so in general $\mathrm{LGD}\frac{1}{n}\mathbb{I}_{\{t\leq\tau_j\leq T_{4M}\}}$ is the index loss generated by name j from t to the index expiry, which the protection seller stands ready to pay. The protection buyer pays a periodic premium, at t_i say, equal to a constant coupon C_t times N_{t_i}, the outstanding notional at t_i,

$$N_{t_i}\equiv 1-\frac{1}{n}\sum_{h=0}^{i}\mathcal{D}\left(t_{h-1},t_h\right)=\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}_{\{\mathrm{Surv}_j\text{ at }t_i\}},\quad N_t\equiv 1,\tag{14.54}$$

where $\mathcal{D}\left(t_{i-1},t_i\right)$ denotes the number of names having defaulted in the the time interval (t_{i-1},t_i),

$$\mathcal{D}\left(t_{i-1},t_i\right)\equiv\sum_{j=1}^{n}\mathbb{I}_{\{\mathrm{Def}_j\in(t_{i-1},t_i)\}},$$

and $\mathcal{D}\left(t_{-1},t_0\right)$ is the number of defaults occurred in (t,t_0).

Given the assumptions that obligors are all the same ex ante, the value at t of the protection leg minus that of the premium leg over the life of the index is the same as in the single name case (see Eq. (14.47)):

$$\begin{aligned}\mathrm{DSX}_t&=\mathbb{E}_t\left(\sum_{j=1}^{n}e^{-\int_t^{\tau_j}r_\tau d\tau}\mathrm{LGD}\frac{1}{n}\mathbb{I}_{\{t\leq\tau_j\leq T_{4M}\}}\right)-\mathbb{E}_t\left(\sum_{i=1}^{4M}e^{-\int_t^{t_i}r_\tau d\tau}C_tN_{t_i}\right)\\&=\mathrm{LGD}\cdot V_{0t}-C_t\cdot V_{1t},\quad\text{for } t\equiv t_0,\end{aligned}\tag{14.55}$$

where V_{0t} and V_{1t} are as in Eqs. (14.48), and can interpreted as the values of securities indexed on default events of an hypothetical representative name.

A CDS index at the time of origination $t\equiv t_0$ is, then, simply, the value of C_{t_0} in Eq. (14.55), which makes $\mathrm{DSX}_{t_0}=0$, viz

$$C_{t_0}\left(M\right)=\mathrm{LGD}\cdot\frac{V_{0,t_0}}{V_{1,t_0}}.$$

Next, consider a *forward starting credit default index*, which is an index starting at time t_0 and decided at some point $t \leq t_0$. Clearly, the value of the protection leg minus that of the premium leg is the same as that in Eq. (14.55), for a generic $t \leq t_0$. Moreover, the value of the forward starting credit default index decided at t is, for any time $\tau \in (t, t_0)$,

$$\mathrm{DSX}_\tau(t, t_0) \equiv N_\tau \cdot (\mathrm{LGD} \cdot V_{0\tau} - C_t \cdot V_{1\tau}), \tag{14.56}$$

where N_τ denotes the outstanding notion in Eq. (14.54) (see Appendix 7).

Finally, an *index default swaption payer* with strike K, gives the holder the option to enter a CDS index as a protection buyer with an index strike spread equal to K. Upon exercise, the protection buyer would also receive a *front-end protection*, defined as the losses occurring from the option origination to the exercise date. Let t be the option origination and $T = t_0$ the maturity of the swaption. The front end protection is, $F_T = \mathrm{LGD}\frac{1}{n}\mathcal{D}(t, T)$. In Appendix 7, we show that the value of the front-end protection is

$$V_\tau^{\mathrm{F}} = \mathbb{E}_\tau\left(e^{-\int_\tau^T r_u du} F_T\right) = \mathrm{LGD}\left(\frac{1}{n}\mathcal{D}(t,\tau) P(\tau, T) + N_\tau (P(\tau, T) - P_{\mathrm{def}}(\tau, T))\right), \tag{14.57}$$

where $P(t, T)$ and $P_{\mathrm{def}}(t, T)$ denote the price of a non-defaultable and a defaultable zero expiring at time T, with zero recovery value and default intensity equal to that of the representative firm, λ. The underlying of a default swaption payer equals $\mathrm{DSX}_T(t, T) + F_T$. Accordingly, we can define the *loss-adjusted forward default swap index*, as $\mathrm{DSX}_\tau^L(t, T) \equiv \mathrm{DSX}_\tau(t, T) + V_\tau^{\mathrm{F}}$, and find the value of $\overline{\mathrm{CDS}}_\tau(M)$ such that newly issued forwards are worthless, $\mathrm{DSX}_\tau^L(\tau, T) = 0$, denoted as $\mathrm{CDX}_\tau(M)$, which is

$$\mathrm{CDX}_\tau(M) = \mathrm{LGD}\frac{V_{0\tau}}{V_{1\tau}} + \frac{V_\tau^{\mathrm{F}}}{N_\tau V_{1\tau}}, \tag{14.58}$$

such that

$$\mathrm{DSX}_\tau^L(t, T) = N_\tau V_{1\tau}(\mathrm{CDX}_\tau(M) - C_t).$$

We wish to use $N_\tau V_{1\tau}$ as a numéraire such that $\mathrm{CDX}_t(M)$ is a martingale under a suitable probability,[15] similarly as for the probability Q_{sc} in Eq. (14.52). Define the probability $\hat{Q}_{\mathrm{sc}}$ through the Radon-Nikodym derivative:

$$\left.\frac{d\hat{Q}_{\mathrm{sc}}}{dQ}\right|_{\mathbb{F}_T} = e^{-\int_\tau^T r_u du}\frac{N_T V_{1T}}{N_\tau V_{1\tau}}. \tag{14.59}$$

The probability $\hat{Q}_{\mathrm{sc}}$ is the index counterpart to Q_{sc} in Eq. (14.52). For simplicity, we shall keep on referring to $\hat{Q}_{\mathrm{sc}}$ as the "survival contingent probability." Appendix 7 contains a proof that $\hat{Q}_{\mathrm{sc}}$ does indeed integrate to one. It also shows that $\mathrm{CDX}_t(M)$ in Eq. (14.58) is a martingale under $\hat{Q}_{\mathrm{sc}}$. The price of a swaption payer with strike K is, for any $\tau \in [t, T]$,

$$\mathrm{SW}_\tau^p(K, T) \equiv \mathbb{E}_\tau\left(e^{-\int_\tau^T r_u du} N_T V_{1T}(\mathrm{CDX}_T(M) - K)^+\right) = N_\tau V_{1\tau} \cdot \hat{\mathbb{E}}_\tau^{\mathrm{sc}}(\mathrm{CDX}_T(M) - K)^+,$$

[15] A technical issue with the definition of $\mathrm{CDX}_t(M)$ in Eq. (14.58) relates to a "denominator problem"—the possibility of a total collapse of the index, $N_\tau = 0$. The occurrence of such an event has been taken into account by Rutkowski and Armstrong (2009) and Morini and Brigo (2011).

where $\hat{\mathbb{E}}_{\tau}^{\text{sc}}(\cdot)$ denotes the time τ conditional expectation under the survival contingent probability $\hat{Q}_{\text{sc}}$ in Eq. (14.59). We know $\text{CDX}_{\tau}(M)$ is a martingale under $\hat{Q}_{\text{sc}}$. We can use Black (1976) to evaluate the previous expression, once we assume that under $\hat{Q}_{\text{sc}}$, $\text{CDX}_{\tau}(M)$ is a geometric Brownian motion with constant volatility.

[Explain the post Big-Bang corrections]

14.3.3 Collateralized Debt Obligations (CDOs)

14.3.3.1 A crash description of securitization

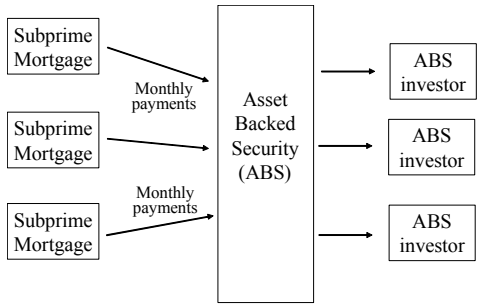
The US government has perhaps provided one of the historically most important inputs to securitization through a series of financial innovation initiatives introduced during the 1980s. To date, and until the 1970s, the financial system is in a "buy to hold" regime, whereby financial intermediaries (such as commercial banks) would make loans to businesses or individuals, which they would typically hold in their balance sheets. During the 1970s, another trend begins, whereby government sponsored enterprises such as the Government National Mortgage Association ("GNMA" or "Ginnie Mae") would buy the mortgages from banks to incentivize them to extend more loans, with the purpose of making houses accessible to families. In a second step, GNMA would then sell securities based on the cash flows generated by these mortgages.

The Federal National Mortgage Association ("Fannie Mae") and the Federal Home Loan Mortgage Corporation ("Freddie Mac") would subsequently take securitization to a higher level while securitizing the assets through "tranching." Once the tranching model was initially developed, investment banks would apply this same idea to other kinds of assets, such as corporate bonds, student loans, small business loans, automobile loans, etc.

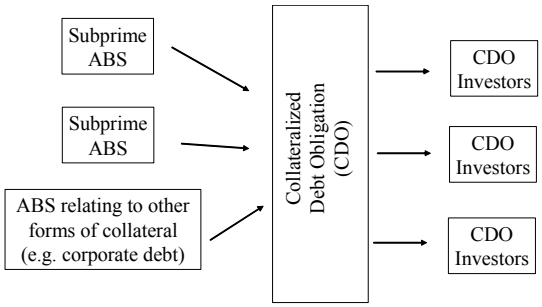
How does tranching work? Tranching relies on CDOs, which are securitized shares in pools of assets. Collateral assets include loans or debt instruments. A CDO may be a collateralized loan obligation (CLO) or collateralized bond obligation (CBO) according to whether it relies only on loans or bonds, respectively. CDO investors bear the credit risk of the collateral. Multiple tranches of securities are issued by the CDO, offering investors various maturity and credit risk characteristics. Tranches are categorized according to their degree of credit risk, and can be senior, mezzanine or junior.[16] When defaults occur, or otherwise the CDO's collateral underperforms, scheduled payments to senior tranches take precedence over those of mezzanine tranches, and scheduled payments to mezzanine tranches take precedence over those to junior tranches. Typically, senior tranches are rated, with ratings of A to AAA. Mezzanine are also rated, typically with ratings of B to BBB. In principle, these ratings should reflect both the credit quality of the collateral and the protection a given tranche is given by the tranches subordinating to it.

CDOs are part of a more complex securitization process, which could also involve the inclusion of assets of different nature. The stylized example in the diagram below illustrates this process. In a first step, subprime mortgages are securitized. In a second step, a CDO is created out of the securitized subprime mortgages and additional Asset Backed Secutities (ABS); in a third step, the structuring process involves creating seniority rules.

[16] Junior tranches are sometimes referred to as "subordinated" or "equity."



Step 1



Steps 2 and 3

Investors in CDOs *senior tranches* include banks and pension funds, which might benefit from the expertise of the asset managers, and the risk-return profiles difficult to find in the market. Investors in *junior tranches* are hedge funds searching for highly risky investment opportunities that at the same time, are quite rewarding and certainly unavailable in the market. Additional investors in junior tranches were dedicated off-balance-sheets entities such as "SIV," "conduits," and "SIV-lites," which will be reviewed in Section 14.3.7.

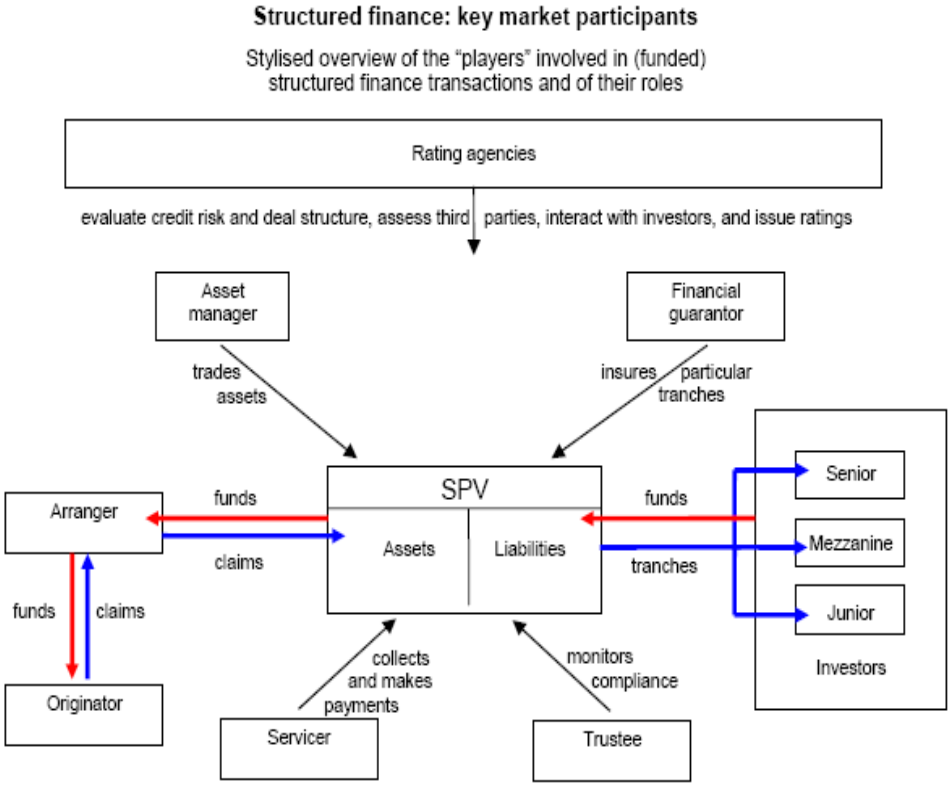
Typical CDOs underwriters are investment banks. They work closely with the asset manager and create the "right" debt/equity ratio and perform collateral quality tests. They liase with law firms and create the special purpose vehicle (possibly in some tax heaven system) that will purchase the assets and issue the tranches, price the various tranches, and obviously find the investors. Fees to underwriters are generous due to the complexity of the CDOs.[17]

Involved into the structuring process are also (i) trustee and collateral administrator, who distribute noteholder reports, check compliance and execute priority of payments; (ii) accountants, who perform due diligence on the CDOs collateral pool, verifying for example credit ratings for each asset; and (iii) rating agencies, which we shall discuss in the next subsection.

The economics behind structured finance is interesting. An originator may have private information about the quality of certain assets and/or a comparative advantage in evaluating these assets relative to other market participants. If the originator intents to sell some of its assets, an adverse selection problem arises: because investors do not know the true quality of the assets, they will demand a premium to purchase them or, even worse, a market might fail to arise (see Section 5.2 in Chapter 5 for a simplified model of this moral hazard problem).

Structured finance helps originators mitigate this problem. First, by pooling the assets, diversification benefits can be achieved. Second, tranching allows relatively poorly informed investors to access senior tranches, and be relatively protected from default. In the process, the originator or arranger may retain subordinated exposure to alleviate investors' concerns about incentive compatibility. The following scheme summarizes the structuring process.

[17] According to Thomson Financial, top underwriters in 2006 were: Bear Sterns, Merrill Lynch, Wachovia, Citigroup, Deutsche Bank, and Bank of America Securities.



Source: Committee on the Global Financial System: "The role of ratings in structured finance: issues and implications," January 2005.

14.3.3.2 The role of rating agencies

Structured finance has always been a "rated" market. Issuers of structured instruments had a natural appetite for a rating to occur at a scale comparable with that applying to debt: the main reason is that rating should facilitate the sale of these products to investors bound by ratings-based constraints defined by their investment mandates.

However, the involvement of rating agencies into the delivery of their opinion about credit risk differs from that related to traditional bonds. As regards traditional instruments, rating agencies simply aim to assess the risk of default as given, which they take as given. As regards structured finance transactions, rating agencies play a much more ex-ante, reverse engineering role. A tranche rating reflects a view about both the credit risk of the asset pool and the extent of credit support to be provided. These two elements are organized to reverse engineer the tranche rating targeted by the deal's arrangers. Deal origination involves rating agencies into the structuring process.

14.3.3.3 Types of CDOs

In practice, CDOs are considerably more complex than the stylized examples outlined earlier. We have a number of cases. We say that a CDO is *static*, if it holds the same set of assets. Instead, a CDO is *managed*, if the asset manager is allowed to change the composition of assets. If the claims to the CDO arise from the cash flows originated by the assets, we have a *cash-flow CDO*. If the claims to the CDO arise from the cash flows originated by the assets and/or active asset management, we have a *market-value CDO*. CDOs can also be created to carve out balance sheets, in which case we have *balance-sheet CDOs*. Moreover, and interestingly, CDOs

can be created (i) to achieve investment grade bonds through a pool of noninvestment grade bonds, and (ii) to create riskier securities than those in the asset pool. In these cases, we have *arbitrage CDOs*. Naturally, "arbitrage" CDOs do not give rise to any arbitrage opportunity. These instruments merely "reshuffle" risk and returns of the assets in the pool, as illustrated by the examples in the next section. Arbitrage CDOs differ from balance sheet CDOs, because issuers of arbitrage CDOs do not necessarily hold the underlying collateral in advance, which is obviously the case for issuers of balance-sheet CDOs. Therefore, the assets to be put into the an arbitrage CDO pool have to be reasonably liquid.

Furthermore, we have *synthetic CDOs*, which are exposed to a pool of assets that are not strictly owned or in the asset pool, typically through CDS underwriting. Like a cash-flow CDO, the vehicle receives payments (the premium), which is then transferred to the tranche holders. Naturally, there can be default events, which are also passed through to the investors, according to the prespecified seniority rules. A synthetic CDO is *funded*, if the relevant tranche holders are to pay for in the case of a credit event related to the assets the CDO is exposed to. Typically, some funding is made available at the very time of investment. At maturity, the investor receives a payoff equal to the funding minus the realized losses. Junior tranches are typically funded, and senior are typically not. However, senior tranches investors might have to make payments in the unlikely event losses had ever to erode their tranches.

Finally, we have *hybrid CDOs*, which are partly cash-flow CDOs and partly synthetic CDOs. In a *single-tranche CDO*, the entire CDO is structured to accommodate the specific needs of a small group of investors, with some remaining tranche held by the dealer. And we have CDO^2, where a large portion of the assets in the pool are tranches from other CDOs; or more generally, CDO^n.

14.3.3.4 Pricing

CDOs repackage cash flows from a set of assets. We provide a few simple examples to show how to price the cash flows that result from this repackaging process. In these examples, we begin describing the cash flows promised by defaultable bonds. These defaultable bonds can be thought of as being the initial illiquid assets; the resulting CDO then aims to a liquidity enhancement. However, these lectures are silent regarding the process through which the CDO becomes more liquid compared to the initial assets pools.

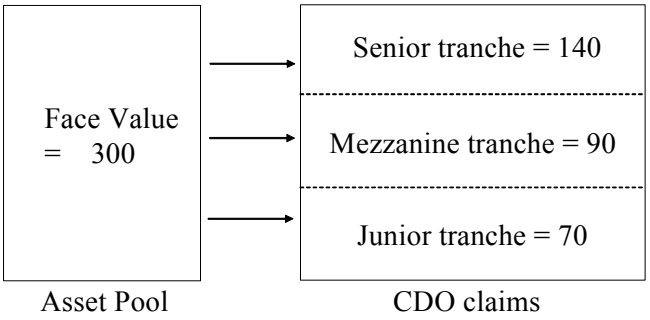
We begin with a simple example, in which we are given three one-year bonds with face value equal to 100. For each of these bonds, the *risk-neutral* probabilities of default equal 10% and the recovery rates are 40. The safe interest rate for one year is 6%. So each bond price is found by standard evaluation:[18]

$$b = e^{-0.06} \cdot (\underbrace{0.10}_{\equiv \text{Def. Prob}} \cdot 40 + \underbrace{0.90}_{\equiv \text{Surv. Prob}} \cdot 100) = 88.526,$$

such that the yield is $-\ln \frac{b}{100} = 12.19\%$.

A CDO restructures the bonds payoffs in such a way to transform the riskiness (and expected returns) of the initial asset base. Consider the following example:

[18]The numerical values of default probabilities and recovery rates in this section are taken from McDonald (2006, p. 583). We use these values as benchmarks to elaborate the additional models in this section.



In this example, each tranche receives the minimum between (i) the nominal value claimed by the tranche and (ii) what is left available to the tranche after having satisfied the other tranches by order of seniority.

Let N_i be the nominal values claimed by the tranches, so that $N_1 = 140$, $N_2 = 90$ and $N_3 = 70$. Let $\tilde{\pi}$ be the realized payoff of the asset pool, defined as,

$$\tilde{\pi} = \text{No. of Defaults} \cdot 40 \; + \; \underbrace{(3 - \text{No. of Defaults})}_{\equiv \text{No. of surviving bonds}} \cdot 100.$$

Naturally, $\tilde{\pi}$ is random because the number of defaults is random. At the expiration,

(i) the senior tranches receives the minimum between N_1 and $\tilde{\pi}$. For example, if only one bond defaults, $\tilde{\pi} = 240$, and the senior tranche receives 140. If, however, three bonds default, $\tilde{\pi} = 120$, which is less than the senior tranch nominal value, and the senior tranche then receives 120. So a quite severe loss is needed to erode the senior tranche claims.

(ii) The mezzanine tranche receives the minimum between N_2 and the "left-over" from the senior tranche.

(iii) Finally, at the expiration, the junior tranche reveives the minimum between N_3 and the "left-over" from the senior and mezzanine tranches.

In general, the payoff to i-th tranche is

$$\pi_i = \min \left\{ \text{Left-over}_{i-1}, N_i \right\},$$

where Left-over_{i-1} denotes the left-over from previous tranches, and up to the $(i-1)$-th tranche, viz

$$\text{Left-over}_{i-1} \equiv \max \left\{ \tilde{\pi} - \sum_{k=1}^{i-1} \pi_k, 0 \right\}.$$

That is,

$$\pi_i = \min \left\{ \max \left\{ \tilde{\pi} - \sum_{k=1}^{i-1} \pi_k, 0 \right\}, N_i \right\}.$$

All we need now is a model for the risk-neutral probability of default for each firm. We assume defaults follow a binomial distribution,

$$\Pr\left(\text{No. of Defaults} = k\right) = \binom{3}{k} p^k \left(1-p\right)^{n-k}, \quad p = 10\%, \quad k \in \{0,1,2,3\},$$

and, initially, we assume that the default events are independent across firms. The following table summarizes the payoffs of each tranche in each state of nature:

Payoffs to CDO tranches, and prices: with independent defaults

Defaults	Pr(Defaults)	π: pool payoff [(1)]		π_1: Senior	π_2: Mezzanine	π_3: Junior
0	0.729	300		140	90	70
1	0.243	240		140	90	10
2	0.027	180		140	40	0
3	0.001	120		120	0	0
			Price	131.8281994	83.40266709	50.34673197
			Yield	0.060142867	0.076129382	0.329561531

[(1)] π: pool payoff = Def*40+(3-Def)*100

N_1 = 140

N_2 = 90

N_3 = 70

The price of each tranche is computed as the tranche payoff, averaged across states, discounted at the safe interest rate. For example, the price of the mezzanine tranche is

$$\mathrm{P_M} = e^{-0.06}\left(0.729 \times 90 + 0.243 \times 90 + 0.027 \times 40 + 0.001 \times 0\right) = 83.403,$$

and its yield is, $\mathrm{Y_M} = -\ln \frac{83.403}{90} = 7.61\%$. Naturally, the sum of the three bond prices, $88.526 \times 3 = 265.58$, is equal to the total value of the three tranches, $131.828 + 83.403 + 50.347 = 265.58$. That is, a CDO doesn't add or destroy value: it is a mere re-packaging device, redistributing risks (and returns).

The assumption defaults among names are uncorrelated is unrealistic, as argued in Section 14.4.4. We now remove this assumption. We first analyze the special case where default events are *perfectly correlated*. In this case, either the three firms all default (with probability 0.10) or none defaults (with probability 0.90), such that the payoff matrix can be summarized by the table below:

Payoffs to CDO tranches, and prices: with perfectly correlated defaults

Defaults	Pr(Defaults)	π: pool payoff [(1)]		π_1: Senior	π_2: Mezzanine	π_3: Junior
0	0.9	300		140	90	70
1	0	NA		NA	NA	NA
2	0	NA		NA	NA	NA
3	0.1	120		120	0	0
			Price	129.9635056	76.28292722	59.33116562
			Yield	0.074388737	0.165360516	0.165360516

[(1)] π: pool payoff = Def*40+(3-Def)*100

N_1 = 140

N_2 = 90

N_3 = 70

Note that mezzanine and junior tranches now yield the same, because they each pay off either their nominal value or zero in exactly the same states of nature. In other words, default clustering implies that good times are really good, in that the probability of not having any default is now 90%, which is much higher than the 72.9% probability in the case of zero correlation.

The previous cases (with independent or perfectly correlated defaults) are extreme, but when defaults are only imperfectly correlated, pricing becomes analytically more complex, requiring a model of default correlations. We rely on a model known as Gaussian copula (reviewed in

Appendixes 9) and simulations. Figure 14.14 depicts the behavior of the yield on each tranche as a function of the default correlation in the CDO assets base.

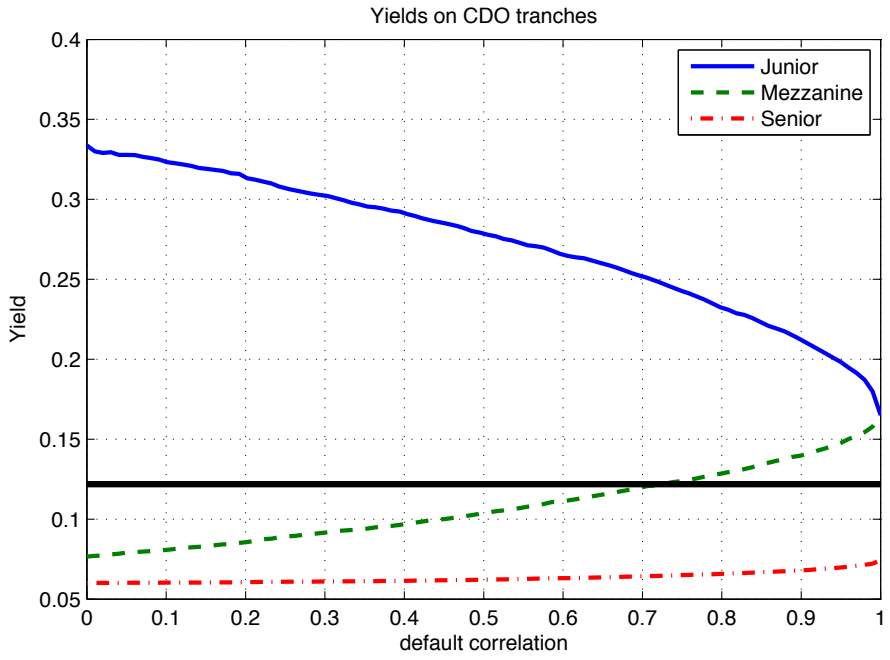


FIGURE 14.15. Yields on the three CDO tranches, as functions of the default correlation among the assets in the structure, with probability of default for each name $p = 10\%$. The thick, horizontal, line is the yield on each securitized asset.

"Arbitrage" CDOs

Figure 14.15 illustrates how arbitrage CDOs work. The CDO has three assets yielding the same, 12.19% (the horizontal line in the picture). However, by restructuring the asset base through a CDO, we can create claims (senior and mezzanine) that yield less than 12.19%, as they are considerably less risky than the asset base. The excess return, $(12.19\% - \mathrm{Y}_{\text{tranche}})$ (with $\mathrm{Y}_{\text{tranche}} \in \{\text{senior}, \text{mezzanine}\}$), is "made available" to Junior tranche holders.[19]

Note that this redistribution of risk works quite effectively when the default correlation is relatively low. As the default correlation in the asset base increases, the situation may change dramatically, with the mezzanine tranche becoming more risky and, then, yielding a higher expected return. Finally, Figure 14.16 depicts the output of a comparative statics where we increase p from 10% to 20%. The yields are obviously higher for each tranche, and the three assets now yield 18.78%, reflecting the higher marginal probability of default for each of the securities in the pool, p.

Correlation assumptions

Figures 14.15 and 14.16 show that the yield on the junior tranche decreases with default correlation. This property follows from the assumption that the probability of default is fixed at $p = 10\%$ for each default correlation ρ (say). As ρ increases, the probability of clustering events increases, which makes the senior and mezzanine tranches relatively less valuable and, hence, the Junior tranches more valuable. An alternative assumption is one for which p increases with

[19]The model does not obviously account for management fees and expenses.

ρ, which captures the idea that in bad times, both default correlation and probability of defaults increase, driven as they should by common factors (e.g., the business cycle).

We now relax the assumption that p and ρ are independent, and assume that they are tied up through the following relation: $\rho = 3.8116 \times \ln(p+1)$, and let p vary from 0.10 to 0.30, such that ρ varies from 0.3633 to 1. Figure 14.17 shows that the pricing outcomes may change dramatically: assumptions on and modeling of default correlation may be of paramount importance for asset pricing.

First, and naturally, the yield on each securitized asset is now increasing in ρ because ρ is also increasing in the probability of default. Second, the Junior tranche has a yield that increases over a wide spectrum of values of ρ. Note that the Junior tranche bends back to lower values as the default correlation gets closer to one, reflecting the fact that default clustering makes this tranche quite valuable in good times, as explained earlier.

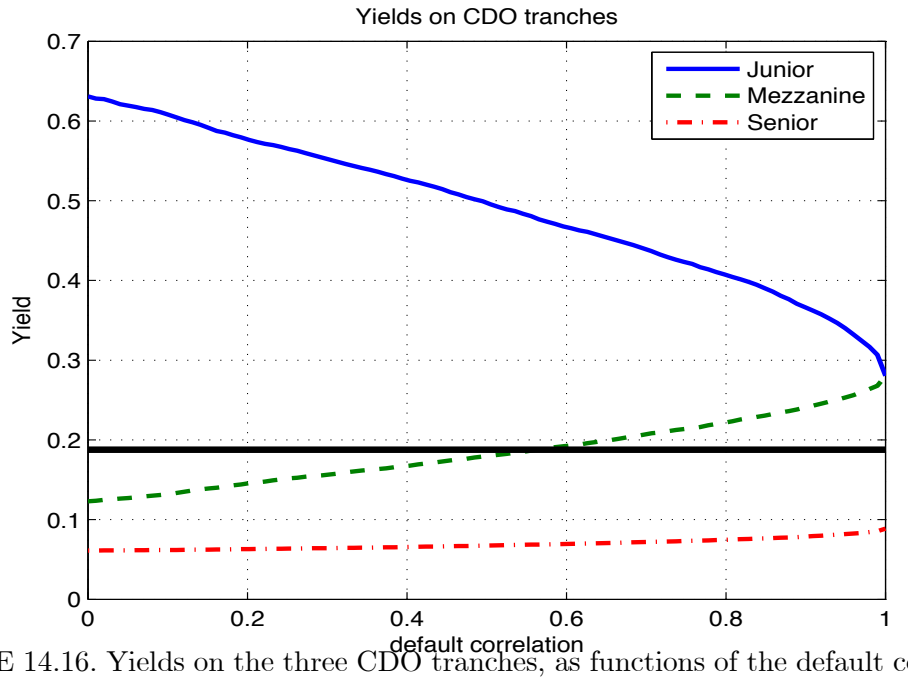


FIGURE 14.16. Yields on the three CDO tranches, as functions of the default correlation in the assets base, with probability of default for each name $p = 20\%$. The thick, horizontal, line is the yield on each securitized asset.

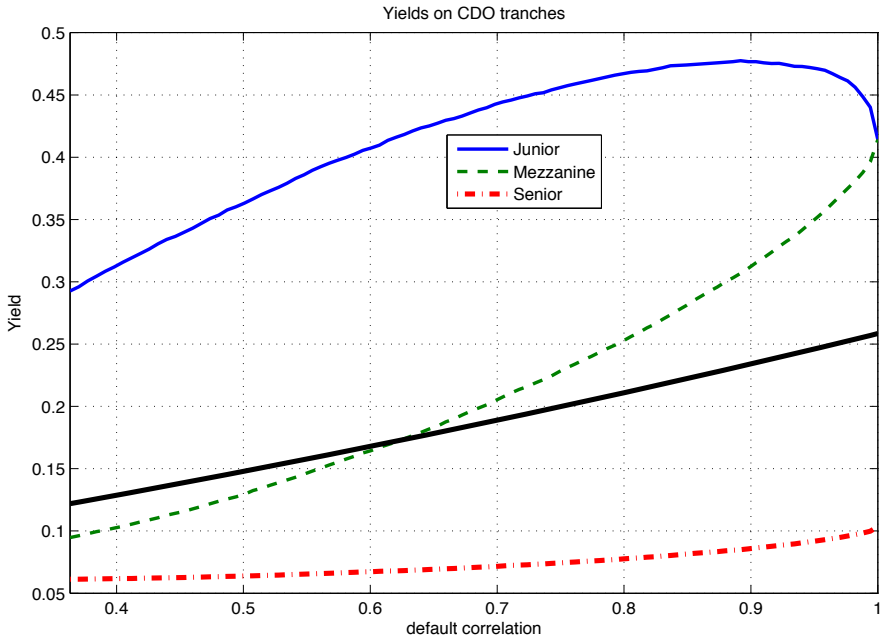


FIGURE 14.17. Yields on the three CDO tranches, as functions of the default correlation in the asset base, with probability of default and default correlation related by $\rho = 3.8116 \times \ln(p+1)$, $p \in [0.10, 0.30]$. The thick curve line depicts the yield on each securitized asset.

14.3.3.5 N-th to default

In this contract, the owner of the 1^{st} to default bears the risk of the first default that occurs in the asset pool, such that his payoff is:

$$\Pi_1 = \Pr(\text{No. of Defaults} \geq 1) \times 40 + \Pr(\text{No. of Defaults} < 1) \times 100.$$

Likewise, the owner of the 2^{nd} to default bears the risk of the second default in the asset pool, and his payoff is:

$$\Pi_2 = \Pr(\text{No. of Defaults} \geq 2) \times 40 + \Pr(\text{No. of Defaults} < 2) \times 100.$$

Finally, the owner of the 3^{rd} to default bears the risk of the third default in the asset pool and his payoff is:

$$\Pi_3 = \Pr(\text{No. of Defaults} = 3) \times 40 + \Pr(\text{No. of Defaults} < 3) \times 100.$$

Let us assume that default correlation is zero for simplicity. We have previously calculated the previous probabilities, which are $\Pr(\#\text{ def} \geq 1) = 0.243 + 0.027 + 0.001 = 0.271$, $\Pr(\#\text{ def} \geq 2) = 0.027 + 0.001 = 0.028$ and $\Pr(\#\text{ def} = 3) = 0.001$, such that the prices of the contracts are

$$\begin{aligned} P_{1^{st}\text{-to-def}} &= e^{-0.06}\left[0.271 \times 40 + (1 - 0.271) \times 100\right] = 78.863 \\ P_{2^{nd}\text{-to-def}} &= e^{-0.06}\left[0.028 \times 40 + (1 - 0.028) \times 100\right] = 92.594 \\ P_{3^{rd}\text{-to-def}} &= e^{-0.06}\left[0.001 \times 40 + (1 - 0.001) \times 100\right] = 94.120 \end{aligned}$$

and the corresponding yields are $Y_{1^{st}\text{-to-def}} = -\ln(0.78863) = 23.74\%$, $Y_{2^{nd}\text{-to-def}} = -\ln(0.92594) = 7.69\%$, and $Y_{3^{rd}\text{-to-def}} = -\ln(0.94120) = 6.06\%$.

14.3.3.6 Numerical examples concerning stylized structured products

A. Defaultable bonds

Suppose we observe the following risk-structure of spreads, related to two bonds maturing in two years:

$$\text{Spread}_A\,(2\text{ years}) = 1.5\%,\quad \text{Spread}_B\,(2\text{ years}) = 2.5\%,$$

where A and B denote the rating classes the bond issuers belong to. Assume that the one-year transition rating matrix, defined under the risk-neutral probability, is:

		To		
		A	B	Def
From	A	0.7	0.3	0
	B	0.3	0.5	0.2
	Def	0	0	1

where "Def" denotes default. We assume that in the event of default, the recovery value of the bond is paid off at the end of the second period. We want to determine the expected recovery rates for the two bonds, and which expected recovery rate is the largest. We have:

$$e^{rT}\frac{D_{0,i}}{N} = \left[\frac{\text{Rec}_i}{N}\mathbb{Q}_i\,(2) + (1 - \mathbb{Q}_i\,(2))\right],\quad i \in \{A, B\}\,.$$

Therefore,

$$\text{Spread}_A\,(2\text{ years}) = 1.5\% = -\frac{1}{2}\ln\left[\frac{\text{Rec}_A}{N}\mathbb{Q}_A\,(2) + (1 - \mathbb{Q}_A\,(2))\right] \tag{14.60}$$

$$\text{Spread}_B\,(2\text{ years}) = 2.5\% = -\frac{1}{2}\ln\left[\frac{\text{Rec}_B}{N}\mathbb{Q}_B\,(2) + (1 - \mathbb{Q}_B\,(2))\right] \tag{14.61}$$

We have to find $\mathbb{Q}_A\,(2)$ and $\mathbb{Q}_B\,(2)$. The transition matrix for two years is,

$$\mathbb{Q}\,(2) = \begin{bmatrix} 0.7 & 0.3 & 0 \\ 0.3 & 0.5 & 0.2 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 0.7 & 0.3 & 0 \\ 0.3 & 0.5 & 0.2 \\ 0 & 0 & 1 \end{bmatrix},$$

such that,

$$\begin{aligned}\Pr\{\text{A defaults in 2 years}\} &= \mathbb{Q}_A\,(2)\\ &= \underbrace{0.70\times 0}_{A\to A\to Def} + \underbrace{0.30\times 0.20}_{A\to B\to Def} + \underbrace{0\times 1}_{A\to Def\to Def}\\ &= 0.06\end{aligned}$$

and

$$\begin{aligned}\Pr\{\text{B defaults in 2 years}\} &= \mathbb{Q}_B\,(2)\\ &= \underbrace{0.20\times 1}_{B\to Def\to Def} + \underbrace{0.50\times 0.20}_{B\to B\to Def} + \underbrace{0.30\times 0}_{B\to A\to Def}\\ &= 0.20 + 0.10 = 0.30.\end{aligned}$$

Hence, using Eqs. (14.60)-(14.61), we have

$$\begin{aligned}\text{Spread}_A\,(2\text{ years}) &= 1.5\% = -\frac{1}{2}\ln\left[\frac{\text{Rec}_A}{N}0.06+(1-0.06)\right]\\ \text{Spread}_B\,(2\text{ years}) &= 2.5\% = -\frac{1}{2}\ln\left[\frac{\text{Rec}_B}{N}0.30+(1-0.30)\right]\end{aligned}$$

Solving, yields

$$\frac{\text{Rec}_A}{N}=50.7\%,\quad \frac{\text{Rec}_B}{N}=83.7\%.$$

The expected recovery rate for the second bond is the largest. This is because the probability firm B defaults is much larger than the probability firm A defaults and yet the two spreads are relatively close to each other. So to rationalize the two spreads, we need a large recovery rate for the second bond.

What would happen to the two credit spreads, once we assume that the recovery rates are the same, and equal to 50%? This question sheds additional light to the previous findings. If the recovery rates are the same and both equal 50%,

$$\begin{aligned}\text{Spread}_A\,(2\text{ years}) &= -\frac{1}{2}\ln\left[0.50\mathbb{Q}_A\,(2)+(1-\mathbb{Q}_A\,(2))\right]\\ \text{Spread}_B\,(2\text{ years}) &= -\frac{1}{2}\ln\left[0.50\mathbb{Q}_B\,(2)+(1-\mathbb{Q}_B\,(2))\right]\end{aligned}$$

Then, using the previously computed transition probabilities for two years, we obtain:

$$\text{Spread}_A\,(2\text{ years})=1.52\%,\quad \text{Spread}_B\,(2\text{ years})=8.12\%.$$

When the recovery rates are the same, the spread on the second bond diverges substantially from that on the first bond.

B. Collateralized debt obligations

Let us keep on using the same framework as before, but use different figures, so as to figure out the implications for CDOs pricing. Consider the following one year transition matrix, under the risk-neutral probability:

		To		
		A	B	Def
From	A	0.7	0.3	0
	B	0.1	0.6	0.3
	Def	0	0	1

where "Def" denotes default. Consider (i) 1 one-year bond issued by a company rated A, and (ii) 3 one-year bonds issued by a company rated B. Both bonds have face value equal to 100. We assume that the recovery values in case of default of all these bonds are the same, and equal to 50. Finally, we assume the safe interest rate is taken to be equal to zero.

Consider a collateralized debt obligation (CDO, in the sequel), which gathers the previous four bonds. Therefore, the CDO has nominal value of 400, and pays off in one year. The CDO has (i) a senior tranche, with nominal value equal to 150; (ii) a mezzanine tranche, with nominal value equal to N_1; and (iii) a junior tranche, with nominal value equal to N_2. We assume that the structure is such that $N_1 > 100$.

First, we determine the price and yields on all the four bonds. Since the safe interest rate is zero, and the company rated A is safe, up to the next year, the price of the A bond is 100, and its yield is zero. As for the three bonds rated B, we have:

$$P_B = 50 \times 0.3 + 100 \times 0.7 = 85.0, \quad Yield_B = -\ln 0.85 = 16.25\%.$$

Second, we determine the yield on the junior tranche, and derive the yield on the mezzanine, as a function of its nominal value N_1. To determine the yield on the tranches, we need to figure out the following table:

No_Def	Pr	Π	π_0	π_1	π_2
0	0.7	400	150	N_1	N_2
1	0	na	na	na	na
2	0	na	na	na	na
3	0.3	250	150	100	0
4	4	na	na	na	na

where No_Def denotes the number of defaults, Pr is the probability of No_Def, Π is the pool payoff, defined as,

$$\Pi = \text{No_Def} \times 50 + (4 - \text{No_Def}) \times 100,$$

and, finally: π_0 is the payoff to the senior tranche, π_1 is the payoff to the mezzanine tranche, and, π_2 is the payoff to the junior tranche. Therefore, the tranches are prices as:

$$P_{\text{mezz}} = 0.70 \times N_1 + 0.30 \times 100, \quad P_{\text{jun}} = 0.70 \times N_2,$$

such that

$$\begin{aligned} Y_{\text{mezz}} &= -\ln\left(\frac{0.70 \times N_1 + 0.30 \times 100}{N_1}\right) = -\ln\left(0.70 + 0.30 \times \frac{100}{N_1}\right) \\ Y_{\text{jun}} &= -\ln\left(\frac{0.70 \times N_2}{N_2}\right) = 35.67\%. \end{aligned}$$

Naturally, we need to have that $Y_{\text{mezz}} < Y_{\text{jun}}$, an inequality that can be easily shown, as follows:

$$Y_{\text{jun}} = -\ln(0.70) > -\ln\left(0.70 + 0.30 \times \frac{100}{N_1}\right) = Y_{\text{mezz}}.$$

A reverse enginnering question is, now, to determine which nominal value of the mezzanine tranche N_1 is needed, to ensure that the yield on the mezzanine tranche is equal to or greater than the yields on the bonds issued by the company with credit rating B? The answer is $N_1 = 200$, for in this case, the mezzanine tranche would have the same payoff structure as the bond rated B: it would deliver (i) the face value, in the event the company rated B does not default; and (ii) half of its nominal value, 100, in the event the company rated B does default.

Finally, we ask which nominal value of the mezzanine tranche N_1 is needed, to ensure that the yield on the mezzanine is equal to 18%? And what is the corresponding nominal value of the junior tranche, N_2? To address these issues, we first want that:

$$\text{Yield_mezzanine} = -\ln\left(\frac{0.70 \times N_1 + 0.30 \times 100}{N_1}\right) = 18\%.$$

Solving for N_1 yields, $N_1 = 221.78$. Therefore, $N_2 = 400 -$ Nominal_value_senior $- N_1 = 400 - 150 - 221.78 = 28.22$.

C. CDO: A second example

Consider the following one-year transition rating matrix, defined under the risk-neutral probability:

		To		
		A	B	Def
From	A	0.9	0.10	0
	B	0.3	0.50	0.2
	Def	0	0	1

We assume that available for trading are one bond rated A and two bonds rated B. The two bonds rated B are issued by the same firm. Each of these three bonds expires in one year, and has a face value of 100. We assume that the recovery rates for each bond are the same, and equal to 50%. Finally, we take the safe, short-term rate equal to zero.

Next, consider the following stylized CDO, one that pools the bond rated A and the two bonds rated B, and has, then, nominal value equal to 300. The CDO pays off in one year, and has three tranches: (a) a senior tranche, with nominal value N_1 equal to 150; (b) a mezzanine tranche, with nominal value equal to N_2; and (c) a junior tranche, with nominal value equal to N_3.

First, we determine the price of the bonds, which are $P_A = 100$ and $P_B = 50Q + 100\,(1-Q) = 90$, where we have used the fact that the risk-neutral probability of default for the second bond in one year is $Q = 0.2$.

Next, suppose that we want to set the nominal value of the mezzanine tranche, N_2, such that $N_2 \geq 50$. We wish to determine the price of each tranche, as a function of N_1, N_2 and N_3. Note that the realized pool payoff, Π, say, is

$$\Pi = \text{N_def} * 50 + (3 - \text{No_def}) * 100,$$

with obvious notation. We know the bond rated A will not default. So we may only have that the two bonds rated B may default, if any. Moreover, the two bonds rated B are issued by the same firm such that, if one of these two bonds defaults, the other defaults as well. More generally, we know that with probability 0.2, there will be two defaults, and with probability 0.80, there will be no defaults at all. To summarize, we have the default probabilities in the following table:

No_Def	Pr	Π	π_1	π_2	π_3
0	0.8	300	150	N_2	N_3
1	0	NA	NA	NA	NA
2	0.2	200	150	50	0
3	0	NA	NA	NA	NA

where π_1 is the payoff to the senior tranche, π_2 is the payoff to the mezzanine tranche, and π_3 is the payoff to the junior tranche.

Therefore, the price of the three tranches are: $P_{\text{sen}} = N_1 = 150$, $P_{\text{mezz}} = 0.8N_2 + 0.2 \cdot 50 = 0.8N_2 + 10$ and $P_{\text{jun}} = 0.8N_3$. Let us determine the yields on the junior and mezzanine tranche. The yield on the junior tranche is determined as $Y_{\text{jun}} = -\ln\left(\frac{P_{\text{jun}}}{N_3}\right) = -\ln(0.8) = 0.22314$. The yield on the mezzanine tranche cannot be determined before we choose its nominal value N_2:

$$Y_{\text{mezz}} = -\ln\left(\frac{P_{\text{mezz}}}{N_2}\right) = -\ln\left(\frac{0.8N_2 + 10}{N_2}\right).$$

The mezzanine yield is increasing in its face value. Intuitively, the higher N_2, the higher the loss-given-default for this tranche. We can solve for N_2 while taking as given a desired yield on the mezzanine tranche. For example, if $Y_{\text{mezz}} = 12\%$, the nominal value of the junior tranche N_3 is $N_2 = 115.05$, which implies the face value of the junior tranche to be $N_3 = 150 - 115.05 = 34.95$.

14.4 Managing loan losses

How do financial institutions hedge against losses that may originate from holding illiquid assets such as a portfolio of loans? Obviously, banks can securitize loans; but the economics of structured finance suggests that hedging through securitization may fail due to concerns for moral hazard or due to adverse selection problems (see Chapter 4), as the 2007 subprime crisis vividly reminds us (see Section 14.5). Alternatively, how much capital should a bank set apart to cover such expected losses? Understanding and modeling "value at risk" is a standard tool in the equity space. Equally standard did value-at-risk become in the credit space, albeit the details are more elaborated in this case.

This section provides an introduction to the modeling of loan losses and foundations of risk-management. Section 14.4.1 provides discussion regarding the extant regulatory framework; this overview also helps framing procyclicality issues dealt with later in this Chapter (see Section 14.5). Section 14.4.2 is an introduction to standard risk-management tools, tools that are mostly suitable to deal with equity products. Section 14.4.3 deals with the relatively more challenging case of credit risk.

14.4.1 Regulatory framework

Banks are required to set up capital buffers to guarantee the debt they issue against their risky activities of lending and investing. The Basel Committee on Banking Supervision (BCBS)[20] drafts accords aiming to create an international standard for the capital necessary to cope with these risks, together with rigorous tools for risk measurement and management. Quite simply, the greater the risk a bank is exposed to, the greater the amount of capital the bank needs to hold to safeguard its solvency, in the interest of overall economic stability. The main issue, then, is to correctly measure this risk.

The first accord of 1988, known as Basel I, focussed on minimal capital requirements to cope with *credit risk*, and was enforced by law by the Group of Ten in 1992 and then by more than 100 countries. It relied on the so-called Cooke ratio (after Peter Cooke of the Bank of England), a minimum capital adequacy standard of 8% of the total risk-weighted assets. The accord was quite coarse, in that it considered five broad classes of credit risk with which to weigh the assets, which did not discriminate about the credit quality across classes. For example, corporate loans had 100% weightings and loans to OECD countries had zero weigthings, independently of the ratings of the borrowing entities.

The first amendment to Basel I occurred in 1996, and aimed to include tools to cope with *market risk*. In 1999, a first consultative paper was drafted on a new accord, known as Basel II. One of the main issues under reform was the "one-size-fits-all" approach of Basel I—the fact

[20]The BCSB is a committee of banking supervisory authorities established by the central bank governors of the Group of Ten countries in 1975. It consists of senior representatives of bank supervisory authorities and central banks from Belgium, Canada, France, Germany, Italy, Japan, Luxembourg, the Netherlands, Spain, Sweden, Switzerland, the United Kingdom, and the United States. It usually meets at its permanent Secretariat, located at the Bank for International Settlements in Basel.

that default risk could be substantially lower for some of the assets within the same class of risk in the banks accounts. For example, banks could have securitized the loans with default risk lower than that implied by the flat rate within the same class, and hold those loans with higher default risk. This might have led to an increase in the overall riskiness of financial institutions. In 1998, the Federal Reserve Chairman Alan Greenspan pointed to the existence of incentives left to bank to implement regulatory arbitrage:

> "Banks arbitrage away inappropriately high capital requirements on their safest assets by removing these assets from the balance sheet via securitization. The issue is not solely whether capital requirements on the bank's residual risk in the securitized assets are appropriate. We should also be concerned with the sufficiency of regulatory capital requirements on the assets remaining on the book. In the extreme, such 'cherry picking' would leave on the balance sheet only those assets for which economic capital allocations are greater than the 8 percent regulatory standard."[Greenspan, 1998 p. 166]

There is a consensus that Basel II did indeed considerably mitigate these issues, by paying more attention to risk-sensitivity by means of a more precise set of indications about classes of risk and, also, distinguishing among credit risk, market risk and even operational risks.[21] Moreover, the Basel II accords aimed to a flexible supervisory system whereby banks could choose to measure and manage risks by following a "standardized" or an "internal" rating approach. The standardized approach would rely on the ratings supplied by dedicated rating agencies approved by national supervisors. Capital incentives would also be available to move towards more advanced, internal approaches, where banks calculate weightings through proprietary models. However, moving from mechanical rules to propr-generated ones requires a particular care as to the quality of the models, which have to be approved by national supervisors.[22] Therefore, Basel II encourages banks to improve their risk measurement and managament systems.

Basel II relies on three "three pillars," which impose increasingly stringent rules on capital adequacy:

> *Pillar I: Minimal capital requirements.* It defines a set of basic rules for measuring credit risk, market risk and operational risks (both "standardized" and "internal").
>
> *Pillar II: Supervisory review of capital adequacy.* It increases the role of banking supervisors, by setting rules of conduct national supervisors should maintain to ensure that banks have capital adequacy over and above the minimal capital requirements of Pillar 1.
>
> *Pillar III: Market disclosure.* It increases the role of market discipline and disclosure, by relying on the publication of information by banks, relating to issues such as risk measurements, risk-rating processes, or risk-management systems.

Pillar I defines a clear separation of three types of risk: (a) credit risk; (b) operational risk and (c) market risk. The calculation rules rely on the concept of total risk-weighted assets

[21] Operational risk is defined as the risk of losses resulting from inadequate or failed internal processes, people and systems, or external events. Examples of operational risk include two famous cases of "rogue trading": Nick Leeson, who in 1995 led Baring Bank to bankruptcy, through a loss of £1.3bn, and Jérôme Kerviel, who in 2008 led Société Genérale to a loss of €5bn.

[22] Note that within the internal rating approach, banks are not allowed to use internal models of credit risk. Banks that have received supervisory approval to use the internal approach may rely on their own internal estimates of risk components in determining the capital requirement for a given exposure. The risk components include: (i) measures of the probability of default, (ii) loss given default, (iii) exposure at default, (iv) effective maturity.

(Total RWA), defined as

$$\text{Total RWA} = \text{Cr} + \text{MnO},$$

where

$$\begin{aligned} \text{Cr} &= \text{Risk-weigthted assets for } \textit{credit} \text{ risk} \\ \text{MnO} &= \text{Assets weighted for } \textit{market} \text{ and } \textit{operational} \text{ risk} \\ &= (\text{Capital requirements for } \textit{market} \text{ and } \textit{operational} \text{ risk}) \times \underbrace{12.5}_{=1/0.08} \end{aligned}$$

For example, the capital requirements for market risk can be determined through dedicated VaR models, such as (and, possibly, more sophisticated versions of) those surveyed in Section 14.4. The (total) minimum capital requirements are taken to be 8% of the total risk-weighted assets,

$$\frac{\text{Regulatory capital}}{\text{Total RWA}} \geq 8\%.$$

One immediate issue arising with Basel II is its heavy reliance on credit rating agencies for what pertains the standardized approach to credit risk. This approach might be misguided due to conflicts of interest between credit rating agencies and the firms these agencies rate the debt of. At the time of writing, rating agencies are mostly unregulated, with the credit risk estimates quality being obviously observable only with lags and, importantly, too late should serious mispricing take place.

A second issue is "procyclicality": in bad times, banks reduce lending, exacerbating the current economic developments, which makes banks reduce lending even further, over a vicious circle. Section 14.5 deals deal with procyclicality issues. The dangers of procyclicality in a regulatory context is that in bad times, when risks intensify, the banking system is given additional burden due to regulatory capital, which might lead to a further lending shrinkage. Basel III introduces a regulatory device that allow to mitigate cyclicality by requiring banks to build up capital buffers in good times with which to cope in adverse times.

After a number of additional consultative papers, national regulators indicated to the Financial Stability Institute (FSI)[23] that they would implement Basel II by 2015. During the process when the European Union was implementing Basel II through its EU Capital Adequacy Directive (CAD III), the global financial crisis following the 2007 subprime events determined a re-thinking of Basel II, leading to a new set of rules, known as Basel III. The main innovations of Basel III are summarized by the following points: (i) new capital requirements, such as those summarized in the table below, as well as a new mandatory capital conservation buffer of 2.5% of the Total RWA, to face economic stress; (ii) new rules that allow national supervisors to require banks to set up capital up to 2.5% of the Total RWA in times of high credit growth—countercyclical capital buffers; and (iii) a target for the leverage ratio, defined as the ratio of Tier 1 capital (i.e. equity plus reserves minus intangible assets) over total assets net of intangible assets, to be at least 3% ($\frac{\text{Tier 1}}{\text{Assets}} > 3\% \Leftrightarrow \text{Lev} \equiv \frac{\text{Non Tier 1}}{\text{Tier 1}} < \frac{1}{0.03} - 1$), as well as additional liquidity ratios. The following table summarizes the main differences in capital requirements that Basel III introduces against Basel II.

[23]The FSI, headquartered by the Bank for International Settlements in Basel, was established in 1999 in response to the Asian crisis of 1997.

Capital requirements as a % of the Total RWA		
	Basel II	Basel III
Common equity	2%	4.5%
Tier 1	4%	6%
Total capital	8%	8%
Common equity (conservation)	–	2.5%

The next sections provide analytical tools to deal with capital requirements regarding the occurrence of loan losses. We begin with fundamental tools that are relevant in the equity space.

14.4.2 Foundations of risk-management

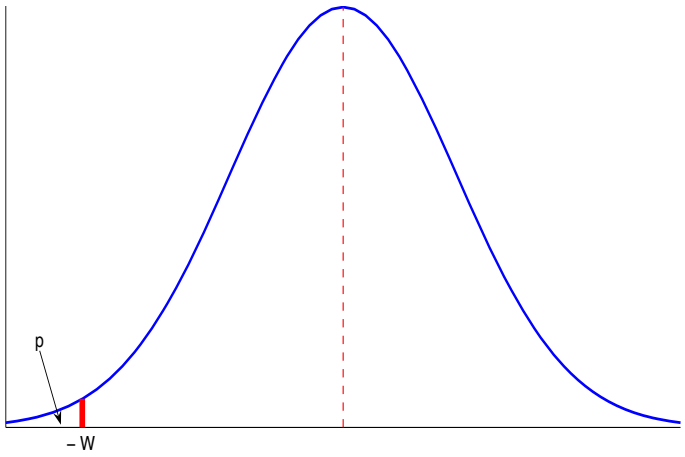
This section develops standard concepts regarding assessment and measurement of market risk. The following section extends these concepts to the credit case.

14.4.2.1 Value at Risk (VaR)

We need to review Value at Risk (VaR), in general. VaR is a method of assessing risk that uses statistical techniques. Useful for supervision and management of financial risks. Origins: reaction to financial disasters in the early 1990s involving Orange County, Barings, Metallgesellschaft, Daiwa, etc. *VaR* measures the worst expected loss over a given horizon under normal market conditions at a given confidence level.

Consider the following definition:

DEFINITION I: *We are* $(1-p)\%$ *certain that a given portfolio will not suffer of a* loss *larger than* $£W$ *over the next* N *weeks,* $\Pr(\text{Loss} < -W) = p$. *That is,* $£\text{VaR}_p = £W$.



Equivalently, note that the (negative) portfolio return is $\frac{\text{Loss}}{V_0} = \frac{\Delta V}{V_0}$, where ΔV is the change in value of the portfolio over the next N days and V_0 is the current value of the portfolio. Hence,

$$p = \Pr\left(\text{Loss} < -\text{VaR}_p\right) = \Pr\left(\frac{\Delta V}{V_0} < -\frac{\text{VaR}_p}{V_0}\right).$$

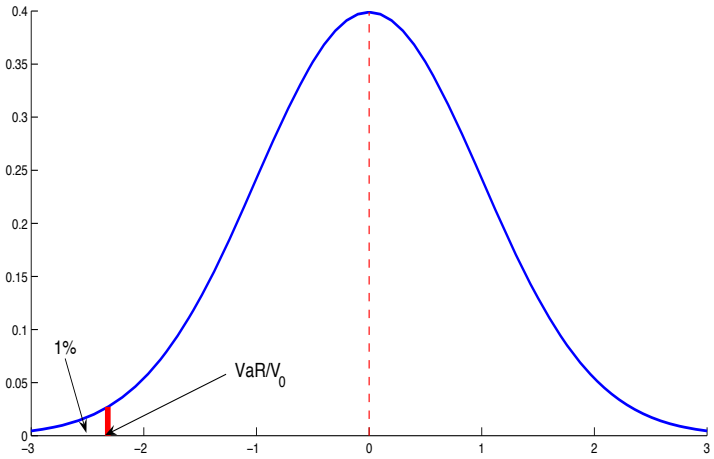
This formulation leads us to the following alternative definition:

DEFINITION II: *We are* $(1-p)\%$ *certain that a given portfolio will not experience a* relative loss *larger than* $\frac{\text{VaR}_p}{V_0}$ *over the next* N *weeks.*

So in practice, we shall have to find the relative loss, ℓ_p, for a given confidence p, as follows:

$$p = \Pr\left(\frac{\Delta V}{V_0} < -\ell_p\right), \quad \text{where } \ell_p = \frac{\text{VaR}_p}{V_0}.$$

The corresponding VaR_p is just $\text{VaR}_p = \ell_p \cdot V_0$. For example, suppose that the portfolio return over the next 2 weeks, $\frac{\Delta V}{V_0}$, is normally distributed with mean zero and unit variance. We know that $0.01 = \Pr(\frac{\Delta V}{V_0} < -2.32)$. Hence, $\text{VaR}_p = 2.32 \cdot V_0$.



We are 99% certain that our portfolio will not suffer of a *loss* larger than -2.32 times its current value over the next 2 *weeks*. We are 99% certain that our portfolio will not experience a *relative loss* larger than -2.32 over the next 2 weeks.

As a second example note that the previous assumption about the portfolio return was extreme. Assume, instead, the porfolio return over the next 2 weeks, $\frac{\Delta V}{V_0}$, is normally distributed with mean zero and variance $\sigma^2 = \frac{2}{52}\sigma^2_{\text{year}}$, where σ^2_{year} is the annualized variance. We assume that $\sigma^2_{\text{year}} = 0.15^2$. We have to re-scale the previous formulas, as follows. First, we introduce a variable $\tilde{\epsilon} \sim N(0,1)$, i.e. $\tilde{\epsilon}$ is normally distributed with mean zero and variance $= 1$. So we can write,

$$\frac{\Delta V}{V_0} \overset{d}{=} \tilde{\epsilon} \cdot \sigma \sim N\left(0, \sigma^2\right),$$

and, hence,

$$0.01 = \Pr\left(\tilde{\epsilon} < -2.32\right) = \Pr\left(\Delta V < -2.32 \cdot V_0 \cdot \sigma\right),$$

whence, $\text{VaR}_p = 2.32 \cdot V_0 \cdot \sigma$. We know the annualized variance, $\sigma^2_{\text{year}} = 0.15^2$, from which we can derive the two-week standard deviation, $\sigma^2 = \frac{2}{52}\sigma^2_{\text{year}} \approx 0.03^2$, and, hence, $\frac{\text{VaR}_p}{V_0} = 2.32 \cdot \sigma = 2.32 \cdot 0.03 \approx 7\%$. That is, we are 99% certain that our portfolio will not suffer of a *loss* larger than 7% times its current value over the next 2 weeks. We are 99% certain that our portfolio will not experience a *relative loss* larger than 7% over the next 2 weeks.

More generally, we may assume the porfolio return over the next 2 weeks, $\frac{\Delta V}{V_0}$, is normally distributed with mean μ and variance σ^2. In this case,

$$\frac{\Delta V}{V_0} \overset{d}{=} \mu + \tilde{\epsilon} \cdot \sigma \sim N\left(\mu, \sigma^2\right),$$

and, hence,

$$0.01 = \Pr\left(\tilde{\epsilon} < -2.32\right) = \Pr\left(\Delta V < -V_0 \cdot (2.32 \cdot \sigma - \mu)\right)$$

whence, $\text{VaR}_p = V_0 \cdot (2.32 \cdot \sigma - \mu)$. In practice, μ is very small if the horizon is as short as two weeks.

14.4.2.2 Challenges to VaR

Challenges related to *distributional assumptions, nonlinearities, or conceptual difficulties.*

Distributional assumptions

The assumption that data are generated by a normal distribution does not describe asset returns well. In previous chapters of this Part, and in Part II, we explain that we need ARCH effects, stochastic volatility and multifactor models. More generally, data can exhibit changes in regimes, nonlinearities and fat tails. Fat tails are particularly important to understand, since this is what we're interested in after all. More in general, it is quite challenging to understand what the data generating process is, especially in so far as we consider portfolios of assets. Asset returns and volatilities are typically correlated, with correlation rising in bad times—correlation is stochastic.

We may make distributional assumptions but then, these assumptions have to be carefully assessed through, for example, backtesting (to be explained below). We may proceed with nonparametric methods, and this is indeed a promising avenue, but with its caveats.

How do nonparametric methods work? These methods rely on an old and idea, which is to estimate the data distribution through histograms. These histograms can be readily used to compute VaR. This approach is nonparametric in nature, as it does not rely on any model. A more refined method replaces "rough" histograms with "smoothed" histograms, as follows. Suppose to have access to a time series of data x_n, which are drawn from a certain probability law, with density $f(x)$. We may define the following estimate of the density $f(x)$,

$$\hat{f}_N(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{\lambda}K\left(\frac{x - x_n}{\lambda}\right),$$

where N is the sample size, and K is some symmetric function integrating to one. We may think of $\hat{f}_N(x)$ as a smoothed histogram, with window bin equal to λ. It is possible to show that as N goes to infinity and λ goes to zero at a certain rate, $\hat{f}_N(x)$ converges "in probability" to $f(x)$, for all x. But we are not done, since there are not obvious rules to choose λ and K?

The choice of λ is notoriously difficult. Unfortunately, the "bias," $\hat{f}_N(x) - f(x)$, tends to be large exactly on the tails of $f(x)$, which do represent the region we're interested in. In general, we can use Montecarlo simulations out of a smoothed density like this to compute VaR.

Nonlinearities

Finally, portfolios of assets can behave in a nonlinear fashion, especially when the portfolio contains derivatives. In general, the value of a portfolio including M assets is,

$$P = \sum_{j=1}^{M} \alpha_i S_i,$$

where α_i is the number of the i-th asset in the portfolio, and S_i is the price of the i-th asset in the portfolio. Holding α_i constant, the variation on the portfolio return is simply a weigthed average of all the asset returns,

$$\Delta P_T \equiv P_T - P_t = \sum_{j=1}^{M} \alpha_i \Delta S_i \Longleftrightarrow \frac{\Delta P}{P} = \sum_{j=1}^{M} \left(\frac{\alpha_i S_i}{P} \right) \frac{\Delta S_i}{S_i},$$

where the variations relate to any time interval. Often, the prices S_i are rational functions of the state variables, or are interlinked through arbitrage restrictions. Use factors to determine the risk associated with fixed income securities. When the horizon of the VaR is large, it is unlikely that α_i is constant. Typically, we shall need to go for numerical methods, based, for example, on Monte Carlo simulations. So all in all, we need to have a careful understanding of the derivatives in the book, and proceed with back testing and stress testing.

VaR as an appropriate measure of risk

There are technical difficulties with the very definition of VaR. VaR suffers from some statistic-theoretic foundation. VaR tells us that 1% of the time, losses will exceed the VaR figure, but it does not tell us the entity of the loss. So we need to compute the *expected shortfall*. Any risk measure should enjoy a number of sensible properties. Artzner et al. (1999) have noted a number of properties, and showed that VaR does not enjoy the so-called *subadditivity* property, according to which the sum of the risk measures for any two portfolios should be larger than the risk measure for the sum of the two portfolios. VaR doesn't satisfy the subadditivity property, but expected shortfall does satisfy the subadditivity property.

Define expected shortfall.

[In progress]

14.4.2.3 Backtesting

How well the VaR estimate would have performed in the past? How often the loss in a given sample exceeded the reference-period 99% VaR? If the exceptions occur more than 1% of the time, there is evidence that the models leading to VaR estimates are "misspecified"—a nice word for saying "bad" models.

The mechanics of backtesting is as follows. Suppose the models leading to the VaR are "good". By construction, the probability the VaR number is exceeded in any reference period is p, where p is the coverage rate for the VaR. Next, we go to our sample, which we assume it comprises N days, and let M be the number of days the VaR is exceeded. We wish to test whether the number of exceptions we observe in the sample "conforms" to the expected number

of exceptions based on the VaR. For example, it might be that the number of exceptions we have observed, M, is larger than the expected number of exceptions, $p \cdot N$. We want to make sure this circumstance arose due to sample variability, rather than model misspecification. A simple one-tail test is described below.

Let us compute the probability that in N days, the VaR is exceeded for M or more days. Assuming exceptions are binomially distributed, this probability is,

$$\Pi_p = \sum_{k=M}^{N} \frac{N!}{k!\,(N-k)!} p^k \,(1-p)^{N-k}\,.$$

Then, we can say the following. If $\Pi_p \leq 5\%$ (say), we reject the hypothesis that the probability of exceptions is p at the 5% level—the models we're using are misspecified. If $\Pi_p > 5\%$ (say), we cannot reject the hypothesis that the probability of exceptions is p at the 5% level—we can't say the models we're using are misspecified. This test is reviewed in more detail by Hull (2007, p. 208). Other tests are reviewed by Christoffersen (2003, p. 184).

14.4.2.4 Stress testing

Stress testing is a technique through which we generate artificial data from a range of possibles *scenarios*. Stress scenarios help cover a range of factors that can create extraordinary losses or gains in trading portfolios, or make the control of risk in those portfolios very difficult. These factors include low-probability events in all major types of risks, including the various components of credit, market, and operational risks. Stress scenarios need to shed light on the impact of such events on positions that display both linear and nonlinear price characteristics (i.e. options and instruments that have options-like characteristics).

Possible scenarios include simulating (i) shocks that although rare or even absent from the historical database at hand, are likely to happen anyway; and (ii) shocks leading to structural breaks and/or smooth transition in the data generating mechanism. One possible example is to set the percentage changes in all market variables in the portfolio equal to the worst percentage changes having occured in ten days in a row during the subprime crisis 2007-2008.

This example on the subprime crisis is related to the *historical simulation approach* to generate scenarios. This approach consists can be explained through a single formula. Let v_t the value of some market variable i in day t in our sample, where $t = 0, \cdots, T$ (say). We can generate T scenarios for the next day, $T+1$, as follows.

(i) The first scenario is that in which each variable grows by the same amount it grew at time 1, $v_{T+1} = v_T \cdot \frac{v_1}{v_0}$.

(ii) The second scenario is that in which each variable grows by the same amount it grew at time 2, $v_{T+1} = v_T \cdot \frac{v_2}{v_1}$.

(iii) $\cdots$

(iv) The T-th scenario is that in which each variable grows by the same amount it grew at time T, $v_{T+1} = v_T \cdot \frac{v_T}{v_{T-1}}$.

(v) The T scenarios are generated for all the market variables, which would give us an artificial multivariate sample of T observations. We can use this sample for many things, including VaR.

14.4.2.5 Measures of systemic risk

This subsection is important, also in light of Section 14.5
[In progress]

14.4.3 Credit risk and VaR

We can use the tools in Section 14.2 to assess the likelihood of default for a given name. The important thing to do is to use the physical probability of default, not the risk neutral one. The risk neutral probability of default is likely to be larger than the physical one. Therefore, using the risk neutral probability leads to too conservative estimates.

VaR for credit risks pose delicate issues as well. The key issue is the presence of default correlation. In practice, defaults among names or loans are likely to be correlated, for many reasons. First, there might be direct relationships or, more generally, network effects, among names. Second, firms performance could be driven by common economic conditions, as in the one factor model which we now describe. This one factor model, developed by Vasicek (1987), is at the heart of Basel II. In the appendix, we provide additional technical details about how this model is related to a modeling tool known as copulae functions. We now proceed to develop this model in an intuitive manner. Let us define the following variable:

$$z_i = \sqrt{\rho} F + \sqrt{1-\rho}\epsilon_i, \tag{14.62}$$

where F is a common factor among the names in the portfolio, ϵ_i is an idiosynchratic term, and $F \sim N(0,1)$, $\epsilon_i \sim N(0,1)$. As we explain in the Appendix, $\rho \geq 0$ is meant to capture the default correlation among the names.

Next, assume that the *physical* probability each firm defaults, by T, say $\mathbb{P}(T)$, is the same for each firm within the same class of risk, and given by,

$$\mathbb{P}(T) = \Phi(\zeta_{\text{PD}}) \equiv \text{PD},$$

where PD is the probability of default, and Φ is the cumulative distribution of a standard normal variable. That is, by time T, each firm defaults any time that,

$$z_i < \zeta_{\text{PD}} \equiv \Phi^{-1}(\text{PD}),$$

where Φ^{-1} denotes the inverse of Φ. One economic interpretation of Eq. (14.62) is that z_i is the value of a firm and that the firm defaults whenever this value hits some exogenously given barrier ζ_{PD}.

Conditionally upon the realization of the macroeconomic factor F, the probability of default for each firm is,

$$p(F) \equiv \Pr(\text{Default}|\, F) = \Phi\left(\frac{\Phi^{-1}(\text{PD}) - \sqrt{\rho}F}{\sqrt{1-\rho}}\right). \tag{14.63}$$

By the law of large numbers, this is quite a good approximation to the default rate for a portfolio of a large number of assets falling within the same class of risk.

We see that this conditional probability is decreasing in F: the larger the level of the common macroeconomic factor, the smaller the probability each firm defaults. Hence, we can fix a value of F such that $\Pr(\text{Default}|\, F) = \text{Default rate}$ is what we want. Note, the probability F is larger than $-\Phi^{-1}(x)$ is just x! Formally,

$$\Pr\left(F > -\Phi^{-1}(x)\right) = \Pr\left(-F < \Phi^{-1}(x)\right) = \Phi\left(\Phi^{-1}(x)\right) = x.$$

Then, with probability x, the default rate will not exceed

$$\text{VaR}_{\text{Credit Risk}}(x) = \Phi\left(\frac{\Phi^{-1}(\text{PD}) + \sqrt{\rho}\Phi^{-1}(x)}{\sqrt{1-\rho}}\right).$$

It is easy to see that $\text{VaR}_{\text{Credit Risk}}(x)$ increases with ρ. Basel II sets $x = 0.999$ and, accordingly, it imposes a capital requirement equal to,

$$\text{LGD} \times (\text{VaR}_{\text{Credit Risk}}(0.999) - \text{PD}) \times \text{Maturity adjustment}.$$

The reason Basel II requires the term $\text{VaR}_{\text{Credit Risk}}(0.999) - \text{PD}$, rather than just $\text{VaR}_{\text{Credit Risk}}$, is that what is really needed here is the capital in excess of the 99.9% worst case loss over the expected idiosyncratic loss, PD. Well functioning capital markets should already discount the idiosyncratic losses.

Finally, Basel II requires banks to compute ρ through a formula in which ρ is inversely related to PD. The formula is based on empirical research (see Lopez, 2004): for a firm which becomes less creditworthy, the PD increases and its probability of default becomes less affected by market conditions. Basel II requires banks to compute a maturity adjustment factor that takes into account that the longer the maturity the more likely it is a given name might eventually migrate towards a more risky asset class.

The previous model can be further elaborated. We ask: (i) What is the unconditional probability of defaults, and (ii) what is the density function of the fraction of defaulting loans?

First, note that conditionally upon the realization of the macroeconomic factor F, defaults are obviosly independent, being then driven by the idiosyncratic terms ϵ_i in Eq. (14.62). Given N loans, and the realization of the macroeconomic factor F, these defaults are binomially distributed as:

$$\Pr(\text{No of defaults} = n|\, F) = \binom{N}{n} p(F)^n (1 - p(F))^{N-n},$$

where $p(F)$ is as in Eq. (14.63). Therefore, the unconditional probability of n defaults is:

$$\Pr(\text{No of defaults} = n) = \int_{-\infty}^{\infty} \Pr(\text{No of defaults} = n|\, F)\, \phi(F)\, dF,$$

where ϕ denotes the standard normal density. This formula provides a valuable tool analysis in risk-management. It can be shown that VaR levels increase with the correlation ρ.

Next, let ω denote the fraction of defaulting loans. For a large portfolio of loans, $\omega = p(F)$, such that:

$$\Pr(\omega \leq x) = \int_{-\infty}^{\infty} \Pr(\omega \leq x|\, F)\, \phi(F)\, dF = \int_{-\infty}^{\infty} \mathbb{I}_{p(F)\leq x} \phi(F)\, dF = \Phi(F^*), \tag{14.64}$$

where $\mathbb{I}$ denotes the indicator function, and F^* satisfies, by Eq. (14.63), $-F^* : x = p(-F^*) = \Phi\left(\frac{\Phi^{-1}(\text{PD}) + \sqrt{\rho}F^*}{\sqrt{1-\rho}}\right)$. Solving for F^* leaves:

$$F^* = \frac{\sqrt{1-\rho}\Phi^{-1}(x) - \Phi^{-1}(\text{PD})}{\sqrt{\rho}}.$$

It is the threshold value taken by the macroeconomic factor that guarantees a frequency of defaults ω less than x. Replacing F^* into Eq. (14.64) delivers the cumulative distribution function for ω. The density function $f(x)$ for the frequency of defaults is then:

$$f(x) = \sqrt{\frac{1-\rho}{\rho}} e^{\frac{1}{2}\left(\Phi^{-1}(x)\right)^2 - \frac{1}{2\rho}\left(\sqrt{1-\rho}\Phi^{-1}(x) - \Phi^{-1}(\mathrm{PD})\right)^2}.$$

This model can be generalized to one where the asset value of the firm, Z_i, is given by a multifactor model,

$$Z_i = \sum_{j=1}^{K} \beta_{i,j} Y_j + \epsilon_i,$$

and accordingly interpret the event of default as that occurring as soon as $Z_i < \zeta_{\mathrm{PD},i}$, where $\zeta_{\cdot,i}$ now can change across firms, similarly as explained by Schönbucher (2003, Chapter 10).

14.5 Procyclicality, credit crunches and quantitative easing

How is it that some relatively small losses in the banking system may have the power to generate a spectacular financial crisis such as that of the late 2000s? After all, these initial losses did not amount to substantially more than one trillion USD. This episode was equally devastating regarding its effects on the real sphere of the economy, which experienced one of the deepest recessions after World War II, the "Great Recession."

"Procyclicality" can be one explanation for these developments: a trend in asset prices feeds some automatic mechanism, which, in turn, reinforces the asset price trend, thereby creating a feedback loop between the asset price trend and the mechanism. The equilibria in these markets can be quite different from those arising in the absence of automatic mechanisms and, sometimes, no equilibrium may even exist in the first place. These mechanisms are automatic in that, once in place, they are not under control of any decision makers. Examples of procyclicality were given throughout the whole lectures. For example, Chapter 11 explains that program trading is most likely the reason underlying the Black Monday crash of October 19th, 1987; and the flash crash of May 6th, 2010 is the "high-tech" counterpart to the 1987 crash (see also Chapter 11).

Chapter 9 deals with an instance of procyclicality, which is more closely related to the topics studied in this chapter: the amplification of business cycles due to capital market frictions (see Section 9.14), a phenomenon described as financial accelerator. The mechanism operates through agency problems, which lead to an increase in the cost of external funds in bad times and, then, a decreased availability of these funds, whence, a further deterioration of the economic conditions, over a vicious circle.

This section provides discussions of similar procylicality problems, which arise through the balance sheets of financial intermediaries. Section 14.5.1 reviews a few institutional facts sorrounding the 2007 subprime turmoil. Section 14.5.2 develops a few models where the amplification of small shocks occurs because financial intermediaries have concerns regarding the structure of their books. Thus, following a negative shock affecting the assets in the balance sheet, banks need to restore their Tier 1 and Tier 2 capital (in short, their "top tier capital") and leverage ratios. Since they cannot raise fresh capital in the short-run, they cash-in by selling some of their assets. These sales create a vicious feedback loop where banks sell assets,

contributing to a further drop in the assets value, triggering further sales into a depressed market. Sometimes, this loop leads to a complete market dry-up. The more severe capital market frictions are, the more likely it is that a liquidity shock can turn into spots of liquidity black holes.

Even in the less extreme situations, in the equilibria of these markets, an initial small loss in the banking system is amplified, to an extent possibly leading to a very substantial lending shrinkage, a credit crunch. Section 14.5.3 discusses the monetary policy implemented in an attempt to mitigate the credit crunch originating from the 2007 subprime crisis. The standard policy action against a recession is to target low interest rates in the interbank markets for mandatory reserves. However, the cost of capital that matters to a recovery in the economic activity is that faced by firms whilst demanding new funds to banks (through loans) and/or the market (through issuance of corporate bonds). This cost can be substantially higher than the interest rates targeted by the monetary authority, due to the credit crunch. "Quantitative easing" is an unconventional policy action, where the monetary authority engages into the purchase of some of the assets held by banks (including the most illiquid ones), so as to give banks incentives to start lending again. Section 14.5.4 provides a discussion regarding the extent to which this unconvential policy may lead to bubbles.

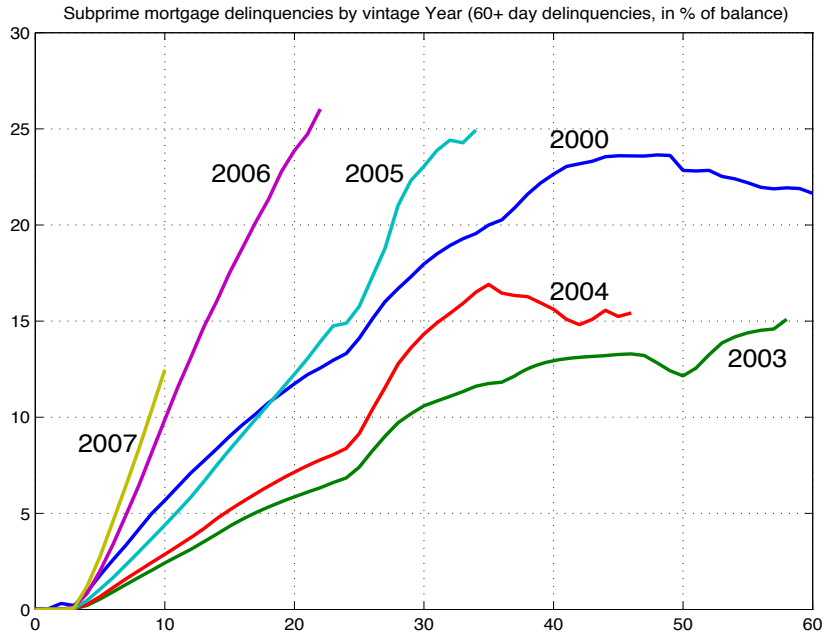
14.5.1 The 2007 subprime crisis

Underlying the 2007 subprime crisis is an explosive mixture of a number of factors. One factor relates to institutional details: MBS securities were traded through a "shadow banking system" that escaped the official financial community, including supervisors. The microstructure architecture underlying this trading system implied a high asset-liability duration mismatch and, accordingly, liquidity risks, carrying a potential for market failures. The exposure of financial institutions to these risks was not properly understood in the marketplace.

Model misspecification is a second factor. There is substantial evidence of moral hazard regarding mortgage origination[24] and final investors may have inadvertently had acquired more exposure than desired in this space. Concrete instances of model misspecification are visible in an evaluation framework that incorporated unrealistic assumptions such as stability of delinquencies, reliance on expected (linear) actuarial losses (not tail risk losses), or generous risk-aversion and liquidity adjustments. For example, the picture below shows delinquencies were actually traveling fast over the relatively newly created subprime mortgages; in retrospect, this of piece information could have helped predict an imminent crisis in the mortgage market. Another instance of model misspecification was a reliance on an inappropriate rating "mapping" system: rating agencies tended to rate structured products based on MBS, by using rating systems previously designed to assess corporation risks. Finally, correlation of default risks were underestimated: a deterioration in economic conditions leads to more clustered defaults.

This section attempts at a very succinct account of the subprime crisis while discussing the previous elements. The crisis erupted while MBS derivatives were producing losses in a market where the identity of the players was unclear as these products had been channelled through the off-balance-sheet vehicles of the shadow banking system.

[24] See Keys, Mukherjee, Seru, and Vig (2010).



Source: IMF, Global Financial Stability Report, April 2008.

14.5.1.1 Off-balance-sheet entities: "SIV," "conduits," and "SIV-lites"

Asset backed securities are traded through entities that it is useful to describe. A preliminary distinction is order. A Special Purpose Vehicle (or "SPV") is an entity that organizes the initial securitization of an asset pool. A Structured Investment Vehicle (or "SIV") is a *fund* that manages asset backed securities. SIV are virtual banks somehow, in that they borrow through low-interest securities and invest in longer term securities, thereby promising large rewards (in exchange for high risk), as further discussed below.

Cousins to SIVs are conduits. SIVs and conduits usually have both an open-ended lifespan. On the funding side, SIVs issue long-maturity notes; on the asset side, they invest into assets that are relatively more complex than those conduits are invested in. Moreover, SIVs tended to be more leveraged than conduits. At the extreme lie "SIV-lites," which are less conservatively managed than SIV and are structured with greater leverage. Moreover, their portfolios are not much diversified, and are much smaller in size than SIVs. SIV-lites had a finite lifespan, with a one-off issuance vehicle. They were greatly exposed to the U.S. subprime market, more so than SIVs.

Off-balance-sheet entities borrow in the shorter term, typically through commercial paper or auction rate securities with average maturity of 90 days, as well as medium term notes with average maturity of a year. They purchase long-maturity debt, such as financial corporate bonds or asset-backed securities, which is high-yielding. Naturally, the profits made by these entities are paid to the capital note holders, and the investment managers. The capital note holders are, of course, the first-loss investors.

The obvious risk incurred by these entities is *solvency*, a risk that materializes when the value of long-term assets falls below the value of short-term liabilities. This risk has great chances to materialize when the pricing of the assets is "informal," as argued below. A second risk

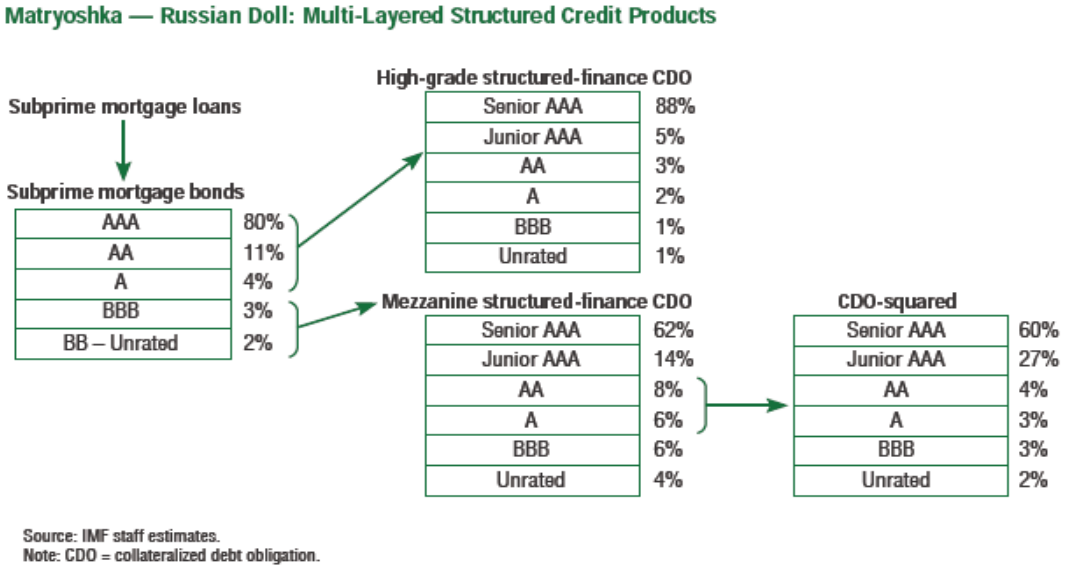
is *funding liquidity*, the risk related to duration mismatch: refinancing occurs on a short-term frequency, but if short-term market conditions are bad, the entities need to sell the assets into a depressed market. To cope with this risk, the sponsoring banks would grant credit lines.[25]

14.5.1.2 Credit ratings and model misspecification

The role of credit ratings was crucial in determining the riskiness of MBS-related derivatives (see Section 14.3.3). Credit ratings were endogenous in the securitization and tranching process: they would not be applied to an exogenously given tranching scheme; rather, they would be used to determine the riskiness of the tranching scheme.[26]

Another important point is that many of the MBS-based securitized assets were illiquid. Illiquidity does not facilitate pricing and the standard solution in these cases is to look for somehow similar products that are priced in relatively more liquid markets. Thus, the pricing of structured products would rely on the pricing of products that were similarly rated and for which quotes were available. For example, the price of AAA ABX subindices would be used to estimate the value of AAA-rated tranches of MBS. Or, the price of BBB subindices would be used to value BBB-rated MBS tranches. This is the "mapping role" credit ratings played for the pricing of customized or illiquid structured credit products.

Yet it is well-known that the risk profile of structured products differs from that of corporate bonds. Even if a tranche has the same expected loss as an otherwise similar corporate bond, *unexpected loss or tail risk can* be much larger than that for corporate bonds given the complexity of the product (see the Matryoska - Russian doll scheme below). All in all, it may be misleading to extrapolate structured products ratings from corporate bonds ratings. Typically, corporate bond ratings only capture the first moments of the distribution. Finally, credit rating inertia for bonds does not necessarily work for structured products. Rating deterioration for structured products can travel very fast.[27]



[25] Well-known sponsors at the time were Citibank ($100bn), JP Morgan Chase ($77bn), Bank of America ($60bn). In the European Union: HBOS ($42bn), ABN Amro ($40bn), HSBC ($32bn).

[26] After the crisis, some rating agencies adopted internal reforms to mitigate conflicts of interest arising during their process of interaction with originators. Amongst them, notable are those that aim to separate rating activities from sales, or the interdiction of agencies analysts to participate into the originators' business development activities.

[27] See IMF 2008 report.

Finally, correlation. Past data suggested a weak correlation between regional mortgages, which made investors perceive a sense of "diversification." However, the housing market grinding to a halt turned up to be a nation-wide phenomenon.

At the basis of all these technical complications was an erosion in lending standards: statistical models were based on historically low mortgage default and delinquency rates that used to arise in credit environments with tight credit standards. All in all, the products were "over-rated," that is, they appeared less risky than they were in reality. In this sense, these products were probably "toxic", in that they would not fit into many of the investors optimal composition of their risk-return trade-offs; or, these investors, if made aware of potential risks, would only require much lower doses of these products to be in their portfolios.

14.5.1.3 The meltdown

A crucial feature of the crisis was the market fear of contagion from the rising level of defaults in subprime underlying instruments, many of which were incorporated in complex products. Fears of contagion began to regard safer tranches too. They came from the investors' understanding the pricing models were misspecified, and their lack of trust vis-à-vis the rating agencies.

Banks were affected for a number of reasons: (i) they had invested in subprime securities directly; (ii) they had provided credit lines to SIV (indebted through commercial paper) and conduits that held these securities, thereby creating a shadow banking system, which escaped accounting and supervision rules; and (iii) this very same shadow system generated banks' loss of confidence in the ability of their counterparties to meet their contractual obligations. So the Asset Backed Commercial Paper market dried up, triggering credit lines. The result was a sell-off of anything related to structured finance, from junk to AAA, which led to a complete "liquidity black hole," and a severe reappraisal of structured finance.

The reappraisal of structured finance determined severe writedowns, also due to a "liquidity black hole" fueled by a difficult repricing. Indeed, in the absence of a liquid market, writedowns largely rely on marking-to-model. But investors begun not to trust the models and the rating process leading to them. Meanwhile, credit agencies proceeded to severe downgrades, confirming the investors' beliefs that previous ratings were based on misspecified assumptions, a quite self-reinforcing mechanism. These events escalated to a complete dry up in September-October 2008, partly restored by painful bank bail-outs and recapitalizations.

[In progress, explain Lehman's experience]

14.5.2 Top tier capital ratio targets and endogenous volatility

Treating volatility or credit risk as exogenous could be a good approximation whilst living in good times. The quality of this approximation deteriorates in times of crisis. The implicit assumption made in many instances of this and previous chapters, is that one's own actions, based on a volatility forecast, do not affect future volatility, just like forecasting weather does not influence future weather. Arguably, the actions of many heterogeneous market participants should tend to cancel with each other, during periods of "calm." However, market participants tend to cluster their decision rules in periods of crisis. The literature on this "endogenous risk" is quite fascinating, as are the surveys in Shin (2010), or the recent modeling framework put forward by Danielsson, Shin and Zigrand (2011).

This section develops a simple model of endogenous risk, where markets can be destabilized by one instance of procyclicality, arising because financial institutions need to comply with a given "top tier capital" (i.e. the capital comprising Tier 1 and Tier 2, using the terminology of

Section 14.5.1) ratio. After a negative shock in the value of the assets on the balance sheet, a financial instititution needs to restore its top tier capital ratio. In the short-run, the institution can only restore this ratio through asset sales. Because every institution is doing the same, these asset sales have a market impact, collectively, determining a further fall in the value of the assets, and so on. The final outcome is an increased volatility of the risky asset price, as well as a disproportionate assets sell-off, if compared to the initial shocks triggering it. The model is useful to think about the subprime events in 2007, as well as the ensuing credit crunch and the new solutions that monetary authorities have experimented to help mitigate these adverse developments, as we shall discuss.

14.5.2.1 Model

We consider a model with many identical financial institutions complying with regulation or, more generally, concerned with a pre-specified target of top tier capital ratio against risky assets. Each institution has the following balance sheet.

Balance sheet, Time 1	
A	E
C_b	D

The notation is a bit more elaborated than that used in Section 14.2: whilst we still define E as equity (including past retained earnings) and D as debt, we now define A as the value of risky assets, no matter how liquid these can be. Moreover, we define C_b to equal cash and reserves. We suppose the 8% Cooke ratio is in place, or $E \geq 0.08A$ and to simplify let $E = 0.08A$. Note that a top tier capital rule does not determine a leverage rule: there are, obviously, many leverage ratios, $\frac{D}{E}$, consistent with a given top tier capital ratio $\frac{E}{A}$. In fact, the new Basel III expicitly considers leverage ratios, thereby innovating upon Basel II, as discussed in Section 14.5.1. We shall deal with the procyclicality induced by this new rule later in this section.

Next, assume that some exogenous shock takes place, which makes the value of the risky assets decrease by some amount, ΔA, after which each institution would have the following balance sheet.

Balance sheet, Time 2	
$A - \Delta A$	$E - \Delta A$
C_b	D

But, each institution has to comply with its top tier ratio target. Therefore, at time 2, the new top tier capital, $E - \Delta A$, must be at least 8% of the risky assets, $A - \Delta A$, or,

$$E - \Delta A = 0.08 \left(A - \Delta A \right). \tag{14.65}$$

At time 1, the financial institution had set $E = 0.08A$. Therefore, Eq. (14.65) cannot hold, as a simple computation reveals. The intuition behind this impossibility is simple. As the value of the risky assets falls, the value of equity falls by a larger percentage than that of the risky assets, due to leverage—the value of risky assets falls by $-\frac{\Delta A}{A}$, whereas the value of equity falls by a larger percentage, $-\frac{\Delta A}{E}$, due to $E < A$.

Two solutions are available to the financial institution: (i) to inject fresh capital; (ii) to sell some of the risky assets. The first solution is not quite viable in the short-run. Let us analyze the second solution. We are looking for some quantity of the risky asset to sell, such that the reduction in value of the assets, say X^{s}, is able to meet the top tier capital ratio target. In terms of the balance sheet, we have the following situation.

Balance sheet, Time 3	
$A - \Delta A - X^s$	$E - \Delta A$
$C_b + X^s$	D

How much of the risky asset value should the financial institution precisely get rid off? To maintain the new top tier ratio, X^s must satisfy:

$$E - \Delta A = 0.08 \left(A - \Delta A - X^s \right).$$

Using $E = 0.08A$, and solving for X^s yields,

$$X^s = \frac{1 - 8\%}{8\%} \Delta A = 11.5 \cdot \Delta A.$$

That is, roughly, the number of risky assets to sell is proportional to the percentage loss in their value. In general,

$$\frac{X^s}{A} = \frac{1 - \kappa}{\kappa} \frac{\Delta A}{A}, \quad \kappa \equiv \frac{E}{A}, \tag{14.66}$$

where κ denotes the top tier capital ratio against risky assets. This result is intuitive: following a negative shock affecting the value of the risky assets, the amount of asset sales is decreasing in the pre-specified top tier capital ratio, κ, because the closer κ is to 100%, the easier the adjustment is to maintain the same κ. Eq. (14.66) would end the description of this market, if the financial institution had no price impact.

We assume that financial institutions have a market impact, collectively: while the behavior of one single institution cannot affect the price of the risky asset, many institutions doing the same thing at the same time more likely could, thereby creating price pressures, with the price of assets falling and triggering new sales into a depressed market, over a vicious feedback loop. Is there an equilibrium for this loop? The answer relies on the way we think of selling pressure. We model selling pressure by assuming that there is a continuum of financial institutions, and that the asset value changes according to:

$$\frac{\Delta A}{A} = \tilde{\epsilon} + \lambda \frac{X^s}{A}, \tag{14.67}$$

where λ is the "price impact," in the language of market microstructure (see Chapter 10). In words, the percentage change in the asset value is the product of two components: (i) the initial shock, $\tilde{\epsilon}$, and (ii) a selling pressure component, which will pump up endogeneous volatility, as we shall show. Naturally, if the the market was perfectly liquid, $\lambda = 0$. Moreover, in a world where financial institutions had no concerns about top tier capital rules, we would have $X^s = 0$ in the first place, and so:

$$\frac{\Delta A}{A} = \tilde{\epsilon}. \tag{14.68}$$

Note that this solution is also that arising when the market is perfectly liquid, $\lambda = 0$. However, we assume the existence of price pressure, as formalized by Eq. (14.67), determined by the concern financial institutions have to comply with a top tier capital rule, leading them to a sale $\frac{X^S}{A}$ satisfying Eq. (14.66). The loop we have created is, then, the following. After an initial shock affecting the risky asset value, $\tilde{\epsilon}$, financial institutions sell risky assets, to an extent proportional to the percentage change in the asset value, according to Eq. (14.66). In turn, the sell-off entails a further percentage drop in the asset value, as determined by Eq. (14.67). An equilibrium,

provided it exists, is a fixed point to this feedback-loop, i.e. a situation where the loop stops because the drops in the risky asset value do not happen anymore, and the assets sell-offs are interrupted as a result, thereby providing no reasons for the asset value to fall any further.

To find this fixed point, we replace Eq. (14.66) into Eq. (14.67) and solve for $\frac{\Delta A}{A}$, leaving:

$$\frac{\Delta A}{A} = c_A \cdot \tilde{\epsilon}, \quad c_A \equiv \frac{\kappa}{\kappa - \lambda(1-\kappa)}. \tag{14.69}$$

Note that in order for this equilibrium to exist, we need that λ, the slope of the line $\frac{X^s}{A} \mapsto \frac{\Delta A}{A}$ in Eq. (14.67), be less than $\frac{\kappa}{1-\kappa}$, the slope of the line $\frac{X^s}{A} \mapsto \frac{\Delta A}{A}$ in Eq. (14.66), or that $\lambda(1-\kappa) < \kappa$. Intuitively, if the price impact λ was too large, the feedback from asset sales to the asset value drops would create a perverse spiral such that the market would collapse.

Eq. (14.69) shows, crucially, that the shock multiplier, $c_A > 1$. When financial institutions have a concern over top tier capital ratios, the ultimate change in the asset value resulting from an initial shock, is larger than that we would have observed otherwise, say in Eq. (14.68). For example, assuming a price impact $\lambda = 0.05$ implies a multiplier $c_A = 2.4$. Naturally, these effects become less important as the market becomes more liquid, and do not matter anymore in the limit case where the market is perfectly liquid, $\lambda = 0$.

What is the amount of asset sales resulting from this loop? Replacing Eq. (14.69) into Eq. (14.66) yields:

$$\frac{X^s}{A} = c_X \cdot \tilde{\epsilon}, \quad c_X \equiv \frac{1-\kappa}{\kappa - \lambda(1-\kappa)}. \tag{14.70}$$

Note that the feedback loop might extert quite substantial effects into the amount of asset sales. Assuming $\kappa = 0.08$, the multiplier c_X would equal 11.5 in the absence of feedback, $\lambda = 0$, as previously noted. The same multiplier more than doubles in the presence of feedbacks and a price impact of just $\lambda = 0.05$, attaining a value $c_X = 27.1$.

This model thus formalizes the idea that even a small shock affecting the risky assets held by financial institutions might lead to large sale adjustments and price corrections, similarly as for the developments inherent the subprime events described in the previous section. In the model, the concerns financial institutions have about top tier capital ratios leads them to substantial asset sales in response to a shock, which are even more amplifed in the presence of feedback effects induced by liquidity frictions.

14.5.2.2 Multiple equilibria and market break-ups

In the model we analyze, an equilibrium exists under parameter restrictions that are independent of the realization of the initial shock, $\tilde{\epsilon}$. As noted, we simply need that the denominators of the multipliers c_A and c_X in Eqs. (14.69) and (14.70) be strictly positive. We now present a variant of the model, where an equilibrium fails to exist when the initial shock $\tilde{\epsilon}$ is sufficiently large. We simply assume that the price pressure is nonlinear, differently from the linearity assumption underlying Eq. (14.67). It is:

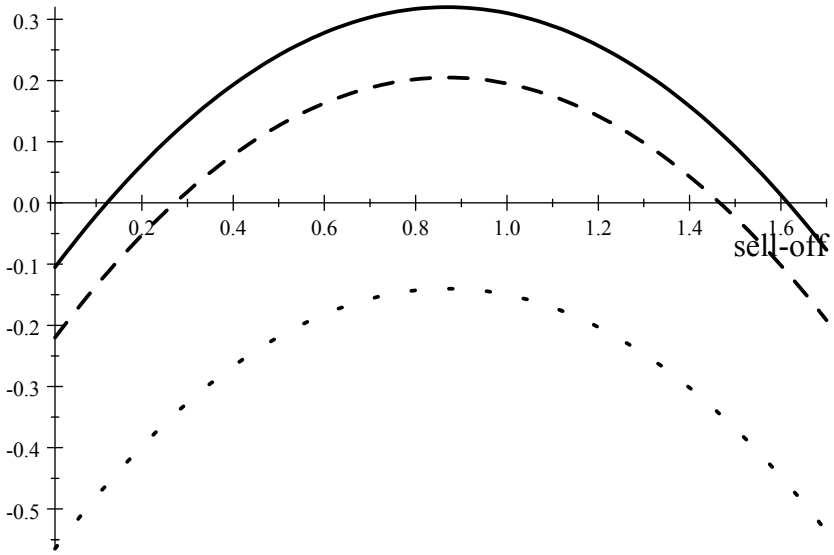
$$\frac{\Delta A}{A} = \tilde{\epsilon} + \lambda \left(\frac{X^s}{A}\right)^2. \tag{14.71}$$

The quadratic term in Eq. (14.71) formalizes the idea that the price impact of asset sales does not matter too much when the asset sales are limited, but becomes disproportionately high when the asset sales are at a large scale. This convexity translates into a non-linearity of the

resulting feedback loop. Replacing Eq. (14.71) into Eq. (14.66), we find that in any equilibrium, the amount of asset sales is solution to the following quadratic equation

$$0 = \varphi\left(\frac{X^{\mathrm{S}}}{A}\right) \equiv \frac{X^{\mathrm{S}}}{A} - \frac{1-\kappa}{\kappa}\lambda\left(\frac{X^{\mathrm{S}}}{A}\right)^2 - \frac{1-\kappa}{\kappa}\tilde{\epsilon}. \tag{14.72}$$

We hypothesize the market is hit by a series of shocks. Initially, we assume that $\tilde{\epsilon} = 0$, such that two equilibria are possible: one, where the sell-off is just zero, $\frac{X^{\mathrm{S}}}{A} = 0$; and another, where the sell-off is $\frac{\kappa}{\lambda(1-\kappa)}$. We assume the sell-off is zero. Then, we assume that a first positive shock hits the market, with $\tilde{\epsilon} = 1\%$. The solid line in the next picture is the graph of $\varphi\left(\frac{X^{\mathrm{S}}}{A}\right)$ in this case. There are two equilibria, corresponding to $\frac{X^{\mathrm{S}}}{A} : \varphi\left(\frac{X^{\mathrm{S}}}{A}\right) = 0$. We assume that the market coordinates towards the leftmost one.



The three curves depict the graph of $\varphi\left(\frac{X^{\mathrm{S}}}{A}\right)$ in Eq. (14.72), when the initial shock is $\tilde{\epsilon} = 1\%$ (solid line), $\tilde{\epsilon} = 2\%$ (dashed line), and $\tilde{\epsilon} = 5\%$ (dotted line), obtained assuming a top tier capital ratio against risky assets $\kappa = 8\%$, and a price impact $\lambda = 0.05$. The equilibrium sales are those where the curves intersect the horizontal axis, if any.

As the risky asset value is hit by one additional shock, say with $\tilde{\epsilon} = 2\%$, the graph of $\varphi\left(\frac{X^{\mathrm{S}}}{A}\right)$ shifts to South-East, to the dashed line, and the asset sales increase as a result, still being the leftmost zero of $\varphi\left(\frac{X^{\mathrm{S}}}{A}\right)$. The market collapses when the shock is $\tilde{\epsilon} = 5\%$, leading the graph of $\varphi\left(\frac{X^{\mathrm{S}}}{A}\right)$ to a further shift to South-East, where no equilibria are left at all.

14.5.2.3 Deleveraging

So far, we have assumed that, following a shock affecting the assets in the balance sheets, financial institutions sell additional assets to the extent their top tier capital ratios are restored. This section investigates additional adjustments that aim to comply with leverage ratio targets. Denote the leverage ratio as $L \equiv \frac{D}{E}$. The timing of the shock and banks reaction is as in Section 14.5.3.1, with the exception that we now have one additional constraint: at time-3, financial

institutions also call portions of their debt, X^{l} say, so as to comply with leverage ratio targets, and achieve the following balance sheet.

Balance sheet, Time 3
under a deleveraging scenario with deep liquidity buffers

$A-\Delta A-X^{\mathrm{s}}$	$E-\Delta A$
$C_b+X^{\mathrm{s}}-X^{\mathrm{l}}$	$D-X^{\mathrm{l}}$

The term X^{s} is the usual sell-off needed to maintain top tier capital targets. Now, financial institutions also target to deleverage for an amount equal to $X^{\mathrm{l}}: \frac{D-X^{\mathrm{l}}}{E-\Delta A}=L$, or

$$\frac{X^{\mathrm{l}}}{A}=L\cdot\frac{\Delta A}{A}, \tag{14.73}$$

and we initially assume that while doing so, they do not exaust their liquidity buffers,

$$C_b+X^{\mathrm{s}}-X^{\mathrm{l}}\geq 0. \tag{14.74}$$

Note that if (14.74) holds true, deleveraging does not have a price impact because it would imply banks are simply using cash to repay portions of their debt. In this case, X^{s} is the same as that in Eq. (14.70). However, should C_b+X^{s} be not large enough, financial institutions would have to sell additional assets, ΔX^{s} say, so as to have sufficient cash and meet leverage ratio targets. Precisely, if the inequality in (14.74) does not hold, we need to have that, the asset sales ΔX^{s} leave the bank with some positive liquidity buffers B,

$$\Delta X^{\mathrm{s}}: B\equiv X^{\mathrm{s}}+\Delta X^{\mathrm{s}}+C_b-X^{\mathrm{l}}\geq 0, \tag{14.75}$$

such that the balance sheet that financial institutions face at time-3 would be

Balance sheet, Time 3
under a deleveraging scenario leading to exhausting liquidity buffers

$A-\Delta A-X^{\mathrm{s}}-\Delta X^{\mathrm{s}}$	$E-\Delta A$
$C_b+X^{\mathrm{s}}-X^{\mathrm{l}}+\Delta X^{\mathrm{s}}$	$D-X^{\mathrm{l}}$

To determine the feedback effects of the asset sales, X^{s} and ΔX^{s}, assume that the bank keeps on using the top tier capital ratio condition in (14.66).[28] Replacing Eq. (14.66) and the leverage condition, Eq. (14.73), into Eq. (14.75), and setting for simplicity $B=0$, leaves

$$\frac{\Delta X^{\mathrm{s}}}{A}=\frac{(1+L)\,\kappa-1}{\kappa}\frac{\Delta A}{A}-\frac{C_b}{A}. \tag{14.76}$$

This expression is positive because, by assumption, while deleveraging, banks exhaust their liquidity buffers to the extent (14.74) does not hold. This case arises precisely when leverage L is high enough. For example, assuming $\kappa=0.08$, the loading for $\frac{\Delta A}{A}$ in Eq. (14.76) is positive only when $L>11.5$. We would need to observe values of L larger than 11.5 (and state-dependent, i.e. depending on the realization of $\frac{\Delta A}{A}$), for the condition in (14.74) to break down.

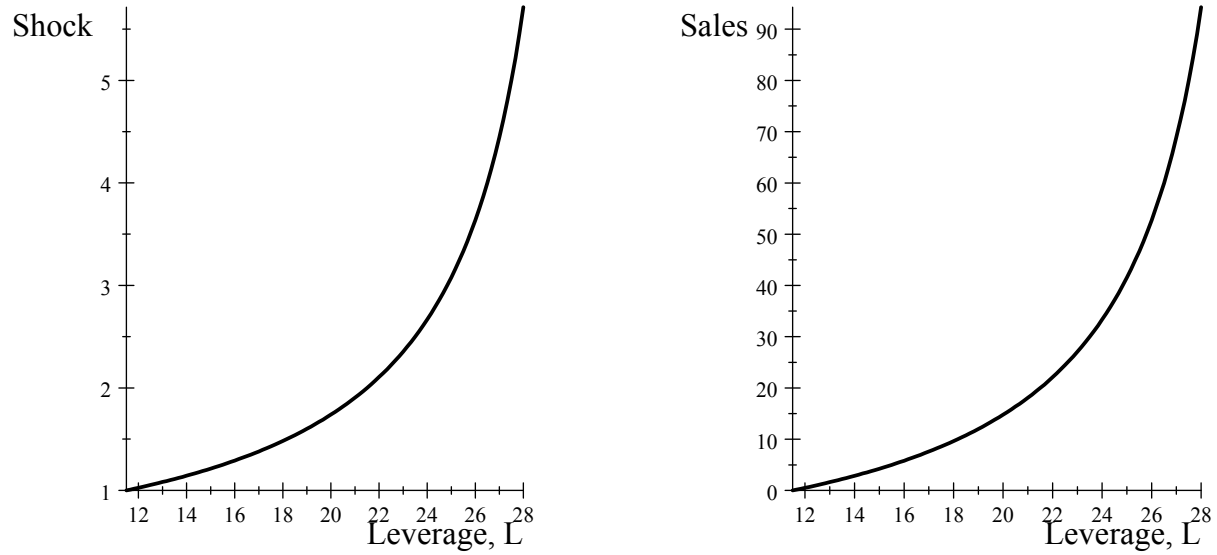
We determine an equilibrium for the feedback loop in this market. Replace Eq. (14.76) into Eq. (14.67), evaluated when the asset sales amount to ΔX^{s}, i.e. $\frac{\Delta A}{A}=\tilde{\epsilon}+\lambda\frac{\Delta X^{\mathrm{s}}}{A}$, and solve for both the asset value drop and sales,

$$\frac{\Delta A}{A}=c_{L,a}\cdot\tilde{\epsilon}-\lambda c_{L,a}\cdot\frac{C_b}{A},\quad \frac{\Delta X^{\mathrm{s}}}{A}=c_{L,b}\cdot\tilde{\epsilon}-\frac{\left((1+L)\,\kappa-1\right)\lambda c_{L,a}+\kappa}{\kappa}\cdot\frac{C_b}{A}, \tag{14.77}$$

[28] In this case, the top-tier condition will a fortiori hold with $\Delta X^{\mathrm{s}}>0$.

and the constants $c_{L,a} \equiv \frac{\kappa}{\kappa-\lambda((1+L)\kappa-1)}$ and $c_{L,b} \equiv \frac{(1+L)\kappa-1}{\kappa-\lambda((1+L)\kappa-1)}$. To ensure an equilibrium exists, we need that λ, the slope of the line $\frac{X^s}{A} \mapsto \frac{\Delta A}{A}$ in Eq. (14.67), be less than $\frac{\kappa}{(1+L)\kappa-1}$, the slope of the line $\frac{X^s}{A} \mapsto \frac{\Delta A}{A}$ in Eq. (14.76), or that the denominators of $c_{L,a}$ and $c_{L,b}$ be strictly positive.

Finally, note that there is a third possibility available to banks: to sell additional assets even when the condition in (14.74) holds. This possibility is relevant when financial institutions worry about maintaining liquidity buffers (some may be mandatory). For example, if the liquidity target $B = C_b$, we replace the inequality in (14.74) with $X^s - X^l \geq 0$. The solutions for the asset value drop and sales are as in Eqs. (14.77), but with the terms involving C_b dropped, i.e., $\frac{\Delta A}{A} = c_{L,a} \cdot \tilde{\epsilon}$, and $\frac{\Delta X^s}{A} = c_{L,b} \cdot \tilde{\epsilon}$. The next picture depicts the two multipliers of $\tilde{\epsilon}$ for $\frac{\Delta A}{A}$ and $\frac{\Delta X^s}{A}$ arising in this case, assuming the top tier capital ratio target is $\kappa = 0.08$, and the price impact of asset sales is $\lambda = 0.05$.



This picture depicts the graph the two multipliers, $c_{L,a}$ and $c_{L,b}$ in Eqs. (14.77), as a function of banks' leverage, L, when the top tier capital ratio is $\kappa = 0.08$, and the price impact of banks sell-offs is $\lambda = 0.05$. When banks have concerns regarding maintaining cash and reserves to their level before the shock, C_b, the overall asset value drop and sales equal $\frac{\Delta A}{A} = c_{L,a} \cdot \tilde{\epsilon}$ and $\frac{\Delta X^s}{A} = c_{L,b} \cdot \tilde{\epsilon}$. The left panel depicts the graph of $c_{L,a}$ and the right panel depicts that of $c_{L,b}$.

First, note that given the assumption that $\kappa = 0.08$, the relevant range of variation for leverage is that for $L \geq 11.5$, as we are studying situations where deleveraging leads to partial exhaustion of liquidity buffers. The effects can be quite substantial.

14.5.3 Credit crunches and quantitative easing

[Show pictures of (i) the balance sheet of FED, ECB and BoE over the last years. (ii) the FED fund rate, and EONIA, (iii) the corporate spreads]

14.5.3.1 The money multiplier

Loans make deposits! The well-known mechanics underlying the creation of money relies on the standard money multiplier, whereby new deposits made available to the banking system are

partially used to extend new loans, which generate further deposits, and so on. Mathematically, the supply of money, say M1 aggregates, includes cash held by the public plus deposits, $M = C_p + D$, with straight forward notation. Instead, the monetary base, or "high potential" money, is made of cash held by the public plus banks cash and reserves, $H = C_p + C_b$, where C_b denotes cash and reserves held by banks, as in the previous sections. Note the leakage banks create over the circuit of money creation: because banks face liquidity needs possibly arising in the short-term (vis-à-vis their clients and/or other banks), and possibly need to maintain mandatory reserves with the central banks of the countries where they operate, they hoard C_b, which escapes the loans-make-deposits loop.

We assume that the ratios $\frac{C_p}{D}$ and $\frac{C_b}{D}$ are constant and equal to c_p and θ, respectively, such that money supply equals a money multiplier times monetary base, viz

$$M = \frac{1 + c_p}{c_p + \theta} \cdot H. \tag{14.78}$$

The value of θ depends on a variety of factors, such as: (i) the discount rate at which a central bank lends money to banks; (ii) the interbank rates such as the LIBOR; (iii) the level of mandatory reserves banks have to keep with their central banks; (iv) the interbank rate for resources to allocate to mandatory reserves—the Federal Funds rate in the US; (v) the risk/return trade-off prevailaing in other markets. Clearly, the value of θ increases with the values of the items from (i) to (iv), and decreases when the trade-off in (v) compares more favourably to banks.

14.5.3.2 Policy actions

Conventional policy strategies consist of actions aiming to affect the value of θ and H. For example, central banks can expand H through open market operations, by purchasing short-term Government bonds. Note that this action likely affects θ as well, as the opportunity costs of holding excess reserves increase as markets are flooded with more and more liquidity. There are, obviously, limits to this action, arising when short-term interest rates get close to zero. Consider the 2007 subprime events. We know that following a shock affecting the banks' books, the overall adjustments can be quite substantial. Consider, for example, the model in the previous section, where banks have concerns over the top tier capital ratio. After the shock takes place, banks aim to a shrinkage in the asset value equal to X^{s}. Moreover, the shrinkage can be even more substantial, $X^{\mathrm{s}} + \Delta X^{\mathrm{s}}$, in markets where banks are high levered, and concerned about not increasing their leverage even more as a result of the shock. The model is silent about which particular assets (liquid or not) or loans are involved into the shrinkage plans, X^{s} or $X^{\mathrm{s}} + \Delta X^{\mathrm{s}}$. We define a "credit crunch" as the situation where banks decide to cut on corporate loans and bonds—they hold more reserves, instead of lending money to the real sector.

A quite mechanical response to a credit crunch is to increase H through, say, open market operations. Put simply, a credit crunch entails a higher value of θ, among other things. Monetary aggregates, M, are destroyed as a result, but can be restored through an injection of high potential money. Precisely, by Eq. (14.78), the expansion of monetary base needed to maintain the same supply of money, M, is $\Delta H = \frac{M}{1+c_p}\Delta\theta$, where $\Delta\theta$ is the increase in θ determined by the credit crunch. This policy action is quite fundamental as it helps keep interest rates low, yet it may not be enough when the credit crunch is so particularly severe to lead to very substantial shrinkages in the economic activity, as we now explain.

The effects of a credit crunch on the real sector of the economy are quite obvious, with an increase in the cost of capital and a subsequent shrinkage in the economic activity. For example,

the recession following the subprime events was spectacular, with industrial production falling by approximately 13% on a yearly basis in March 2009, the highest drop since World War II. The policy action was equally impressive: in less than two years after the subprime events, the FED was capable of pushing short-term rates close to zero although during those periods, it was already clear that this policy would not be likely to prevent an even deeper recession. All in all, even if the Federal Funds rate and short-term rates on safe assets were close to zero, the cost of capital firms had to bear were quite substantial as a result of the credit crunch.[29] Note that at that time, the credit crunch was also exasperated by a freeze in the interbank lending market, arising from concerns financial institutions had about counterparty risk.[30]

The events following the 2007 turmoil can be described as those of a "liquidity trap," where banks hold abundant liquidity and short-term rates are close to zero. Note that the nature of this liquidity trap is different from the standard Keynesian liquidity trap, as formalized in the Appendix of Chapter 1: the Keynesian trap arises when money demand is flat as a result of the expectations investors have that future interest rates can only increase. In this case, agents simply absorb any liquidity injections made by the monetary authority, and interest rates remain "trapped" at some minimum rate, coinciding with the lowest, shadow interest rate beyond which no investor is ready to bet against a decrease. The liquidity trap we are analyzing is different, and stems from the mere and mechanical circumstance that money supply is so abundant to have made short-term rates close to zero in the first place. However, in both cases, the economy is trapped in that a further increase in the monetary base would have no effects on short-term interest rates.

What was the initial policy reaction to this liquidity trap? Note that a further issue arising within the case we analyze in this section, is that the liquidity trap is accompanied by a surge in corporate spreads, as a result of the credit crunch. "Quantitative easing" (QE, henceforth) is a policy action that aims to restore the shrinkage in credit supply, and possibly reduce these corporate spreads, so as to mitigate the adverse effects of the credit crunch on to the real economic activity. Consider the following balance sheets. The balance sheet on the left-hand side is that arising after financial institutions have reacted to a shock in their asset value, as formalized in previous sections. The portion of X^{s} that includes corporate loans and bonds is what we are terming credit crunch.

Balance sheets, Time 3 and beyond

$A-\Delta A-X^{\mathrm{s}}$	$E-\Delta A$	$\rightarrow$	$A-\Delta A-X^{\mathrm{s}}+Q^{\mathrm{s}}$	$E-\Delta A$
C_b+X^{s}	D	$\leftarrow$	$C_b+X^{\mathrm{s}}-Q^{\mathrm{s}}$	D

The effects of quantitative easing can be seen through the balance sheet on the right-hand side. Banks can purchase new corporate bonds or extend new loans (securitized loans), which the central bank immediately purchases, leading to the "left arrow" of the previous diagram. The ideal, spontaneous, resolution of a credit crunch is when financial institutions are willing to extend corporate loans or purchase corporate bonds, to restore their credit shrinkage, at least

[29] A high cost of capital to firms might occur during a credit crunch, as a result of one additional effect: a credit crunch leads to a contraction in aggregate demand, which makes defaults more likely.

[30] A shrinkage in the economy activity suggests a natural extension of the models in the previous section, with one additional element of procyclicality: as the real economy plummets as a result of the credit crunch, the value of the assets decreases even more, thereby deepening the credit crunch, over a vicious feedback loop. Note that this feedback mechanism would be, quite naturally, part of the financial accelerator hypothesis, although distinct from the mechanism mentioned at the beginning of this section, and surveyed in Chapter 9 (see Section 9.14). In the version we suggest, the credit crunch is determined by concerns financial intermediaries have about their top tier capital ratios and leverage exposures, rather than agency problems occurring over their relationships with clients. We do not examine this additional source of procyclicality to keep the analysis as simple as we can.

partially, by some amount Q^s. However, we know this resolution cannot occur until the value of the assets remains depressed. However, the central bank could step in to purchase the assets the banking system is disliking, to an extent equal to at least Q^s, so as to leave banks with the excess reserves they wish and comply with their top tier capital ratio targets. This action is the essence of quantitative easing, with assets typically involved being ABS and even long-term bonds. Its effects include both (i) an increase in H, equal to the extent of the liquidity injection, and (ii) higher incentives given to banks to refinance the real sector, due to the liquidity buffers supplied by the central bank.

14.5.4 Where did QE go?

Plot P/E ratios over the years from 2006 to 2016, to analyze the effects of QE, in the US, Europe and the UK. Take Germany as representative of Europe. Developments in the P/E and the real economy provide a sense to the reasoning. So also plot (i) Treasuries, and MBS purchased + asset purchase facilities by the FED, UK and ECB (Balance sheet proportions) (ii) Target FED fund rate (U.S.), interest rate on MRO (main refinancing operations) (Euro area) and Bank of England base rate (U.K.) (iii) P/E ratios in the U.S., in the U.K. and in Germany; (iv) Industrial production in the U.S., in the U.K. and in Germany; (v) Corporate spreads in the U.S., in the U.K. and in Germany.

14.5.4.1 Bubbles and crises

These lectures have dealt with various frameworks that help rationalize the existence of bubbles (see, e.g., Chapters 3, 5 and 9), but none of these frameworks explains that bubbles may actually link to the mere availability of credit to investors with limited liability. Yet we know from Chapter 9 (see Section 9.14) that, when allowed to default on debt obligations, investors are incentivized to take more risks than they would without the possibility of bankruptcy opened to them. But wouldn't credit availability, then, lead investors to price assets above their fundamentals? Allen and Gale (2007) survey a number of models that formalize these ideas.

The economic rationale behind these models captures a few traits of some relatively famous bubbles such as the Japanese stock and real estate bubbles of the 1980s-1990s or the South-East Asia bubbles of the 1990s. Typically, a period of credit expansion (fueled, for example, by financial liberalization) leads investors to take on risks that they would certainly not take were they fully liable of their investment decisions. The mechanism is standard, and is convexity: limited liability leads investors to be interested in the upper part of the payoffs distribution because the risks regarding the lower part are shifted to the lenders. Such a *risk-shifting* (from investors to banks) may be responsible of bubbles, with investors indebted at relatively cheap conditions and, thus, incentivized to take positions in assets overvalued and in finite supply. Eventually, a crisis erupts in bad times, i.e., when the investments performance is poor, and indebted investors default on the debt they had used to buy overvalued assets. These defaults can, eventually, lead to banking crises, with possible spillovers into the real economy. These mechanisms can also be part of the explanations regarding credit market developments in the late 2000s (see, e.g., Reinhart and Rogoff, 2009). Relatively more controversial seems to be the question whether these ideas help explain market developments subsequent to QE.

[Provide comments to the previous plots]

The deliberate scope of QE is to increase lending. But abundant credit availability also creates incentives to invest in risky activities more than it would be the case without the possibility opened to investors to default on their debt. In the next section, we shall actually argue that

these incentives increase as investments become riskier. At least theoretically, QE may then detract resources from the original targets set by central banks, and lead to bubbles. We now present a model along the lines of the work of Allen and Gale to illustrate how bubbles arise as a result of risk-shifting mechanisms.

14.5.4.2 Models of credit and bubbles

We consider a market for a risky security in fixed supply. First, there is a safe, fixed income security, in that investing F into it delivers $F(1+r)$ after one period, with $r \geq 0$. The risky security (say, real estate or stocks) pays off $\tilde{x} = x_u$ in the good state with probability p and $\tilde{x} = x_l$ with complementary probability, with $\frac{1}{S}x_l < 1 + r < \frac{1}{S}x_u$, where S is the price of this security. Investors are risk-neutral, and they can borrow £1 to purchase $\frac{1}{S}$ units of the risky security. Their total expected returns are $\frac{1}{S}E(\tilde{x}) - (1+r)$, where E is the expectation under p. Due to risk-neutrality, it must be that this total expected return is zero, such that the price of the risky security is $S_{\mathrm{F}} = \frac{E(\tilde{x})}{1+r}$. We refer to S_{F} as the *fundamental value* of the risky asset.

Next, consider a friction, notably that borrowers can default on their obligations. We also assume that borrowing and lending rates may differ, with borrowing rates lower than lending rates: borrowers can borrow £1 at an interest rate $r_0 \leq r$, as a result of a credit expansion or, say, a non-standard monetary policy, QE. We assume that $r - r_0 = \sigma r_0$, where $\sigma \geq 0$ is a spread parameter linked to "capital competition": the more fierce competition banks face in acquiring funds, the lower σ and in the limit, $\sigma = 0$. We actually assume that banks can also invest into fixed income securities: they may also finance at r_0 from other banks and invest at r, achieving in this way an "intermediation margin" equal to σr_0. In this case, one can also think of σ approaching zero as being the result of a "rate curve compression" arising by leveraged purchases of securities fueled by resources made available through QE.

We assume that banks have not the ability to invest directly into risky projects; only borrowers can do so. This market "specialization" typically occurs in the case of real estate investments, and may actually result by bans on proprietary trading by commercial banks, as with the "Volcker rule" (i.e., Section 619 of the Dodd-Frank Wall Street Reform and Consumer Protection Act, 2010). Finally, we assume that a collateral C is required to back any lending, and that this collateral is small, i.e., $\frac{1}{S}x_l + C < 1 + r_0$. This low collateral condition simply means that, for a given price S, a borrower long in the risky security does default on his obligations in the bad state of nature.

How do these new assumptions affect asset evaluation? To begin, assume that $\sigma = 0$; in this case, the total expected return from borrowing £1 and investing it into the risky asset satisfies

$$E(\tilde{\pi}; S) \equiv E\left(\max\left\{\frac{1}{S}\tilde{x} - (1+r_0), -C\right\}\right) = 0, \tag{14.79}$$

due to risk-neutrality. An equilibrium price S^* satisfies $E(\tilde{\pi}; S^*) = 0$. But by Jensen's inequality, $E(\tilde{\pi}; S_{\mathrm{F}}) \geq 0$, such that $S^* \geq S_{\mathrm{F}}$. The difference, $S^* - S_{\mathrm{F}}$, is a *bubble*. Moreover, it is easily seen that (due to convexity), S increases after a mean-preserving spread in $\tilde{x}$: the riskier the asset, the larger the bubble. We now characterize this bubble.

Consider again the general case $\sigma \geq 0$. The "carry trade" profits from investing in the fixed income asset are $\pi \equiv \frac{1}{F} - (1+r_0) = \sigma r_0$, such that the wedge between the expected returns from investing in the risky security and in the riskless is

$$\phi(S) \equiv E(\tilde{\pi}; S) - \pi = p\left(\frac{1}{S}x_u - (1+r_0)\right) + (1-p)(-C) - \sigma r_0.$$

The equilibrium price satisfies $S^* : \phi(S^*) = 0$. Any price $S < S^*$ is not an equilibrium, in that investors would then only buy the risky security, bidding up the price to S^* due to the asset being in finite supply. A bubble arises whenever $S^* > S_{\mathrm{F}}$. Solving for the equilibrium price leaves

$$S^* = \frac{px_u}{p(1+r_0) + (1-p)C + \sigma r_0}. \tag{14.80}$$

A bubble arises under parameter restrictions. Assume, for example, that $x_l = 0$ and $r_0 = r$ (or $\sigma = 0$), in which case,

$$S^* = S_{\mathrm{F}} \cdot \mathcal{B}, \quad \mathcal{B} \equiv \frac{1+r_0}{p(1+r_0) + (1-p)C} > 1,$$

where the inequality follows by the assumptions on collateral requirements, $C < 1 + r_0$.

Note that an "ultra-low interest rate environment" does not necessary lead to bubbles in this model. Bubbles arise because credit facilities are made available to investors who can possibly default. In fact, in the previous example, $\mathcal{B}$ is actually increasing in r_0, although then both S_{F} and S^* are decreasing in r_0.[31] Their existence in this model does actually represent a true "distortion," in that the possibility of defaulting on a credit expansion leads the asset price to not factor in all risks (notably, x_l). In fact, bubbles can be mitigated while increasing collateral requirements, C. Or they may even not exist when both C and x_l are sufficiently large. But they do arise otherwise, as in this example.

[31] The bubble can be decreasing in r_0, when the latter is determined endogenously, as in the following example (see, e.g., Allen and Gale, 2007, p. 243). Assume that the central bank makes available a level of credit $\mathcal{C}$, which is used in the economy for investments in a productive activity (x) and the risky asset (S), such that $\mathcal{C} = x + S^*$. The interest rate $1 + r_0$ equals the marginal product of capital $f'(x)$, where $f(x) = 3\sqrt{x}$. Thus, $1 + r_0 = \frac{3}{2\sqrt{\mathcal{C} - S^*}}$, which replaced into Eq. (14.80) leaves a nonlinear equation for S^*, which is solved in closed-form once $C = 0$.

14.6 Appendix 1: Details on strategic defaulting

Present values contingent on future bankruptcy. The value of debt in Leland's (1994) model can be written as:

$$D(A) = \mathbb{E}\left(\int_0^{T_B} e^{-rs} C ds\right) + \mathbb{E}\left(e^{-rT_B}(1-\alpha) A^B\right), \tag{14A.1}$$

where T_B is the time at which the firm is liquidated. Eq. (14A.1) simply says that the value of debt equals the expected coupon payments plus the expected liquidation value of the bond. We have:

$$\mathbb{E}\left(e^{-rT_B}\right) = \int_0^{\infty} e^{-rt} f\left(t; A, A^B\right) dt \equiv p_B(A), \tag{14A.2}$$

where $f\left(t; A, A^B\right)$ denotes the density of the first passage time from A to A^B. It can be shown that $p_B(A)$ is exactly as in Eq. (14.12) of the main text. Similarly,

$$\begin{aligned}
\mathbb{E}\left(\int_0^{T_B} e^{-rs} C ds\right) &= C \cdot \mathbb{E}\left(\int_0^{T_B} e^{-rs} ds\right) \\
&= C \cdot \int_0^{\infty} \left(\int_0^{t} e^{-rs} ds\right) f\left(t; A, A^B\right) dt \\
&= C \cdot \int_0^{\infty} \frac{1-e^{-rt}}{r} f\left(t; A, A^B\right) dt \\
&= \frac{C}{r} \cdot \left(1 - p_B(A)\right).
\end{aligned} \tag{14A.3}$$

Replacing Eq. (14A.2)-(14A.3) into Eq. (14A.1) yields Eq. (14.11).

Equity value as a real option. By results in Chapter 5 (see Section 5.7), the equity value $E(A)$ in Eq. (14.15) is solution to

$$LE(A) + \delta A - (1-\theta) C - rE = 0 \tag{14A.4}$$

where $LE = \frac{1}{2}\sigma^2 A^2 E'' + (r-\delta) A E'$ denotes the infinitesimal generator, and

$$E\left(A^B\right) = 0, \quad E'\left(A^B\right) = 0. \tag{14A.5}$$

The first equality in Eqs. (14A.5) is a value matching condition (equity value is zero when the firm has defaulted, $A < A^B$, and at the very same moment it defaults). The second equality is a smooth pasting condition. The general solution to Eq. (14A.4) is

$$E(A) = -(1-\theta)\frac{C}{r} + A + k_1 A^{x_1} + k_2 A^{x_2}, \tag{14A.6}$$

where x_1 and x_2 are the two roots of $0 = \frac{1}{2}\sigma^2 x^2 + \left(r - \delta - \frac{1}{2}\sigma^2\right) x - r$, i.e.,

$$x_{1/2} = \frac{-\left(r - \delta - \frac{1}{2}\sigma^2\right) \mp \sqrt{\left(r - \delta - \frac{1}{2}\sigma^2\right)^2 + 2r\sigma^2}}{\sigma^2}.$$

We have $x_2 > 0$ and, accordingly, we set $k_2 = 0$ in (14A.6), meaning that the effects of the option to default weaken as the assets value grows large. It is equity analogue assumption to the boundary condition for the bond formulated in the main text, by which the bond value is $\frac{C}{r}$ when the assets value is so large to make the likelihood of bankruptcy negligible. To determine k_1 and A^B, we use the value matching condition and the smooth pasting condition in (14A.5), which leaves the solution in (14.13), with A^B as in (14.14) and $\gamma = -x_1$.

14.7 Appendix 2: Proof of selected results

Alternative derivation of Eq. (14.26). Under the risk-neutral probability, the expected change of any bond price must equal zero when the safe short-term rate is zero,

$$\frac{\partial B\left(t\right)}{\partial t}+\lambda\left(\text{Rec}-B\left(t\right)\right)=rB=0,\quad \text{with } B\left(T\right)=N,$$

where the first term, $\frac{\partial B(t)}{\partial t}$, reflects the change in the bond price arising from the mere passage of time, and $\lambda\left(\text{Rec}-B\left(t\right)\right)$ is the expected change in the bond price, arising from the event of default, i.e. the probability of a sudden default arrival, λ, times the consequent jump in the bond price, $\text{Rec}-B\left(t\right)$.

The solution to the previous equation is,

$$B\left(0\right)=\int_{0}^{T}\text{Rec}\cdot\underbrace{\lambda e^{-\lambda t}}_{=\Pr\{\text{Default at }t\}}dt+Ne^{-\lambda T},$$

which is Eq. (14.26).

Proof of Eq. (14.27). The spread is given by:

$$s\left(T\right)=-\frac{1}{T}\ln\left(\frac{\text{Rec}_{T}\left(1-e^{-\lambda T}\right)+Ne^{-\lambda T}}{N}\right).$$

With $N=1$, and $\text{Rec}_{T}=R\cdot e^{-\kappa T}$, we have,

$$s\left(T\right)=-\frac{1}{T}\ln\left(Re^{-\kappa T}\left(1-e^{-\lambda T}\right)+e^{-\lambda T}\right)=\lambda-\frac{1}{T}\ln\left(Re^{-(\kappa-\lambda)T}\left(1-e^{-\lambda T}\right)+1\right),$$

or equivalently,

$$s\left(T\right)=-\frac{1}{T}\ln\left(Re^{-\kappa T}\left(1-e^{-\lambda T}\right)+e^{-\lambda T}\right)=\kappa-\frac{1}{T}\ln\left(R\left(1-e^{-\lambda T}\right)+e^{-(\lambda-\kappa)T}\right),$$

Therefore, if $\kappa\geq\lambda$, then, $\lim_{T\to\infty}s\left(T\right)=\lambda$, and if $\kappa\leq\lambda$, $\lim_{T\to\infty}s\left(T\right)=\kappa$.

14.8 Appendix 3: Transition probability matrices and pricing

Consider the matrix $P(T-t)$ for $T-t \equiv \Delta t$, $P(\Delta t)$, and write,

$$P(\Delta t)_{ij} \equiv \begin{cases} 1+\lambda_{ij}\Delta t, & i=j \\ \lambda_{ij}\Delta t, & i \neq j \end{cases} \tag{14A.7}$$

We are defining the constants λ_{ij} as they were the counterparts of the intensity of the Poisson process in Eq. (14.24). Accordingly, these constants are simply interpreted as the instantaneous probabilities of migration from rating i to rating j over the time interval Δt. Naturally, for each i, we have that $\sum_{j=1}^{Z} P(\Delta t)_{ij} = 1$, and using into Eq. (14A.7), we obtain,

$$\lambda_{ii} = -\sum_{j=1, j\neq i}^{Z} \lambda_{ij}. \tag{14A.8}$$

The matrix Λ containing the elements λ_{ij} defined in Eqs. (14A.7) and (14A.8) is called the *generating matrix*.

Next, let us rewrite Eq. (14A.7) in matrix form,

$$P(\Delta t) = I + \Lambda \Delta t.$$

Suppose we have a time interval $[0, T]$, which we chop into n pieces, so to have $\Delta t = \frac{T}{n}$. We have,

$$P(T) = P(\Delta t)^n = \left(I + \Lambda \frac{T}{n}\right)^n.$$

For large n,

$$P(T) = \exp(\Lambda T), \tag{14A.9}$$

the matrix exponential, defined as, $\exp(\Lambda T) \equiv \sum_{n=0}^{\infty} \frac{(T\Lambda)^n}{n!}$.

To evaluate derivatives "written on states," we proceed as follows. Suppose F_i is the price of derivative in state $i \in \{1, \cdots, Z\}$. Suppose the Markov chain is the only source of uncertainty relevant for the evaluation of this derivative. Then,

$$dF_i = \frac{\partial F_i}{\partial t} dt + [F_{\tilde{R}} - F_i],$$

where $\tilde{R} \in \{1, \cdots, Z\}$, with the usual conditional probabilities. In words, the instantaneous change in the derivative value, dF_i, is the sum of two components: one, $\frac{\partial F_i}{\partial t} dt$, related to the mere passage of time, and the other, $[F_{\tilde{R}} - F_i]$, related to the discrete change arising from a change in the rating.

Suppose that $r = 0$. Then,

$$rF_i = 0 = \frac{E(dF_i)}{dt} = \frac{\partial F_i}{\partial t} + \sum_{j=1}^{Z} \lambda_{ij}(F_j - F_i) = \frac{\partial F_i}{\partial t} + \sum_{j\neq i} \lambda_{ij}(F_j - F_i),$$

with the appropriate boundary conditions.

As an example, consider defaultable bonds. In this case, we may be looking for pricing functions having the following form,

$$F_i(T-t) = xQ_i(T-t) + 1 - Q_i(T-t),$$

and then solve for $Q_i(T-t)$, for all $i \in \{1, \cdots, Z\}$. Naturally, we have

$$\begin{aligned} 0 &= xQ_i' - Q_i' + \sum_{j\neq i} \lambda_{ij} [x(Q_j - Q_i) - (Q_j - Q_i)] \\ &= x\left(Q_i' + \textstyle\sum_{j\neq i} \lambda_{ij}(Q_j - Q_i)\right) - \left(Q_i' + \textstyle\sum_{j\neq i} \lambda_{ij}(Q_j - Q_i)\right), \end{aligned}$$

which holds if and only if,

$$Q_i' = -\sum_{j\neq i}\lambda_{ij}\left(Q_j - Q_i\right) = -\sum_{j\neq i}\lambda_{ij}Q_j + \sum_{j\neq i}\lambda_{ij}Q_i = -\left(\sum\nolimits_{j\neq i}\lambda_{ij}Q_j + \lambda_{ii}Q_i\right).$$

That is, $Q' = -\Lambda Q$, which solved through the appropriate boundary conditions, yields precisely Eq. (14A.9).

14.9 Appendix 4: Bond spreads in markets with stochastic default intensity

We derive Eq. (14.40), by relying on the pricing formulae of Chapter 13. If the short-term is constant, the price of a defaultable bond derived in Section 13.3.7 of Chapter 13 can be easily extended to (and using the notation of the present chapter)

$$P(y,N)=e^{-rN}\mathbb{E}_0\left(e^{-\int_0^N\lambda_t dt}\right)+\int_0^N e^{-rt}\underbrace{\mathbb{E}_0\left(\lambda_t e^{-\int_0^t\lambda_u du}\right)}_{=\Pr\{\text{Default}\in(t,t+dt)\}}\text{Rec}(t)\,dt. \tag{14A.10}$$

The term indicated inside the integral of the second term, is indeed the density of default time at t, because,

$$P_{\text{default by time }t}(\lambda)=1-\mathbb{E}\left(e^{-\int_0^t\lambda_s ds}\right),$$

such that by differentiating with respect to t, yields, under the appropriate regularity conditions, that $\Pr\{\text{Default}\in(t,t+dt)\}$ is just the term indicated in Eq. (14A.10). So Eq. (14.40) follows. Naturally,

$$\Pr\{\text{Default}\in(t,t+dt)\}=-\frac{\partial}{\partial t}P_{\text{surv}}(\lambda,t).$$

Replacing this into Eq. (14A.10),

$$\begin{aligned}P(y,N)&=e^{-rN}\mathbb{E}_0\left(e^{-\int_0^N\lambda_t dt}\right)+\text{Rec}\int_0^N e^{-rt}\left(-\frac{\partial}{\partial t}P_{\text{surv}}(\lambda,t)\right)dt\\&=1-\text{LGD}\left(1-e^{-rN}P_{\text{surv}}(\lambda,N)\right)-(1-\text{LGD})\int_0^N re^{-rt}P_{\text{surv}}(\lambda,t)\,dt,\end{aligned}$$

where the second equality follows by integration by parts and the assumption of constant recovery rates. Setting $r=0$, produces Eq. (14.41).

Appendix 5: Bond and CDS spreads

We provide some analytical details regarding the behavior of bonds and CDS spreads depicted in Figure 14.14. First, we show that it holds, approximately, that bonds spreads are dominated by CDS spreads. We have:

$$\begin{aligned}\frac{-1}{N}\ln P(\lambda,N) &= \frac{-1}{N}\ln\left[1-\text{LGD}\cdot\left(1-P_{\text{surv}}(\lambda,N)\right)\right]\\ &\approx \text{LGD}\cdot\frac{1-P_{\text{surv}}(\lambda,N)}{N}\\ &\leq \text{LGD}\cdot\frac{1-P_{\text{surv}}(\lambda,N)}{\frac{1}{4}\sum_{i=1}^{4N}P_{\text{surv}}(\lambda,t_i)}\\ &= 4\times\text{CDS}_0(N).\end{aligned}$$

Second, we show that approximately, bonds and CDS spreads are bounded away by a function decreasing in time to maturity once the current λ is close to its long-term average under the risk-neutral probability, $\bar{\lambda}$. We illustrate this property while relying on arguments similar to those utilized in Chapter 13 to address a related topic (see Section 13.3.4). For bond spreads, since $\mathbb{E}_0(\lambda_s) = \bar{\lambda}+e^{-\phi s}\left(\lambda-\bar{\lambda}\right)$, we have, approximately, that:

$$\begin{aligned}\frac{-1}{N}\ln P(\lambda,N) &= \frac{-1}{N}\ln\left[1-\text{LGD}\cdot\left(1-P_{\text{surv}}(\lambda,N)\right)\right]\\ &= \frac{-1}{N}\ln\left[1-\text{LGD}\cdot\left(1-\mathbb{E}_0\left(e^{-\int_0^N\lambda_s ds}\right)\right)\right]\\ &\leq \frac{-1}{N}\ln\left[1-\text{LGD}\cdot\left(1-e^{-\int_0^N\mathbb{E}_0(\lambda_s)ds}\right)\right]\\ &\approx \text{LGD}\cdot\frac{1-e^{-\int_0^N\mathbb{E}_0(\lambda_s)ds}}{N}\\ &= \text{LGD}\cdot\frac{1-e^{-\bar{\lambda}N-\left(\lambda-\bar{\lambda}\right)\frac{1-e^{-\phi N}}{\phi}}}{N}.\end{aligned}$$

Therefore, even if $\lambda=\bar{\lambda}$, bond spreads are bounded away by a function decreasing in N. Naturally, this property does not mechanically imply that bond spreads are decreasing in N too, although the existence of such a bounding function helps this happening. As for the CDS spreads, we have, approximately that:

$$4\times\text{CDS}_0(N)=\text{LGD}\cdot\frac{1-P_{\text{surv}}(\lambda,N)}{\frac{1}{4}\sum_{i=1}^{4N}P_{\text{surv}}(\lambda,t_i)}\leq\text{LGD}\cdot\frac{1-P_{\text{surv}}(\lambda,N)}{N\cdot P_{\text{surv}}(\lambda,N)}\approx-\frac{1}{N}\ln P_{\text{surv}}(\lambda,N),$$

such that for $\lambda=\bar{\lambda}$, $\text{CDS}_0(N)$ is bounded away by a function decreasing in N, due to the same arguments made to for the bond spreads.

14.10 Appendix 6: Conditional probabilities of survival

We prove Eqs. (14.43)-(14.45). First, for (t_{i-1}, t_i) small, the numerator in Eq. (14.42) can be replaced by

$$-\frac{\partial}{\partial t}P_{\text{surv}}(\lambda, t) \equiv \mathbb{E}_0\left(\lambda_t e^{-\int_0^t \lambda_s ds}\right),$$

and rescaled by dt. Regularity conditions under which we can perform this differentiation can be found in a related context developed in Mele (2003). Eqs. (14.43)-(14.44) follow.

As for Eq. (14.45), the proof follows the same lines of reasoning as that in Appendix 3 of Chapter 13. That is, we can define a density process,

$$\eta_T(\tau) = \frac{e^{-\int_0^\tau \lambda_s ds} P_{\text{surv}}(\lambda_\tau, \tau, T)}{\mathbb{E}\left(e^{-\int_0^T \lambda_s ds}\right)}, \quad P_{\text{surv}}(\lambda_\tau, \tau, T) \equiv \mathbb{E}\left(e^{-\int_\tau^T \lambda_s ds}\middle| \mathcal{F}_\tau\right).$$

It is easy to show that the drift of P_{surv} is $\lambda_\tau d\tau$, such that by Itô's lemma,

$$\frac{d\eta_T(\tau)}{\eta_T(\tau)} = -\left(-\text{Vol}\left(P_{\text{surv}}(\lambda_\tau, \tau, T)\right)\right) dW_\tau,$$

where,

$$-\text{Vol}\left(P_{\text{surv}}(\lambda_\tau, \tau, T)\right) \equiv -\frac{\frac{\partial}{\partial \lambda}P_{\text{surv}}(\lambda_\tau, \tau, T)}{P_{\text{surv}}(\lambda_\tau, \tau, T)}\sigma\sqrt{\lambda_\tau} = B(T-\tau)\,\sigma\sqrt{\lambda_\tau},$$

where the second line follows by the closed-form expression of P_{surv} in Eq. (14.37). Therefore, $W_{\lambda,\tau}$ is a Brownian motion under Q_λ, where

$$dW_{\lambda,\tau} = dW_\tau + B(T-\tau)\,\sigma\sqrt{\lambda_\tau}d\tau,$$

and Eq. (14.45) follows.

14.11 Appendix 7: Details regarding CDS index swaps and swaptions

Proof of Eq. (14.56). We have

$$\mathrm{DSX}_{\tau}\left(t,t_{0}\right)=\mathbb{E}_{\tau}\left(\sum_{j=1}^{n}e^{-\int_{\tau}^{\tau_{j}}r_{u}du}\mathrm{LGD}\frac{1}{n}\mathbb{I}_{\{t\leq\tau_{j}\leq t_{4M}\}}\right)-\mathbb{E}_{\tau}\left(\sum_{i=1}^{4M}e^{-\int_{\tau}^{t_{i}}r_{u}du}C_{t}N_{t_{i}}\right). \tag{14A.11}$$

The expectation of the first term in Eq. (14A.11) is proportional to

$$\begin{aligned}&\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_{\tau}\left(e^{-\int_{\tau}^{\tau_{j}}r_{u}du}\mathbb{I}_{\{\mathrm{Surv}_{j}\text{ at }\tau\}}\mathbb{I}_{\{\mathrm{Def}_{j}\in(t_{0},t_{4M})\}}\right)\\&=\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}_{\{\mathrm{Surv}_{j}\text{ at }\tau\}}\mathbb{E}_{\tau}\left(e^{-\int_{\tau}^{\tau_{j}}r_{u}du}\mathbb{I}_{\{\mathrm{Def}_{j}\in(t_{0},t_{4M})\}}\right)=N_{\tau}\mathbb{E}_{\tau}\left(e^{-\int_{\tau}^{\tau_{*}}r_{\tau}d\tau}\mathbb{I}_{\{\tau_{*}\in(t_{0},t_{4M})\}}\right),\end{aligned} \tag{14A.12}$$

where the last equality follows by the definition of the outstanding notional value in Eq. (14.54), and the fact that the expectation in the first equality is the same for each name j, due to the assumption that the index names have the same credit quality.

The expectation of the second term in Eq. (14A.11) follows by noticing that

$$\begin{aligned}&\mathbb{E}_{\tau}\left(e^{-\int_{\tau}^{t_{i}}r_{\tau}d\tau}\mathbb{I}_{\{\mathrm{Surv}_{j}\text{ at }t_{i}\}}\right)\\&=\mathbb{E}_{\tau}\left(e^{-\int_{\tau}^{t_{i}}r_{\tau}d\tau}\mathbb{I}_{\{\mathrm{Surv}_{j}\text{ at }\tau\}}\mathbb{I}_{\{\mathrm{Surv}_{j}\text{ at }t_{i}|\mathrm{Surv}_{j}\text{ at }\tau\}}\right)=\mathbb{I}_{\{\mathrm{Surv}_{j}\text{ at }\tau\}}\mathbb{E}_{\tau}\left(e^{-\int_{\tau}^{t_{i}}r_{\tau}d\tau}\mathbb{I}_{\{\mathrm{Surv}_{j}\text{ at }t_{i}|\mathrm{Surv}_{j}\text{ at }\tau\}}\right),\end{aligned}$$

and summing over the reset dates and all names, and using the fact that all names have the same default intensities.

Proof of Eq. (14.57). The derivation of Eq. (14A.12) relies on default events occurring after the swap origination, i.e. over the reset dates, after $T=t_{0}$. In evaluating the *front-end protection*, we need to price securities that pay off over defaults possibly occurring over the life of the swaption, i.e. before time $T=t_{0}$. We have,

$$\begin{aligned}V_{\tau}^{\mathrm{F}}&=\mathbb{E}_{\tau}\left(e^{-\int_{\tau}^{T}r_{u}du}F_{T}\right)\\&=\mathrm{LGD}\frac{1}{n}\mathbb{E}_{\tau}\left(e^{-\int_{\tau}^{T}r_{u}du}\sum_{j=1}^{n}\left(\mathbb{I}_{\{\mathrm{Def}_{j}\in(t,\tau)\}}+\mathbb{I}_{\{\mathrm{Surv}_{j}\text{ at }\tau\}}\mathbb{I}_{\{\mathrm{Def}_{j}\in(\tau,T)\}}\right)\right)\\&=\mathrm{LGD}\frac{1}{n}\mathcal{D}\left(t,\tau\right)P\left(\tau,T\right)+\mathrm{LGD}\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}_{\{\mathrm{Surv}_{j}\text{ at }\tau\}}\mathbb{E}_{\tau}\left(e^{-\int_{\tau}^{T}r_{u}du}\left(1-\mathbb{I}_{\{\mathrm{Surv}_{j}\text{ at }T|\mathrm{Surv}_{j}\text{ at }\tau\}}\right)\right)\\&=\mathrm{LGD}\left(\frac{1}{n}\mathcal{D}\left(t,\tau\right)P\left(\tau,T\right)+N_{\tau}\left(P\left(\tau,T\right)-P_{\mathrm{def}}\left(\tau,T\right)\right)\right),\end{aligned}$$

where the third equality holds by the assumption that the names have the same credit quality, and $P\left(\tau,T\right)=\mathbb{E}_{\tau}(e^{-\int_{\tau}^{T}r_{u}du})$ and $P_{\mathrm{def}}\left(\tau,T\right)=\mathbb{E}_{\tau}(e^{-\int_{\tau}^{T}(r_{u}+\lambda_{u})du})$. Note that the first term in the brackets of the second equality is, obviously, always zero, when the timing of possible defaults does not overlap with the evaluation horizon, as for Eq. (14A.12).

Proofs regarding the survival contingent probability. We show that $\hat{Q}_{\text{sc}}$ in Eq. (14.59) does integrate to one, and that $\text{CDX}_t(M) = \text{LGD}\frac{V_{0\tau}}{V_{1\tau}} + \frac{V_\tau^{\text{F}}}{N_\tau V_{1\tau}}$ in Eq. (14.58), is a martingale under $\hat{Q}_{\text{sc}}$. We rely on results that generalize the following equality:

$$\begin{aligned}\mathbb{E}_\tau(N_T) &= \mathbb{E}_\tau\left(\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}_{\{\text{Surv}_j \text{ at } T\}}\right)\\ &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}_{\{\text{Surv}_j \text{ at } \tau\}}\mathbb{E}_\tau\left[\mathbb{E}\left(\mathbb{I}_{\{\text{Surv}_j \text{ at } T|\text{Surv}_j \text{ at } \tau\}}\middle|\mathbb{F}_T^r\right)\right]\\ &= N_\tau\mathbb{E}_\tau\left(e^{-\int_\tau^T \lambda_u du}\right).\end{aligned}$$

For example, regarding the survival contingent probability $\hat{Q}_{\text{sc}}$, we have that, under regularity conditions,

$$\begin{aligned}\mathbb{E}_\tau\left(e^{-\int_\tau^T r_u du}N_T V_{1T}\right) &= \mathbb{E}_\tau\left[\mathbb{E}\left(e^{-\int_\tau^T r_u du}N_T V_{1T}\middle|\mathbb{F}_T^r\right)\right]\\ &= N_\tau\mathbb{E}_\tau\left(e^{-\int_\tau^T (r_u+\lambda_u)du}V_{1T}\right)\\ &= N_\tau V_{1\tau},\end{aligned}$$

where the second equality follows by Eq. (14.51) and the third by the definition of Q_{sc} in Eq. (14.52).

As for the martingale property of $\text{CDX}_t(M)$ under $\hat{Q}_{\text{sc}}$, let $\hat{\mathbb{E}}_\tau^{\text{sc}}(\cdot)$ the time τ conditional expectation operator under the the survival contingent probability $\hat{Q}_{\text{sc}}$. We have, using the definition of $\hat{Q}_{\text{sc}}$ in Eq. (14.59),

$$\begin{aligned}&\hat{\mathbb{E}}_\tau^{\text{sc}}(\text{CDX}_T(M)) = \text{LGD}\cdot\hat{\mathbb{E}}_\tau^{\text{sc}}\left(\frac{V_{0T}}{V_{1T}}\right) + \hat{\mathbb{E}}_\tau^{\text{sc}}\left(\frac{V_T^{\text{F}}}{N_T V_{1T}}\right)\\ &= \text{LGD}\cdot\mathbb{E}_\tau\left(e^{-\int_\tau^T r_u du}\frac{N_T V_{1T}}{N_\tau V_{1\tau}}\frac{V_{0T}}{V_{1T}}\right) + \mathbb{E}_\tau^{\text{sc}}\left(e^{-\int_\tau^T r_u du}\frac{N_T V_{1T}}{N_\tau V_{1\tau}}\frac{V_T^{\text{F}}}{N_T V_{1T}}\right)\\ &= \text{LGD}\frac{1}{V_{1\tau}}\cdot\mathbb{E}_\tau\left(e^{-\int_\tau^T r_u du}\frac{N_T}{N_\tau}V_{0T}\right) + \frac{1}{N_\tau V_{1\tau}}\mathbb{E}_\tau\left(e^{-\int_\tau^T r_u du}V_T^{\text{F}}\right)\\ &= \text{LGD}\frac{V_{0\tau}}{V_{1\tau}} + \frac{V_\tau^{\text{F}}}{N_\tau V_{1\tau}},\end{aligned}$$

where the last equality follows by the Law of Iterated Expectations and Eq. (14.53),

$$\begin{aligned}\mathbb{E}_\tau\left(e^{-\int_\tau^T r_u du}\frac{N_T}{N_\tau}V_{0T}\right) &= \mathbb{E}_\tau\left[\mathbb{E}\left(e^{-\int_\tau^T r_u du}\frac{N_T}{N_\tau}V_{0T}\middle|\mathbb{F}_T^r\right)\right]\\ &= \mathbb{E}_\tau\left(e^{-\int_\tau^T (r_u+\lambda_u)du}V_{0T}\right)\\ &= V_{0\tau}.\end{aligned}$$

14.12 Appendix 8: Modeling correlation with copulae functions

A. Statistical independence and correlation

Two random variables are always uncorrelated, provided they are independently distributed. Yet there might be situations where two random variables are *not* correlated and still exhibit statistical dependence. As an example, suppose a random variable y relates to another, x, through $y = kx^3$, for some constant k, and x can take on $2N+1$ values, $x \in \{-x_N, x_{N-1}, \cdots, -x_1, 0, x_1, \cdots, x_{N-1}, x_N\}$, and $\Pr\{x_j\} = \frac{1}{2N+1}$. Then, we have that $\text{Cov}(x,y) \propto \sum_{j=1}^{N}(-x_j)x_j^3 + \sum_{j=1}^{N}(x_j)x_j^3 = 0$ and yet, y and x are obviously dependent. This example might be interpreted, economically, as one where y and x are two returns on two asset classes. These two returns are not correlated, overall. Yet the comove in the same direction in both very bad and in very good times. This appendix is a succinct introduction to copulae, which are an important tool to cope with these issues.

Consider two random variables Y_1 and Y_2. We may relate Y_1 to another random variable Z_1 and we may relate Y_2 to a second random variable Z_2, on a percentile-to-percentile basis, viz

$$F_i(y_i) = G_i(z_i), \quad i = 1, 2, \tag{14A.13}$$

where F_i are the cumulative marginal distributions of Y_i, and G_i are the cumulative marginal distributions of Z_i. That is, for each y_i, we look for the value of z_i such that the percentiles arising through the mapping in Eq. (14A.13) are the same. Then, we may assume that Z_1 and Z_2 have a joint distribution and model the correlation between Y_1 and Y_2 through the correlation between Z_1 and Z_2. This indirect way to model the correlation between Y_1 and Y_2 is particularly helpful. It might be used to model the correlation of default times, as in the main text of this chapter. We now explain.

B. Copulae functions

We begin with the simple case of two random variables, This simple case shall be generalized to the multivariate one with a mere change in notation. Given two uniform random variables U_1 and U_2, consider the function $C(u_1, u_2) = \Pr(U_1 \leq u_1, U_2 \leq u_2)$, which is the joint cumulative distribution of the two uniforms. A copula function is any such function C, with the property of being capable to aggregate the marginals F_i into a summary of them, in the following natural way:

$$C(F_1(y_1), F_2(y_2)) = F(y_1, y_2), \tag{14A.14}$$

where $F(y_1, y_2)$ is the joint distribution of (y_1, y_2). Thus, a copula function is simply a cumulative bivariate distribution function, as $F(Y_1)$ and $F(Y_2)$ are obviously uniformly distributed. To prove Eq. (14A.14), note that

$$\begin{aligned} C(F_1(y_1), F_2(y_2)) &= \Pr(U_1 \leq F_1(y_1), U_2 \leq F_2(y_2)) \\ &= \Pr\left(F_1^{-1}(U_1) \leq y_1, F_2^{-1}(U_2) \leq y_2\right) \\ &= \Pr(Y_1 \leq y_1, Y_2 \leq y_2) \\ &= F(y_1, y_2). \end{aligned} \tag{14A.15}$$

That is, a copula function evaluated at the marginals $F_1(y_1)$ and $F_2(y_2)$ returns the joint density $F(y_1, y_2)$. In fact, Sklar (1959) proves that, conversely, any multivariate distribution function F can be represented through some copula function.

The most known copula function is the Gaussian copula, which has the following form:

$$C(u_1, u_2) = \Phi\left(\Phi_1^{-1}(u_1), \Phi_2^{-1}(u_2)\right), \tag{14A.16}$$

where Φ denotes the joint cumulative Normal distribution, and Φ_i denotes marginal cumulative Normal distributions. So we have,

$$F(y_1, y_2) = C(F_1(y_1), F_2(y_2)) = \Phi\left(\Phi_1^{-1}(F_2(y_2)), \Phi_2^{-1}(F_2(y_2))\right), \tag{14A.17}$$

where the first equality follows by Eq. (14A.15) and the second equality follows by Eq. (14A.16).

As an example, we may interpret Y_1 and Y_2 as the times by which two names default. A simple assumption is to set:

$$F_i(y_i) = \Phi_i(z_i), \quad i = 1, 2, \tag{14A.18}$$

for two random variables Z_i that are "stretched" as explained in Part A of this appendix. By replacing Eq. (14A.18) into Eq. (14A.17),

$$F(y_1, y_2) = \Phi(z_1, z_2).$$

This reasoning can be easily generalized to the N-dimensional case, where:

$$F(y_1, \cdots, y_N) = C(F_1(y_1), \cdots, F_N(y_N)) = \Phi(z_1, \cdots, z_N),$$

where

$$z_i : F_i(y_i) = \Phi_i(z_i).$$

We use this approach to model default correlation among names, as explained in the main text, and in the next appendix.

14.13 Appendix 9: Details on CDO pricing with imperfect correlation

We follow the copula approach to price the stylized CDOs in the main text of this chapter. For each name, create the following random variable,

$$z_i = \sqrt{\rho}F + \sqrt{1-\rho}\epsilon_i, \quad i = 1, 2, 3, \tag{14A.19}$$

where F is a common factor among the three names, ϵ_i is an idiosynchratic term, and $F \sim N(0,1)$, $\epsilon_i \sim N(0,1)$. Finally, $\rho \geq 0$ is meant to capture the default correlation among the names, as follows. Assume that the *risk-neutral* probability each firm defaults, by T, is given by,

$$\mathbb{Q}_i(T) = \Phi(\zeta_{0.10}) \equiv 10\%,$$

where Φ is the cumulative distribution of a standard normal variable. That is, by time T, each firm defaults any time that,

$$z_i < \zeta_{0.10} \equiv \Phi^{-1}(10\%).$$

Therefore, ρ is the default correlation among the assets in the CDO.

We can now simulate Eq. (14A.19), build up payoffs for each simulation, and price the tranches by just averaging over the simulations, as explained below. Naturally, the same simulation technique can be used to price tranches on CDOs with an arbitrary number of assets. Precisely, simulate Eq. (14A.19), and obtain values $\tilde{z}_{i,s}$, $s = 1, \cdots, S$, where S is the number of simulations and $i = 1, 2, 3$. At simulation no s, we have

$$\tilde{z}_{1,s}, \tilde{z}_{2,s}, \tilde{z}_{3,s}, \quad s \in \{1, \cdots, S\}.$$

We use the previously simulated values as follows:

- For each simulation s, count the number of defaults across the three names, defined as the number of times that $\tilde{z}_{i,s} < \zeta_{0.10}$, for $i = 1, 2, 3$. Denote the number of defaults as of simulation s with Def_s.
- For each simulation s, compute the total realized payoff of the asset pool, defined as,

$$\tilde{\pi}_s = \text{Def}_s \cdot 40 + (3 - \text{Def}_s) \cdot 100.$$

- For each simulation s, compute recursively the payoffs to each tranche, $\pi_{i,s}$,

$$\pi_{i,s} = \min\left\{\max\left\{\tilde{\pi}_s - \sum_{k=1}^{i-1}\pi_{k,s}, 0\right\}, N_i\right\},$$

  where N_i is the nominal value of each tranche ($N_1 = 140$, $N_2 = 90$, $N_3 = 70$).
- Estimate the price of each tranche by averaging across the simulations,

$$\text{Price senior} = e^{-r}\frac{1}{S}\sum_{s=1}^{S}\pi_{1,s}, \quad \text{Price mezzanine} = e^{-r}\frac{1}{S}\sum_{s=1}^{S}\pi_{2,s}, \quad \text{Price Junior} = e^{-r}\frac{1}{S}\sum_{s=1}^{S}\pi_{3,s}.$$

Note, the previous computations have to be performed under the risk-neutral probability Q. Using the probability P in the previous algorithm can only be lead to something useful for risk-management and VaR calculations at best.

Note, this model, can be generalized to a multifactor model where,

$$z_i = \sqrt{\rho_{i1}}F_1 + \cdots + \sqrt{\rho_{id}}F_d + \sqrt{1-\rho_{i1}-\cdots-\rho_{id}}\epsilon_i,$$

with obvious notation.

References

Allen, F. and D. Gale (2007): *Understanding Financial Crises*. Clarendon Lectures in Finance. Oxford: Oxford University Press.

Amato, J. D. (2005): "Risk Aversion and Risk Premia in the CDS Market." BIS Quarterly Review, September, 55-68.

Anderson, R. W. and S. Sundaresan (1996): "Design and Valuation of Debt Contracts." *Review of Financial Studies* 9, 37-68.

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999): "Coherent Measures of Risk." *Mathematical Finance* 9, 203-228.

Berndt, A., R. Douglas, D. Duffie, M. Ferguson and D. Schranz (2005): "Measuring Default Risk-Premia from Default Swap Rates and EDFs." BIS Working Papers no. 173.

Black, F. (1976): "The Pricing of Commodity Contracts." *Journal of Financial Economics* 3, 167-179.

Black, F. and J. Cox (1976): "Valuing Corporate Securities: Some Effect of Bond Indenture Provisions." *Journal of Finance* 31, 351-367.

Black, F. and M. Scholes (1973): "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81, 637-659.

Broadie, M., M. Chernov and S. Sundaresan (2007): "Optimal Debt and Equity Values in the Presence of Chapter 7 and Chapter 11." *Journal of Finance* 62, 1341-1377.

Christoffersen, P. F. (2003): *Elements of Financial Risk Management*. Academic Press.

Cox, J. C., J. E. Ingersoll and S. A. Ross (1985): "A Theory of the Term Structure of Interest Rates." *Econometrica* 53, 385-407.

Danielsson, J., Shin, H. S. and J.-P. Zigrand (2011): "Balance Sheet Capacity and Endogenous Risk." Working paper London School of Economics and Princeton.

Duffie, D. and D. Lando (2001): "Term Structure of Credit Spreads with Incomplete Accounting Information." *Econometrica* 69, 633-664.

Fender, I. and P. Hördahl (2007): "Overview: Credit Retrenchement Triggers Liquidity Squeeze." BIS Quarterly Review (September), 1-16.

Greenspan, A. (1998): "The Role of Capital in Optimal Banking Supervision and Regulation." FRBNY Economic Policy Review, October, 163-168.

Hull, J. C. (2007): *Risk Management and Financial Institutions*. Pearson Education International.

Ingersoll, J. E. (1977): "A Contingent-Claims Valuation of Convertible Securities." *Journal of Financial Economics* 5, 289-321.

International Monetary Fund, (2008): Global Financial Stability Report. April 2008.

Jamshidian, F. (1989): "An Exact Bond Option Pricing Formula." *Journal of Finance* 44, 205-209.

Jarrow, R. A., D. Lando and S. M. Turnbull (1997): "A Markov Model for the Term-Structure of Credit Risk Spreads." *Review of Financial Studies* 10, 481-523.

Jorion, Ph. (2008): *Value at Risk*. New York: McGraw Hill.

Keys, B., T. Mukherjee, A. Seru, and V. Vig (2010): "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans." *Quarterly Journal of Economics* 125, 307-362.

Lando, David, 2004. *Credit Risk Modeling—Theory and Applications*. Princeton: Princeton University Press.

Leland, H. E. (1994): "Corporate Debt Value, Bond Covenants and Optimal Capital Structure." *Journal of Finance* 49, 1213-1252.

Leland, H. E. and K. B. Toft (1994): "Optimal Capital Structure, Endogenous Bankruptcy, and the Term Structure of Credit Spreads." *Journal of Finance* 51, 987-1019.

Lopez, J. (2004): "The Empirical Relationship Between Average Asset Correlation, Firm Probability of Default and Asset Size." *Journal of Financial Intermediation* 13, 265-283.

McDonald, R. L. (2006): *Derivatives Markets*, Boston: Pearson International Edition.

Mele, A. (2003): "Fundamental Properties of Bond Prices in Models of the Short-Term Rate." *Review of Financial Studies* 16, 679-716.

Merton, R. C. (1974): "On the Pricing of Corporate Debt: The Risk-Structure of Interest Rates." *Journal of Finance* 29, 449-470.

Modigliani, F. and M. Miller (1958): "The Cost of Capital, Corporation Finance and the Theory of Investment." *American Economic Review* 48, 261-297.

Morini, M. and D. Brigo (2011): "No-Armageddon Arbitrage-Free Equivalent Measure for Index Options in a Credit Crisis." *Mathematical Finance* 21, 573-593.

Reinhart, C.M. and K.S. Rogoff (2009): *This Time Is Different: Eight Centuries of Financial Folly*. Princeton: Princeton University Press.

Rutkowski, M. and A. Armstrong (2009): "Valuation of Credit Default Swaptions and Credit Default Index Swaptions." *International Journal of Theoretical and Applied Finance* 12, 1027-1053.

Schönbucher, Ph J., 2003. *Credit Risk Pricing Models—Models, Pricing and Implementation*. Chichester, UK: Wiley Finance.

Shin, H. S. (2010): *Risk and Liquidity*. Clarendon Lectures in Finance, Oxford University Press.

Sklar, A. (1959): "Fonction de Répartition à N dimensions et Leurs Marges." *Publications de l'Institut Statistique de l'Université de Paris* 8, 229-231.

Vasicek, O. (1987): "Probability of Loss on Loan Portfolio." Working paper KMV, published in: *Risk* (December 2002) under the title "Loan Portfolio Value."