Notes on Fixed Income Derivative Pricing

Sarthak Bagaria

First published 25 August 2026
Last revised 4 October 2026

Contents

Introduction **10**

I Stochastic Calculus **11**

1 Brownian Motion **12**

1.1 Stochastic Processes 12

1.2 Itô Calculus 18

2 Price Process Characterization **33**

2.1 Why This Form and No Other 33

2.2 Dropping Continuity 39

2.3 Two Kinds of Jump 46

3 The Generator **50**

3.1 Itô's Lemma, Read as a Definition 50

3.2 The Semigroup 53

3.3 The Adjoint, and Why There Are Two Equations 55

3.4 What to Carry Forward 60

4 No-Arbitrage Pricing **61**

4.1 Binomial Model 61

4.2 Continuous Time Itô Diffusion Model 63

4.3 From No Arbitrage to a Martingale Measure 66

4.4 Many Periods 70

4.5 What Changes in Continuous Time 71

4.6 Splitting the Risk in Two 76

4.7 Why Every Model Here Is a Diffusion 79

5 Black Scholes **82**
5.1 Black Scholes Option Pricing . . . 82
5.2 What the Model Was Not Needed For . . . 85
5.3 The Option's Own Diffusion . . . 86
5.4 Theta Is Gamma . . . 87
5.5 Numbers Carried in the Head . . . 90
5.6 Why the Formula Is a Price . . . 91

6 Numeraires **94**
6.1 Change of Measure . . . 94
6.2 Numeraire Measures . . . 102

II Classical Models **109**

7 Curve Construction and Linear Products **110**
7.1 Discount Factors, Zero Rates and Forward Rates . . . 110
7.2 The Year Fraction Is Not a Length of Time . . . 112
7.3 The Instruments . . . 113
7.4 Bootstrapping the Curve . . . 116
7.5 The Curve as a Posterior . . . 121
7.6 The Front End Is a Step Function . . . 128
7.7 Two Curves: Projection and Discounting . . . 132
7.8 The Rest of the Multicurve Stack . . . 135
7.9 Curve Risk . . . 137

8 Term Structure Models **139**
8.1 What We Are Modelling . . . 139
8.2 Heath-Jarrow-Morton Framework . . . 140
8.3 Hull-White Model . . . 143
8.4 Future-Forward Basis . . . 146
8.5 Swaptions . . . 150

9 Implied and Local Volatility **155**
9.1 Implied Volatility . . . 156
9.2 The Prices Are the Distribution . . . 157
9.3 What the Standard Combinations Price . . . 160

9.4 How the Density Moves . . . 161
9.5 Local Volatility and Dupire's Formula . . . 161
9.6 Dupire in the Coordinates We Actually Have . . . 164
9.7 Gyongi's Theorem . . . 166
9.8 Scheduled Events on the Surface . . . 169

10 Smile Dynamics and Stochastic Volatility 173
10.1 What Is Left to Get Wrong . . . 173
10.2 Two Rules of Thumb, and What They Would Mean . . . 174
10.3 The Backbone of a Local Volatility Model . . . 175
10.4 The Backbone the Market Actually Has . . . 179
10.5 The Forward Smile . . . 181
10.5.1 A Local Volatility Model With No Calendar In It . . . 181
10.5.2 What Calibration Puts In . . . 182
10.5.3 Whether the Trade Is Forced . . . 183
10.6 Making Volatility Random . . . 184
10.7 The Smile Does Not Determine the Dynamics . . . 185
10.8 Heston, and What a Term Structure Costs . . . 187
10.9 The Short-Dated Skew, and Rough Volatility . . . 188

11 Local-Stochastic Volatility 196
11.1 Taking the Good Half of Each . . . 196
11.2 The Calibration Condition . . . 197
11.3 The Circularity, and How It Is Broken . . . 200
11.4 The Mixing Weight . . . 202
11.5 What Has Been Bought . . . 204

12 Markovian Term-Structure Models 206
12.1 Why HJM Is Not Enough . . . 206
12.2 What a Finite State Requires . . . 207
12.2.1 Markov Is Exactly Finite Rank . . . 207
12.2.2 Finite Rank Alone Forces Nothing . . . 209
12.2.3 The Second Requirement . . . 209
12.2.4 Rank One Is the Exponential . . . 211
12.2.5 Rank n Is a Differential Equation . . . 212
12.2.6 What Rank Two and Three Look Like . . . 213

12.2.7 Several Levels . . . 214
12.3 Excursion: Affine Models and Why They Solve . . . 220
12.4 Where the Smile Lives . . . 223
12.5 More Than One Factor . . . 227
12.6 Calibration, and the Parameter Nobody Can See . . . 230

13 Market Models **233**
13.1 Modelling What Is Quoted . . . 233
13.2 The Drift, and Why There Has To Be One . . . 234
13.3 The Numeraire Is a Choice, and Jamshidian's Is the Better One . . . 236
13.4 No Small State . . . 237
13.5 The Same Curve in Swap Rate Coordinates . . . 237
13.6 Which Coordinates, and What the Answer Costs . . . 238
13.7 Calibration . . . 241
13.8 Simulating It . . . 242
13.9 What Is Missing, and It Is the Smile . . . 243

14 Solvable Models **248**
14.1 Two Collapses, Often Confused . . . 248
14.2 Invariant Families . . . 251
14.2.1 A Subspace: Polynomials . . . 251
14.2.2 A Manifold: Exponential-Affine Functions . . . 254
14.2.3 What Each One Actually Gives You . . . 257
14.2.4 The Distinction, and What It Decides . . . 259
14.3 Reading the Recipe Backwards . . . 260
14.4 Why the Riccati Equation Is Solvable . . . 261
14.5 The Spectral Route . . . 263
14.6 What This Chapter Is For . . . 266

III Products and Dependence **269**

15 CMS and Static Replication **270**
15.1 A Rate Paid on the Wrong Date . . . 270
15.2 The Annuity Mapping Function . . . 271
15.3 The Payoff as a Portfolio . . . 275
15.4 Cash Settled Swaptions . . . 276
15.5 What This Chapter Did Not Need . . . 278

16 Credit **279**
16.1 What We Already Have . . . 279
16.2 The Credit Default Swap . . . 279
16.3 The Credit Triangle . . . 280
16.4 Bootstrapping the Hazard Curve . . . 282
16.5 Marking a Position . . . 284
16.6 The Two Risk Numbers . . . 285
16.7 Making the Intensity Random . . . 285
16.8 Options on Credit . . . 286
16.9 What One Name Cannot Tell You . . . 287

17 Foreign Exchange, Quantos and Inflation **289**
17.1 An Exchange Rate Is a Ratio of Numeraires . . . 289
17.2 The Quanto Adjustment . . . 291
17.3 Inflation Is Foreign Exchange . . . 293
17.4 Where the Model Comes Back: Year-on-Year Swaps . . . 294
17.5 What Carries Over . . . 297

18 Dependence **298**
18.1 The Gap, Stated Precisely . . . 298
18.2 Correlation Is the Wrong Word . . . 299
18.3 The Number That Actually Matters . . . 301
18.4 Four Things To Use Instead . . . 304
18.4.1 Give the copula tails . . . 306
18.4.2 Give it asymmetric tails . . . 306
18.4.3 Model the cause, not the coupling . . . 306
18.4.4 Do not model dependence at all . . . 309
18.5 The Gaussian Copula, and What It Cost . . . 309
18.6 The Same Gap in Rates . . . 312

IV Numerical and Statistical Technique **315**

19 Numerical Methods **316**
19.1 The Constraint That Chose the Models . . . 316
19.2 Two Ways to Compute . . . 317
19.3 Arbitrage-Free Becomes a Property of the Estimator . . . 320

19.4 Early Exercise, Where the Difficulty Concentrates . . . 321
19.5 What Lifting the Constraint Permits . . . 325
19.6 Why This Matters for Trading and Not Only for Pricing . . . 329
19.7 What Computation Does Not Fix . . . 329

20 Fitting and Testing 332
20.1 Choosing the Model in the First Place . . . 332
20.2 Two Jobs With One Name . . . 337
20.3 A Perfect Fit That Determines Nothing . . . 337
20.4 The Bias Nobody Checks . . . 339
20.4.1 Two Other Ways to Fit the Same Process . . . 342
20.5 Testing Rather Than Estimating . . . 343
20.6 Does the Model Fit At All? . . . 346
20.7 What To Take Away . . . 349

21 Estimation in Practice 350
21.1 What the Newer Methods Change . . . 350
21.1.1 Latent states: filtering, and the likelihood it hands over . . . 350
21.1.2 Particle filters, in more detail . . . 352
21.1.3 Reporting the valley: Bayesian calibration . . . 354
21.1.4 Speed, not knowledge: neural approximation . . . 356
21.1.5 Learned models, and the two questions they do not answer . . . 356
21.2 What Is Actually Latent . . . 357
21.3 Regimes, and Whether They Can Be Caught . . . 359
21.4 Calibrating to Prices and to History at Once . . . 362
21.5 Estimating a Volatility . . . 364
21.6 Estimating a Correlation . . . 368
21.7 Why Implied Exceeds Realised . . . 370
21.8 What To Take Away . . . 373

V Trading and Risk 375

22 Relative Value and the Basis Trade 376
22.1 What This Chapter Is . . . 376
22.2 Who Is On the Other Side . . . 377
22.3 How Long Being Right Takes . . . 379

22.4 Whether the Spread Converges at All . . . 380
22.5 Carry, Roll-Down, and the Rest . . . 381
22.6 Curve Trades . . . 383
22.7 The Cash-Futures Basis . . . 386
22.8 Why It Breaks . . . 387
22.9 Swap Spreads, and a Negative Number That Should Not Exist . . . 388
22.10 Selling Volatility, and What That Is Short . . . 389
22.11 Six More, and What Each One Is Short . . . 390
22.12 What Retail Actually Trades . . . 393
22.13 Where an Edge Lives: Information as a Filtration . . . 395

23 Market Making **402**
23.1 What the Spread Is Payment For . . . 402
23.2 A Solvable Quoting Problem . . . 403
23.3 What the Model Cannot See . . . 407
23.4 Measuring Toxicity . . . 408
23.5 The Race the Short Markout Measures . . . 409
23.6 Quoting on a Rates Desk . . . 411
23.7 A Signal, and Who It Is Worth More To . . . 413
23.8 The Partial Hedge . . . 414
23.9 What Is Handed On . . . 415

24 Risk Management **417**
24.1 A Different Question . . . 417
24.2 What a Risk Measure Should Satisfy . . . 418
24.3 Value at Risk, and What Is Wrong With It . . . 419
24.4 Expected Shortfall . . . 420
24.5 Computing It . . . 421
24.6 The Coordinates a Risk Report Uses . . . 424
24.7 Explaining the Profit and Loss . . . 424
24.8 Attributing the Number . . . 426
24.9 Backtesting . . . 428
24.10 Model Risk, and Turning It Into a Number . . . 430
24.11 Counterparty Risk . . . 431
24.12 One Model or Many . . . 432

25 Solvability, Realism and What Is Left **435**

25.1 What the Book Did 435

25.2 The Constraint That Replaced Compute 436

25.3 What Solvability Still Buys 437

25.4 One Result, From Three Directions 437

25.5 What To Do About It 438

Acknowledgements **440**

Further Reading **441**

Introduction

These are notes on the theory of fixed income derivative pricing, running from the construction of the Itô integral to how a swaption book is quoted, hedged and risk-managed. They aim to be self-contained at graduate level: where a step needs an argument, the argument is given, and where a piece of mathematics is needed — functional analysis, Lie algebras, spectral theory — it is introduced rather than gestured at. They began as my own notes, written while learning interest rate modelling and collected over years from books, papers and colleagues; I make no claim to originality.

What they try to add is organisation. The industry's ideas are scattered across a literature written for other purposes. A reader — a junior quant, or anyone who has to work with interest rate derivatives — after finishing these notes should be able to do more than recite what the models say: choose one for a problem, extend it when it does not fit, know which of its assumptions are load-bearing, and see where the mathematics and the history came from. What they are not is an implementation manual; the accompanying code exists to check the notes' claims rather than to serve as a template for a production system. The modelling choices are the subject.

That emphasis is deliberate. Derivative pricing has accumulated a zoo of models that are not competitors in a contest one of them wins — each was designed under constraints, often how little compute was available at the time, and a model's form is a record of them. So the skill worth building is navigating that zoo: why a model has the shape it has, what it can and cannot be made to fit, what breaks when it is extended, how to test whether it does what one believes, and what its compromises cost on a desk. The same questions recur throughout — where solvability is genuinely needed and where modern numerics has removed it, which old compromises are still necessary, and what a model's failures look like in a profit and loss report rather than a norm. Chapter 25 collects the answers.

Every quantitative claim in the prose — every rate, ratio and comparison — is computed by the Rust crate published alongside the notes and pinned by a test, with closed forms cross-checked against an independent route that shares none of their algebra. Figures come from the same crate. The notes are revised often — chapters are added, split and reordered — so the revision date on the title page is the one to trust. They are also a work in progress rather than a finished text, and the latest version is always at `https://www.sarthakbagaria.com/notes/derivativepricing/`.

These notes are a personal project, not affiliated with, endorsed by, or the responsibility of any employer, past or present. They have also been extended and revised with AI assistance — several chapters, the figures and the accompanying code were written that way — but every figure is drawn by published code with tests beside it, so the numbers can be checked rather than taken on trust. The errors that remain are mine, and I would be glad to hear about any that are found.

Part I

Stochastic Calculus

Chapter 1

Brownian Motion

We build the objects the rest of the book computes with. Filtrations and martingales first, then the Itô integral — which is defined by a limit that is not obviously there, so we show why it exists rather than assuming it, and find that the Itô isometry is what supplies the limit rather than merely describing it. Along the way the Stratonovich integral appears as the alternative that obeys the ordinary chain rule and is not adapted, which is the trade between geometry and predictability. The chapter ends with Itô's lemma and the diffusions it applies to.

1.1 Stochastic Processes

Definition 1.1 (Brownian Motion)**.** A brownian motion is a stochastic process $W_{t\geq 0}$ such that the following properties hold:

- $W_0 = 0$ with probability 1.
- $W_a - W_b \sim N(0, |a-b|)$ where $N(m, v)$ is a Normal distribution with mean m and variance v.
- $W_a - W_b$ and $W_c - W_d$ are independent random variables for $a > b \geq c > d$.

So the process starts at zero, moves by independent normal steps, and a step's variance is exactly its length — which is why the path accumulates randomness at a constant, known rate.

Definition 1.2 (σ-Algebra)**.** A σ-algebra on a set X is a collection Σ of subsets of X such that the following properties hold:

- $X \in \Sigma$.
- $A \in \Sigma \Rightarrow A^c \in \Sigma$ (closed under complement)
- $A_i \in \Sigma \Rightarrow \bigcup_i A_i \in \Sigma$ (closed under countable union)

A probability is a map $\Sigma \to \mathbb{R}$ assigning a number to each event in Σ — so the σ-algebra is exactly the collection of events whose probability is known, and the closure axioms say that knowing an event's probability means knowing its complement's and any countable combination's as well.

Definition 1.3 (Filtration)**.** A filtration is an indexed σ-algebra $\mathcal{F}_t$ of subsets of the probability sample space such that $F_a \supseteq F_b$ for $a \geq b$.

So $\mathcal{F}_t$ is the information available at time t — a finer σ-algebra distinguishes more events — and the nesting condition says that information only grows: nothing is forgotten as t increases.

Definition 1.4 (Measurable Function)**.** Given a σ-algebra $\mathcal{F}$ on the sample space Ω, a real-valued random variable X is $\mathcal{F}$-measurable if the X-inverse of any open set in $\mathbb{R}$ is contained in $\mathcal{F}$ i.e.

$$\{A \in \Omega | X(A) \in (a, b)\} \in \mathcal{F} \qquad \forall a, b \in \mathbb{R} \quad a > b$$

So observing the value of a measurable function identifies an event already in $\mathcal{F}$ — the set of outcomes consistent with that observation — which is what lets $\mathcal{F}$ "know" the function's value.

Definition 1.5 (Adapted Process)**.** A stochastic process X_t is adapted to filtration $\mathcal{F}_t$ if for each t, X_t is F_t measurable.

So a process adapted to $\mathcal{F}_t$ has a value at time t already implied by the information available then — nothing about X_t is still hidden from $\mathcal{F}_t$.

Definition 1.6 (Generated Filtration)**.** The filtration generated by a stochastic process is the smallest filtration to which the process is adapted.

Example 1.1. Let's take an example of 3 unbiased coin tosses with process X_t counting the number of heads until toss t. The probability sample space Ω is

$$\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

The filtration generated by the process is

$$
\begin{aligned}
\mathcal{F}_0 =\{ & \\
& \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}, \quad (X_0 = 0) \\
& \emptyset \\
\} & \\
\mathcal{F}_1 =\{ & \\
& \{HHH, HHT, HTH, HTT\}, \quad (X_0 = 0, X_1 = 1) \\
& \{THH, THT, TTH, TTT\}, \quad (X_0 = 0, X_1 = 0) \\
& \emptyset \\
& \text{and all unions of the above} \\
\} & \\
\mathcal{F}_2 =\{ & \\
& \{HHH, HHT\}, \quad (X_0 = 0, X_1 = 1, X_2 = 2) \\
& \{HTH, HTT\}, \quad (X_0 = 0, X_1 = 1, X_2 = 1) \\
& \{THH, THT\}, \quad (X_0 = 0, X_1 = 0, X_2 = 1) \\
& \{TTH, TTT\}, \quad (X_0 = 0, X_1 = 0, X_2 = 0) \\
& \emptyset \\
& \text{and all unions of the above} \\
\} & \\
\mathcal{F}_3 =\{ & \\
& \{HHH\}, \quad (X_0 = 0, X_1 = 1, X_2 = 2, X_3 = 3) \\
& \{HHT\}, \quad (X_0 = 0, X_1 = 1, X_2 = 2, X_3 = 2) \\
& \{HTH\}, \quad (X_0 = 0, X_1 = 1, X_2 = 1, X_3 = 2) \\
& \{HTT\}, \quad (X_0 = 0, X_1 = 1, X_2 = 1, X_3 = 1) \\
& \{THH\}, \quad (X_0 = 0, X_1 = 0, X_2 = 1, X_3 = 2) \\
& \{THT\}, \quad (X_0 = 0, X_1 = 0, X_2 = 1, X_3 = 1) \\
& \{TTH\}. \quad (X_0 = 0, X_1 = 0, X_2 = 0, X_3 = 1) \\
& \{TTT\}, \quad (X_0 = 0, X_1 = 0, X_2 = 0, X_3 = 0) \\
& \emptyset \\
& \text{and all unions of the above} \\
\} &
\end{aligned}
$$

Definition 1.7 (Expectation conditional on filtration)**.** $\mathbb{E}^{\mathbb{P}}[X|\mathcal{F}]$ is the $\mathcal{F}$ measurable random variable such that

$$\mathbb{E}^{\mathbb{P}}\big[X\,\mathbf{1}_A\big] = \mathbb{E}^{\mathbb{P}}\big[\mathbb{E}^{\mathbb{P}}[X|\mathcal{F}]\,\mathbf{1}_A\big] \quad \forall A \in \mathcal{F} \tag{1.1}$$

So averaging X over any event $\mathcal{F}$ can see gives the same answer whether or not X is first replaced

by its conditional expectation — $\mathbb{E}[X|\mathcal{F}]$ is the $\mathcal{F}$-measurable random variable that matches X's average on every event $\mathcal{F}$ can distinguish, even though it need not match X itself.

In the coin toss example above,

$$\mathbb{E}[X_3|\mathcal{F}_1](s) = \begin{cases} 2 & \text{if } s \in \{HHH, HHT, HTH, HTT\} \\ 1 & \text{if } s \in \{THH, THT, TTH, TTT\} \end{cases}$$

Definition 1.8 (Martingale)**.** A stochastic process X_t is a martingale with respect to a filtration $\mathcal{F}_t$ of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ if

$$\mathbb{E}^{\mathbb{P}}[X_a|\mathcal{F}_b] = X_b \quad \text{for } a \geq b$$

So a martingale's expected future value, conditional on the present, is the present value — no drift that today's information can see coming.

Exercise. Show that in the coin toss example, if we had instead counted number of heads minus number of tails, X_t would have been a martingale. For example, show that $\mathbb{E}[X_3|\mathcal{F}_1]$ would have been equal to X_1.

Relaxing the equality in either direction gives the two processes that drift one way on average.

Definition 1.9 (Supermartingale and submartingale)**.** With the same setup, X_t is a *supermartingale* if

$$\mathbb{E}^{\mathbb{P}}[X_a|\mathcal{F}_b] \leq X_b \quad \text{for } a \geq b,$$

and a *submartingale* if the inequality runs the other way. A martingale is both.

The names are the wrong way round for anyone reading them as a description of the paths: a supermartingale is the one that decreases in expectation. They come from potential theory, where a function u is superharmonic when its value at a point is at least its average over any surrounding sphere. Composing such a u with a Brownian motion averages it over exactly those spheres, so $u(W_t)$ is a supermartingale precisely when u is superharmonic, and the two words describe the same inequality. In the language of chapter 3 the condition reads $\mathcal{L}u \leq 0$: a non-positive drift is what a supermartingale has.

Definition 1.10 (Stopping time)**.** A random time $\tau : \Omega \to [0, \infty]$ is a *stopping time* for the filtration $\mathcal{F}_t$ if $\{\tau \leq t\} \in \mathcal{F}_t$ for every t. Whether it has already happened must be decidable from what is known at t, without looking ahead.

The first time a Brownian motion reaches a level, $\tau_a = \inf\{t : W_t \geq a\}$, is a stopping time — at any t you know whether it has happened. The *last* time before T that the path visits a is not, because deciding it at time t requires knowing that the path never returns. Every trading rule that can actually be followed is measurable in the first sense, which is why the condition appears wherever a strategy or an exercise decision does.

Given a stopping time, $X_t^\tau = X_{t \wedge \tau}$ is the process *stopped* at τ: it follows X until τ and is constant thereafter. Stopping preserves the martingale property, and it is the standard way to tame a process that is badly behaved only in the limit — the definition of a local martingale below is exactly that idea.

Definition 1.11 (Càdlàg path)**.** A path is *càdlàg* (*continue à droite, limite à gauche*) if at every time it is right-continuous and has a left limit. It may jump: at a jump the path takes its value on the right, and the left limit records what was approached from before. In any bounded interval a càdlàg path has at most countably many jumps, and for each $\epsilon > 0$ only finitely many exceeding ϵ. A continuous path is càdlàg; the Poisson and jump processes of chapter 2 are càdlàg without being continuous, and are why the theorem below is stated this way. It is standard, and no real restriction, to take every martingale to have càdlàg paths.

Theorem 1.12 (Optional stopping)**.** *Let X be a càdlàg martingale and τ a stopping time. If τ is bounded, or if $\{X_{t\wedge\tau}\}$ is uniformly integrable, then $\mathbb{E}[X_\tau] = X_0$.*

Proof. Take first a τ with finitely many values $t_1 < \cdots < t_n = T$. The whole argument is that the event $\{\tau = t_i\}$ is known by time t_i, so the martingale property may be used on it:

$$\begin{aligned}
\mathbb{E}[X_T] &= \sum_{i=1}^{n} \mathbb{E}\big[X_T \mathbf{1}_{\{\tau=t_i\}}\big] \\
&= \sum_{i=1}^{n} \mathbb{E}\Big[\mathbb{E}[X_T \mid \mathcal{F}_{t_i}]\, \mathbf{1}_{\{\tau=t_i\}}\Big] && \text{since } \{\tau = t_i\} \in \mathcal{F}_{t_i} \\
&= \sum_{i=1}^{n} \mathbb{E}\big[X_{t_i} \mathbf{1}_{\{\tau=t_i\}}\big] && \text{the martingale property} \\
&= \mathbb{E}[X_\tau],
\end{aligned}$$

and $\mathbb{E}[X_T] = X_0$ because X is a martingale. Notice what did the work. The indicator could be taken inside the conditional expectation only because it was $\mathcal{F}_{t_i}$ measurable, and that is the definition of a stopping time — the only property of τ the calculation used. A random time that looked ahead would break the second line and nothing else.

A general bounded $\tau \le T$ is a limit of such times. Put $\tau_n = \big(\lceil 2^n\tau\rceil/2^n\big) \wedge T$, which takes finitely many values, is a stopping time because $\{\tau_n \le k/2^n\} = \{\tau \le k/2^n\}$, and decreases to τ. Each $\mathbb{E}[X_{\tau_n}] = X_0$ by the case just proved, and $X_{\tau_n} \to X_\tau$ almost surely by right-continuity of the path, the approximating times decreasing to τ. The same case also identifies $X_{\tau_n} = \mathbb{E}[X_T \mid \mathcal{F}_{\tau_n}]$, and a family of conditional expectations of one integrable variable is uniformly integrable, so the convergence holds in L^1 and the expectations pass to the limit.

In the unbounded case the hypothesis does exactly the job boundedness did. The stopped process is a martingale, so $\mathbb{E}[X_{t\wedge\tau}] = X_0$ for every finite t, and $X_{t\wedge\tau} \to X_\tau$ almost surely as $t \to \infty$ on $\{\tau < \infty\}$; uniform integrability is what upgrades that to L^1 and lets the expectation follow the limit. Without it the two cannot be exchanged, and the next paragraph is a case where they cannot. □

A finite stopping time is not enough on its own, and the gap is not a technicality. The doubling strategy of chapter 4 stops a Brownian motion at the first time it reaches \$1, which happens almost surely, and collects $\mathbb{E}[X_\tau] = 1$ from a process starting at zero. What fails is uniform integrability: the paths that have not yet reached the target are ever further below it, and the loss they carry is exactly what the expectation is missing. Ruling that out is the work done by admissibility.

Stopping also supplies the weakest useful version of the martingale property.

Definition 1.13 (Local Martingale)**.** A process X is a local martingale if there are stopping times $\tau_n \uparrow \infty$ such that each stopped process $X_{t\wedge\tau_n}$ is a martingale.

"Locally a martingale" says the process is a martingale up to a sequence of stopping times that eventually exhausts the horizon, which for many processes of interest is all one can guarantee. The gap between local and true is the same gap optional stopping left open, and it is not empty: the doubling strategy has a local martingale whose expectation moves. When the process is bounded below, though, half of the property survives.

The proof rests on two convergence facts for conditional expectation, the conditional forms of monotone convergence and Fatou, which recur whenever a local martingale is pushed to its limit.

Ordinary monotone convergence is the starting point: if $0 \le Y_n \uparrow Y$ then $\mathbb{E}[Y_n] \uparrow \mathbb{E}[Y]$, both sides allowed to be infinite. An increasing sequence of non-negative functions has nowhere to put mass in the limit, so the integral tracks the pointwise limit exactly. Conditioning preserves this.

Lemma 1.14 (Conditional monotone convergence)**.** *If $0 \le Y_n \uparrow Y$ almost surely then $\mathbb{E}[Y_n|\mathcal{G}] \uparrow \mathbb{E}[Y|\mathcal{G}]$ almost surely, for every sub-σ-algebra $\mathcal{G}$.*

Proof. Conditional expectation is monotone, so $\mathbb{E}[Y_n|\mathcal{G}]$ increases almost surely to some $\mathcal{G}$-measurable L. For each $A \in \mathcal{G}$, ordinary monotone convergence on both sides of $\mathbb{E}[Y_n\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[Y_n|\mathcal{G}]\mathbf{1}_A]$ sends the left to $\mathbb{E}[Y\mathbf{1}_A]$ and the right to $\mathbb{E}[L\mathbf{1}_A]$. So $\mathbb{E}[L\mathbf{1}_A] = \mathbb{E}[Y\mathbf{1}_A]$ for every $A \in \mathcal{G}$, which with L being $\mathcal{G}$-measurable is exactly $L = \mathbb{E}[Y|\mathcal{G}]$. □

Lemma 1.15 (Conditional Fatou)**.** *If $Y_n \ge 0$ then for every sub-σ-algebra $\mathcal{G}$,*

$$\mathbb{E}\big[\liminf_n Y_n \,\big|\, \mathcal{G}\big] \le \liminf_n \mathbb{E}[Y_n|\mathcal{G}].$$

Proof. Put $Z_n = \inf_{k\ge n} Y_k$. These are non-negative and increase to $\liminf_n Y_n$, so conditional monotone convergence gives $\mathbb{E}[Z_n|\mathcal{G}] \uparrow \mathbb{E}[\liminf_n Y_n|\mathcal{G}]$. And $Z_n \le Y_k$ for every $k \ge n$, so $\mathbb{E}[Z_n|\mathcal{G}] \le \inf_{k\ge n} \mathbb{E}[Y_k|\mathcal{G}]$, whose right side rises to $\liminf_n \mathbb{E}[Y_n|\mathcal{G}]$ as $n \to \infty$. □

The inequality goes one way because the only grip on the limit is the approximation from below by the Z_n; nothing approximates it from above unless the Y_n are dominated. Equality needs the mass held in place, which is what a dominating function or uniform integrability provides.

Lemma 1.16 (A non-negative local martingale is a supermartingale)**.** *If $X \ge 0$ is a local martingale then $\mathbb{E}[X_a|\mathcal{F}_b] \le X_b$ for $a \ge b$.*

Proof. Let $\tau_n \uparrow \infty$ reduce X, so each stopped process is a genuine martingale and $\mathbb{E}[X_{a\wedge\tau_n}|\mathcal{F}_b] = X_{b\wedge\tau_n}$. Because $\tau_n \uparrow \infty$, almost every path has an N past which $\tau_n > a$, so $X_{a\wedge\tau_n} \to X_a$ and $X_{b\wedge\tau_n} \to X_b$ almost surely. The $X_{a\wedge\tau_n}$ are non-negative, so conditional Fatou gives

$$\mathbb{E}[X_a|\mathcal{F}_b] \le \liminf_n \mathbb{E}[X_{a\wedge\tau_n}|\mathcal{F}_b] = \liminf_n X_{b\wedge\tau_n} = X_b.$$

□

Structure (Where the expectation leaks)**.** Each $X_{a\wedge\tau_n}$ has the same expectation as X_0, being a stopped martingale, and the sequence converges almost surely to X_a — yet $\mathbb{E}[X_a]$ can come out strictly smaller. Splitting on whether the localisation has bitten by time a,

$$\mathbb{E}[X_0] = \mathbb{E}[X_{a\wedge\tau_n}] = \mathbb{E}\big[X_a \mathbf{1}_{\{\tau_n > a\}}\big] + \mathbb{E}\big[X_{\tau_n} \mathbf{1}_{\{\tau_n \le a\}}\big],$$

and the first term increases to $\mathbb{E}[X_a]$ as $\{\tau_n > a\}$ fills out, so

$$\mathbb{E}[X_0] - \mathbb{E}[X_a] = \lim_n \mathbb{E}\big[X_{\tau_n} \mathbf{1}_{\{\tau_n \le a\}}\big] \ge 0.$$

The gap is the mass stranded on the paths already stopped: $\{\tau_n \le a\}$ shrinks to null, but the value X_{τ_n} carried there need not shrink with it. Conditioning on $\mathcal{F}_b$ runs the same split on every event known by time b, which is why X_b comes out as an upper bound and not an equality.

So a non-negative local martingale can lose expectation but never gain it, and it is a true martingale exactly when it loses none. chapter 4 is about which side of that line a discounted price has to be on.

Definition 1.17 ($\mathcal{F}_t$ Brownian motion)**.** A Brownian motion W_t is a $\mathcal{F}_t$ Brownian motion if W_t is adapted to $\mathcal{F}_t$ and $W_{t+s} - W_t$ is independent of $\mathcal{F}_t$ for $s > 0$.

Remark (An $\mathcal{F}_t$ Brownian motion is a martingale with respect to filtration $\mathcal{F}_t$)**.**

$$\begin{aligned} \mathbb{E}[W_a|\mathcal{F}_b] &= \mathbb{E}[W_a - W_b|\mathcal{F}_b] + \mathbb{E}[W_b|\mathcal{F}_b] \\ &= 0 + \mathbb{E}[W_b|\mathcal{F}_b] \\ &= W_b \end{aligned}$$

1.2 Itô Calculus

Definition 1.18 (Itô Integral)**.** Let H_t be a $\mathcal{F}_t$ adapted process and W_t be a $\mathcal{F}_t$ Brownian motion. The Itô integral is defined as

$$\int_0^t H_s dW_s = \lim_{n\to\infty} \sum_{i=0}^{2^n-1} H_{t_i^n} \Delta_i W$$

This is similar to the Riemann integral. One thing to note here is that the integrand always takes the value at the start of the interval. If we make the integrand take the average of the values at interval endpoints, we get

Definition 1.19 (Stratonovich Integral)**.**

$$\int_0^t H_s \circ dW_s = \lim_{n\to\infty} \sum_{i=0}^{2^n-1} (H_{t_i^n} + H_{t_{i+1}^n})\Delta_i W/2 \tag{1.2}$$

The two integrals usually do not give the same results, as can be checked by repeating the calculations we do in these notes for Stratonovich integral as well.

This is the first place stochastic calculus parts company with ordinary calculus, and the reason is visible in a single algebraic step. Take $H_s = W_s$, so both sums are attempts at $\int_0^t W dW$, and look at one interval. The Stratonovich term exceeds the Itô term by

$$\tfrac{1}{2}(W_{i+1} + W_i)(W_{i+1} - W_i) - W_i(W_{i+1} - W_i) = \tfrac{1}{2}(W_{i+1} - W_i)^2,$$

so summing over the partition, the gap between the two integrals is exactly half the sum of squared increments. For an ordinary differentiable function that sum vanishes as the partition refines, and the two definitions agree — which is why nobody distinguishes them in ordinary calculus. For Brownian motion the sum of squared increments converges to t, so the gap does not vanish — it tends to $t/2$, and the Stratonovich integral of $W\,dW$ exceeds the Itô one by exactly that.

Structure (Which integral is the geometric one)**.** The two integrals differ by more than a convention.

Stratonovich obeys the ordinary chain rule: $d\,\varphi(X) = \varphi'(X) \circ dX$, with no second order term. So a Stratonovich equation transforms correctly under a change of coordinates — write the same process in a new variable and the equation you get is the one you would have written directly. Itô does not: its correction term $\frac{1}{2}\varphi''\sigma^2 dt$ depends on the coordinate chosen, so the same process has different-looking Itô equations in different variables.

That makes Stratonovich the integral of choice whenever geometry is involved — diffusions on a sphere, on a manifold of correlation matrices, on a Lie group — because there is no privileged coordinate for the equation to be written in. Finance uses Itô regardless: a trading strategy can only act on information already in hand, which forces evaluation at the left endpoint of each interval — Stratonovich's midpoint rule would let the strategy trade on the increment it is about to receive — and that same left-endpoint rule makes $\int H\,dM$ a martingale whenever M is one. Coordinate-invariance is what is given up in exchange, and finance, whose one natural coordinate is money, rarely needs it.

Both are used later. The point here is only that the choice is between coordinate-invariance and the non-anticipating rule — and the martingale property that comes with it — and one cannot have both.

Two things in the figure stand out. The Stratonovich sum is not converging to anything — it is already exactly right at thirty-two intervals, and at every coarser and finer resolution, because $\frac{1}{2}(W_{i+1} + W_i)(W_{i+1} - W_i) = \frac{1}{2}(W_{i+1}^2 - W_i^2)$ and the sum collapses. And the quadratic variation converges to t rather than to zero, which is the single fact that makes all of this necessary.

In finance the integrand is usually the quantities of assets in the portfolio and the stochastic process the price of those assets. The portfolio is rebalanced on the information available when each interval begins, so the integral is the Itô one.

Let's compute expectations and variances of some of the basic Itô Integrals. We assume that the integrand is square-integrable i.e. $\int_0^t \mathbb{E}[H_s^2]ds < \infty$. In this case we have

The calculations below are all sums over the dyadic partition of $[0, t]$. Two abbreviations make the structure easier to see. Write

$$t_i^n = \frac{i\,t}{2^n}, \qquad \Delta_i W = W_{t_{i+1}^n} - W_{t_i^n}$$

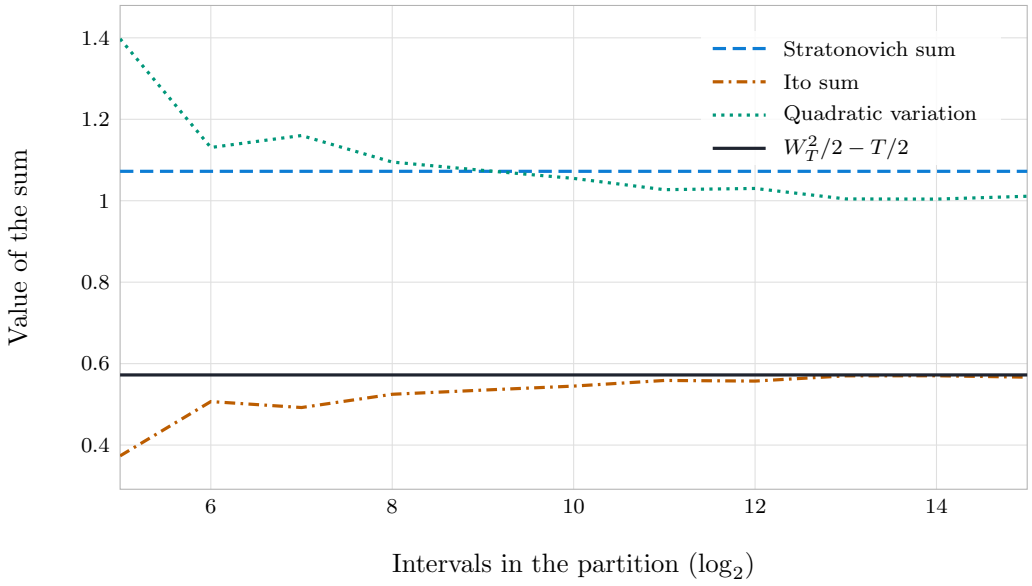

Figure 1.1: The two sums for $\int_0^t W\,dW$, computed on a single Brownian path that is refined rather than redrawn, so that what changes across the picture is the definition and not the randomness. The Stratonovich sum is exactly $W_t^2/2$ at every resolution, because it telescopes — it obeys the ordinary chain rule. The Itô sum sits half the quadratic variation below it and converges to $W_t^2/2 - t/2$. The extra $-t/2$ is not an approximation error; it is what Itô's lemma is about.
Drawn by `sums_over_path` and `brownian_path`.

for the i-th partition point and the Brownian increment across the i-th interval, so that the step length is $t^n_{i+1} - t^n_i = t/2^n$. A subscript on an expectation means conditioning on the information available then, so $\mathbb{E}_{t^n_i}[\,\cdot\,] = \mathbb{E}[\,\cdot \mid \mathcal{F}_{t^n_i}]$.

Calculation 1.20 (Expectation of Itô Integral)**.**

$$\begin{aligned}
\mathbb{E}[\int_0^t H_s dW_s|\mathcal{F}_0] &= \mathbb{E}[\lim_{n\to\infty} \sum_{i=0}^{2^n-1} H_{t^n_i}\Delta_i W|\mathcal{F}_l] \qquad \text{(by definition)} \\
&= \lim_{n\to\infty} \sum_{i=0}^{2^n-1} \mathbb{E}[H_{t^n_i}\Delta_i W|\mathcal{F}_0] \qquad \text{(by square integrability of } H_s) \\
&= \lim_{n\to\infty} \sum_{i=0}^{2^n-1} \mathbb{E}[\mathbb{E}[H_{t^n_i}\Delta_i W|\mathcal{F}_{t^n_i}]|\mathcal{F}_0] \qquad \text{(iterated conditioning)} \\
&= \lim_{n\to\infty} \sum_{i=0}^{2^n-1} \mathbb{E}[H_{t^n_i}\mathbb{E}_{t^n_i}[\Delta_i W]|\mathcal{F}_0] \qquad (H_s \text{ is adapted to } \mathcal{F}_s) \\
&= \lim_{n\to\infty} \sum_{i=0}^{2^n-1} \mathbb{E}[H_{t^n_i}0|\mathcal{F}_0] \qquad (W_s \text{ is } \mathcal{F}_s \text{ Brownian}) \\
&= 0
\end{aligned}$$

By iterated conditioning we also have $\mathbb{E}[\int_0^t H_s dW_s] = \mathbb{E}[\mathbb{E}[\int_0^t H_s dW_s|\mathcal{F}_0]] = \mathbb{E}[0] = 0$.

Calculation 1.21 (Covariance of Itô Integral)**.**

$$
\begin{aligned}
&\mathbb{E}[\int_0^t H_s dW_s \int_0^t K_s dW_s | \mathcal{F}_0] \\
&= \mathbb{E}[\lim_{n,m\to\infty} \sum_{i,j=0}^{2^n-1,2^m-1} H_{t_i^n} \Delta_i W K_{jt/2^m} (W_{(j+1)t/2^m} - W_{jt/2^m}) | \mathcal{F}_0] \\
&\qquad\qquad \text{(by definition)} \\
&= \lim_{n,m\to\infty} \sum_{i,j=0}^{2^n-1,2^m-1} \mathbb{E}[H_{t_i^n} \Delta_i W K_{jt/2^m} (W_{(j+1)t/2^m} - W_{jt/2^m}) | \mathcal{F}_0] \\
&\qquad\qquad \text{(by square integrability of } H_s\text{)} \\
&= \lim_{n,m\to\infty} \sum_{i,j=0}^{2^n-1,2^m-1} \mathbb{E}[\mathbb{E}[H_{t_i^n} \Delta_i W K_{jt/2^m} (W_{(j+1)t/2^m} - W_{jt/2^m}) | \mathcal{F}_{\max(i,j)t/2^n}] | \mathcal{F}_0] \\
&\qquad\qquad \text{(iterated conditioning)} \\
&= \lim_{n\to\infty} \sum_{i,j=0}^{2^n-1} \mathbb{E}[\mathbb{E}[H_{t_i^n} \Delta_i W K_{t_j^n} \Delta_j W | \mathcal{F}_{\max(i,j)t/2^n}] | \mathcal{F}_0] \\
&\qquad\qquad \text{(assuming same partition to simplify algebra)} \\
&= \lim_{n\to\infty} \sum_{i,j=0}^{2^n-1} \begin{cases} \mathbb{E}[H_{t_i^n} K_{t_j^n} \Delta_j W \mathbb{E}_{t_i^n}[\Delta_i W] | \mathcal{F}_0] & \text{if } i > j \\ \mathbb{E}[H_{t_i^n} K_{t_i^n} \mathbb{E}_{t_i^n}[(\Delta_i W)^2] | \mathcal{F}_0] & \text{if } i = j \\ \mathbb{E}[H_{t_i^n} \Delta_i W K_{t_j^n} \mathbb{E}_{t_j^n}[\Delta_j W] | \mathcal{F}_0] & \text{if } i < j \end{cases} \\
&= \lim_{n\to\infty} \sum_{i,j=0}^{2^n-1} \begin{cases} \mathbb{E}[H_{t_i^n} K_{t_j^n} \Delta_j W 0 | \mathcal{F}_0] & \text{if } i > j \\ \mathbb{E}[H_{t_i^n} K_{t_i^n} \mathrm{Var}_{t_i^n}[\Delta_i W] | \mathcal{F}_0] & \text{if } i = j \\ \mathbb{E}[H_{t_i^n} \Delta_i W K_{t_j^n} 0 | \mathcal{F}_0] & \text{if } i < j \end{cases} \\
&= \lim_{n\to\infty} \sum_{i,j=0}^{2^n-1} \begin{cases} 0 & \text{if } i \neq j \\ \mathbb{E}[H_{t_i^n} K_{t_i^n} t/2^n | \mathcal{F}_0] & \text{if } i = j \end{cases} \\
&= \lim_{n\to\infty} \sum_{i=0}^{2^n-1} \mathbb{E}[H_{t_i^n} K_{t_i^n} \, t/2^n | \mathcal{F}_0] \\
&= \mathbb{E}[\lim_{n\to\infty} \sum_{i=0}^{2^n-1} H_{t_i^n} K_{t_i^n} \, t/2^n | \mathcal{F}_0] \\
&= \mathbb{E}[\int_0^t H_s K_s ds | \mathcal{F}_0] \qquad \text{(Riemann sum, with } t/2^n \text{ the width)}
\end{aligned}
$$

Before writing down Itô's Isometry, let's set up the two spaces it relates, since the statement is about a map between them.

Definition 1.22 (The two spaces)**.** $L^2(\Omega)$ is the set of random variables X with finite second moment, $\mathbb{E}[X^2] < \infty$. An element of it is a single random number — the value of something at one

moment — and "finite second moment" is the condition that makes its variance meaningful. It is given the inner product

$$(X, Y)_{L^2(\Omega)} \equiv \mathbb{E}[XY].$$

$L^2_{ad}([0,T]\times\Omega)$ is the set of processes $X_s(\omega)$, defined for times $s \in [0,T]$ and states $\omega \in \Omega$, which are

- *adapted* — that is the subscript "ad" — so X_s is $\mathcal{F}_s$ measurable, meaning its value at time s uses only information available by then; and
- square integrable over both time and chance, $\mathbb{E}\big[\int_0^T X_s^2\, ds\big] < \infty$.

An element of it is a random *function of time* — a whole path — and it carries the inner product

$$(X, Y)_{L^2_{ad}([0,T]\times\Omega)} \equiv \mathbb{E}\bigg[\int_0^T X_s Y_s\, ds\bigg].$$

One contains random numbers, the other random paths, and both are equipped with a notion of length and angle built from an expectation. Adaptedness is doing real work in the second and is not a technical afterthought — it is what rules out an integrand that peeks at the increment it is about to be multiplied by, which is the whole reason the Itô integral has zero mean.

Remark (Itô Isometry)**.** Taking expectations of the above to remove the conditioning, the result is known as the Itô isometry. Write $\mathcal{I}(X) = \int_0^T X_s\, dW_s$ for the Itô integral over $[0,T]$. It takes a random path and returns a random number, so it is a map

$$\mathcal{I} : L^2_{ad}([0,T]\times\Omega) \longrightarrow L^2(\Omega),$$

and it preserves the inner product:

$$\big(\mathcal{I}(H), \mathcal{I}(K)\big)_{L^2(\Omega)} = (H, K)_{L^2_{ad}([0,T]\times\Omega)},$$

which written out is exactly the calculation above with the conditioning dropped:

$$\mathbb{E}\bigg[\int_0^T H_s\, dW_s \int_0^T K_s\, dW_s\bigg] = \mathbb{E}\bigg[\int_0^T H_s K_s\, ds\bigg].$$

In words: the integral carries a random path to a random number without distorting lengths or angles. Two integrands that are orthogonal — meaning $\mathbb{E}[\int H_s K_s\, ds] = 0$ — have uncorrelated integrals, and the size of an integral is the size of its integrand. That is what "isometry" means: a map preserving the geometry, in the way that a rotation of the plane preserves distances.

The isometry is not only a fact about the integral; it is what allows the integral to be defined at all. The argument is worth going through, both because the definition we started from is on shakier ground than it looks and because the same argument is used throughout analysis.

The definition was a limit of sums $\sum_i H_{t_i}(W_{t_{i+1}} - W_{t_i})$. For each partition that sum is a perfectly good random variable. But nothing so far says the sequence of them converges as the

partition refines, and a sequence of random variables does not converge merely because we would like it to. Something has to force it. The construction goes in three steps.

First, the integrands for which there is nothing to prove. Call a process *simple* if it is piecewise constant,

$$H_s = \sum_i h_i \, \mathbb{I}_{(t_i,\, t_{i+1}]}(s),$$

with each h_i already known at time t_i. Its integral is not defined by a limit at all; it is declared outright to be the finite sum

$$\mathcal{I}(H) = \sum_i h_i \big(W_{t_{i+1}} - W_{t_i}\big),$$

and a finite sum of products of random variables raises no questions. The covariance calculation above, applied to two simple processes, is exactly the statement that the isometry holds for these.

Second, approximate. Every $H \in L^2_{ad}([0,T]\times\Omega)$ is a limit of simple processes: there are simple $H^{(n)}$ with $\|H^{(n)} - H\| \to 0$ in that space. We would like to *define* $\mathcal{I}(H)$ as $\lim_n \mathcal{I}(H^{(n)})$ — but that limit has to be shown to exist before it can be used as a definition.

Third, the isometry supplies it. The integral is linear on simple processes, so the difference of two of the approximating integrals is the integral of the difference, and the isometry applies to it:

$$\big\|\mathcal{I}(H^{(n)}) - \mathcal{I}(H^{(m)})\big\|_{L^2(\Omega)} = \big\|\mathcal{I}(H^{(n)} - H^{(m)})\big\|_{L^2(\Omega)} = \big\|H^{(n)} - H^{(m)}\big\|_{L^2_{ad}([0,T]\times\Omega)}.$$

The right hand side tends to zero, because a sequence that converges is Cauchy. So the left hand side does too, which says the integrals $\mathcal{I}(H^{(n)})$ form a Cauchy sequence in $L^2(\Omega)$. And $L^2(\Omega)$ is *complete* — every Cauchy sequence in it has a limit inside it — so the limit exists. The same estimate shows it does not depend on which approximating sequence was chosen, so the definition is unambiguous.

Remark (What the isometry is doing)**.** It is a bridge between two things we can measure. On one side we can make the integrands close, because that is how the approximation was built. On the other we need the integrals to be close, because that is what makes the limit exist. The isometry says those are the same statement.

Without an estimate of that kind the approximating integrals could oscillate forever without settling, and there would be no integral to speak of — the definition would not fail to be useful, it would fail to be a definition.

With $H = K$, the above two results imply

$$\begin{aligned} \mathbb{E}[\int_0^t H_s dW_s] &= 0 \\ \mathrm{Var}[\int_0^t H_s dW_s] &= \mathbb{E}[\int_0^t H_s^2 ds]. \end{aligned}$$

One key feature of Itô calculus is that the second order differentials are not equivalent to 0. In

ordinary calculus we have

$$\begin{aligned}\int_0^t X(s)\,(ds)^2 &\equiv \lim_{n\to\infty}\sum_{i=0}^{2^n-1} X(it/2^n)\,((i+1)t/2^n - it/2^n)^2\\ &= \lim_{n\to\infty}\sum_{i=0}^{2^n-1} X(it/2^n)\,(t/2^n)^2\\ &= \lim_{n\to\infty}\frac{t}{2^n}\sum_{i=0}^{2^n-1} X(it/2^n)\,\frac{t}{2^n}\\ &= \Big(\lim_{n\to\infty}\frac{t}{2^n}\Big)\int_0^t X(s)\,ds = 0,\end{aligned}$$

the sum in the third line converging to the finite integral $\int_0^t X(s)\,ds$ while the factor $t/2^n$ pulled out in front vanishes.

In stochastic calculus we can check the expectation and variance of the distribution of the integral:

Calculation 1.23 (Integral of second order time differential)**.**

$$\begin{aligned}\mathbb{E}\Big[\int_0^t X_s\,(ds)^2\Big] &\equiv \mathbb{E}\Big[\lim_{n\to\infty}\sum_{i=0}^{2^n-1} X_{it/2^n}\,((i+1)t/2^n - it/2^n)^2\Big]\\ &= \lim_{n\to\infty}\sum_{i=0}^{2^n-1} \mathbb{E}[X_{it/2^n}]\,(t/2^n)^2\\ &= \lim_{n\to\infty}\frac{t}{2^n}\sum_{i=0}^{2^n-1} \mathbb{E}[X_{it/2^n}]\,\frac{t}{2^n}\\ &= \Big(\lim_{n\to\infty}\frac{t}{2^n}\Big)\int_0^t \mathbb{E}[X_s]\,ds = 0,\end{aligned}$$

Variance can be similarly computed to be 0 if the stochastic process has finite mean and variance at all times. Therefore the distribution is 0 with probability 1.

However the integral does not necessarily vanish when taken with respect to second order differential of Brownian motion. We have

Calculation 1.24 (Expectation of integral of second order Brownian motion differential)**.**

$$\begin{aligned}
\mathbb{E}[\int_0^t H_s(dW_s)^2 \mid \mathcal{F}_0] &= \mathbb{E}[\lim_{n\to\infty} \sum_{i=0}^{2^n-1} H_{t_i^n} \Delta_i W^2 \mid \mathcal{F}_0] \\
&= \lim_{n\to\infty} \sum_{i=0}^{2^n-1} \mathbb{E}[H_{t_i^n} \Delta_i W^2 \mid \mathcal{F}_0] \\
&= \lim_{n\to\infty} \sum_{i=0}^{2^n-1} \mathbb{E}[\mathbb{E}[H_{t_i^n} \Delta_i W^2 \mid \mathcal{F}_{t_i^n}] \mid \mathcal{F}_0] \\
&= \lim_{n\to\infty} \sum_{i=0}^{2^n-1} \mathbb{E}[H_{t_i^n} \mathbb{E}[\Delta_i W^2 \mid \mathcal{F}_{t_i^n}] \mid \mathcal{F}_0] \\
&= \lim_{n\to\infty} \sum_{i=0}^{2^n-1} \mathbb{E}[H_{t_i^n} t/2^n \mid \mathcal{F}_0] \\
&= \mathbb{E}[\int_0^t H_s \, ds \mid \mathcal{F}_0]
\end{aligned}$$

We therefore have $\mathbb{E}[\int_0^t H_s(dW_s)^2 - \int_0^t H_s \, ds] = 0$. So the two integrals have same expectation. Let's see if their difference has any variance.

Calculation 1.25 (Variance of integral of second order Brownian motion differential)**.**

$$
\begin{aligned}
&\mathbb{E}[\big(\int_0^t H_s(dW_s)^2 - \int_0^t H_s\,ds\big)^2|\mathcal{F}_0] \\
&= \mathbb{E}[\lim_{n\to\infty}\big(\sum_{i=0}^{2^n-1} H_{t_i^n}\big(\Delta_i W^2 - ((i+1)t/2^n - it/2^n)\big)\big)^2|\mathcal{F}_0] \\
&= \mathbb{E}[\lim_{n\to\infty}\big(\sum_{i=0}^{2^n-1} H_{t_i^n}\big(\Delta_i W^2 - t/2^n\big)\big)^2|\mathcal{F}_0] \\
&= \mathbb{E}[\lim_{n\to\infty}\sum_{i,j=0}^{2^n-1} H_{t_i^n}H_{t_j^n}(\Delta_i W^2 - t/2^n)((\Delta_j W)^2 - t/2^n)|\mathcal{F}_0] \\
&= \lim_{n\to\infty}\sum_{i,j=0}^{2^n-1} \mathbb{E}[H_{t_i^n}H_{t_j^n}(\Delta_i W^2 - t/2^n)((\Delta_j W)^2 - t/2^n)|\mathcal{F}_0] \\
&= \lim_{n\to\infty}\sum_{i,j=0}^{2^n-1} \mathbb{E}[\mathbb{E}[H_{t_i^n}H_{t_j^n}(\Delta_i W^2 - t/2^n)((\Delta_j W)^2 - t/2^n)|\mathcal{F}_{\max(i,j)t/2^n}]|\mathcal{F}_0] \\
&= \lim_{n\to\infty}\sum_{i,j=0}^{2^n-1} \begin{cases} \mathbb{E}[H_{t_i^n}H_{t_j^n}\big((\Delta_j W)^2 - t/2^n\big)\mathbb{E}[\Delta_i W^2 - t/2^n|\mathcal{F}_{t_i^n}]|\mathcal{F}_0] & \text{if } i>j \\ \mathbb{E}[(H_{t_i^n})^2\mathbb{E}[(\Delta_i W^2 - t/2^n)^2|\mathcal{F}_{t_i^n}]|\mathcal{F}_0] & \text{if } i=j \\ \mathbb{E}[H_{t_i^n}H_{t_j^n}\big((\Delta_i W)^2 - t/2^n\big)\mathbb{E}[\Delta_j W^2 - t/2^n|\mathcal{F}_{t_j^n}]|\mathcal{F}_0] & \text{if } i<j \end{cases} \\
&= \lim_{n\to\infty}\sum_{i,j=0}^{2^n-1} \begin{cases} \mathbb{E}[H_{t_i^n}H_{t_j^n}\big((\Delta_j W)^2 - t/2^n\big)0|\mathcal{F}_0] & \text{if } i>j \\ \mathbb{E}[(H_{t_i^n})^2\mathbb{E}[\Delta_i W^4 + (t/2^n)^2 - 2(W_{(i+1)t/2^n} - W_{it/2^n})^2 t/2^n|\mathcal{F}_{t_i^n}]|\mathcal{F}_0] & \text{if } i=j \\ \mathbb{E}[H_{t_i^n}H_{t_j^n}\big((\Delta_i W)^2 - t/2^n\big)0|\mathcal{F}_0] & \text{if } i<j \end{cases} \\
&= \lim_{n\to\infty}\sum_{i,j=0}^{2^n-1} \begin{cases} 0 & \text{if } i\neq j \\ \mathbb{E}[(H_{t_i^n})^2(3(t/2^n)^2 + (t/2^n)^2 - 2(t/2^n)^2)|\mathcal{F}_0] & \text{if } i=j \end{cases} \\
&= \lim_{n\to\infty}\sum_{i=0}^{2^n-1} \mathbb{E}[(H_{t_i^n})^2(2(t/2^n)^2)|\mathcal{F}_0] \\
&= \int_0^t 2\mathbb{E}[(H_s)^2|\mathcal{F}_0]\,(ds)^2 \\
&= 0
\end{aligned}
$$

We therefore have $\mathbb{E}[\int_0^t H_s(dW_s)^2 - \int_0^t H_s\,ds] = 0$ and $\mathrm{Var}[\int_0^t H_s(dW_s)^2 - \int_0^t H_s\,ds] = 0$ and therefore $\mathbb{P}(\int_0^t H_s(dW_s)^2 = \int_0^t H_s\,ds) = 1$ i.e.

$$\int_0^t H_s(dW_s)^2 = \int_0^t H_s\,ds \quad \text{almost surely}$$

Remark (Itô Differentials)**.** The above result is also written in differential notation as $(dW)^2 = dt$. We can similarly check that $dtdW = 0$ and $(dW)^3 = 0$. Therefore we have the algebra on

differentials:

$$(dt)^2 = 0$$
$$(dW)^2 = dt$$

The identity $(dW)^2 = dt$ is a statement about Brownian motion in particular, and chapter 2 needs the corresponding statement for processes that are not Brownian motion. The object it names is the accumulated sum of squared increments, which is what the figure above was plotting.

Definition 1.26 (Quadratic variation)**.** Let X be a continuous adapted process and let Π_n be a sequence of partitions $0 = t_0^n < t_1^n < \cdots < t_{k_n}^n = t$ of $[0,t]$ whose mesh $\max_i(t_{i+1}^n - t_i^n)$ tends to zero. The *quadratic variation* of X is

$$\langle X\rangle_t = \lim_{n\to\infty} \sum_{i=0}^{k_n-1} \left(X_{t_{i+1}^n} - X_{t_i^n}\right)^2,$$

the limit taken in probability, when it exists and does not depend on the choice of partitions.

For a continuous local martingale the limit always exists. That is the Doob–Meyer decomposition, and stated in full generality — for submartingales, with no continuity assumed — its proof is a long argument about weak compactness. The case these notes use is the one whose proof is short, and it is short because it is the argument that built the Itô integral, run a second time.

Theorem 1.27 (The bracket of a continuous square integrable martingale)**.** *Let M be a continuous martingale with $M_0 = 0$ and $\mathbb{E}[M_t^2] < \infty$ for every t. Then the sums of squared increments converge in L^2, so $\langle M\rangle_t$ exists. It is continuous, non-decreasing, adapted and zero at the origin; $M^2 - \langle M\rangle$ is a martingale; and $\langle M\rangle$ is the only process with those properties.*

Proof. Write $Q_n(t) = \sum_i \left(M_{t_{i+1}^n} - M_{t_i^n}\right)^2$ for the sum over the partition Π_n of $[0,t]$. Everything follows from one piece of algebra. Since $b^2 - a^2 = (b-a)^2 + 2a(b-a)$, summing over the partition telescopes the left side and gives

$$M_t^2 = Q_n(t) + 2\sum_i M_{t_i^n}\left(M_{t_{i+1}^n} - M_{t_i^n}\right). \tag{1.3}$$

The last term is a martingale in t for each fixed partition, being a sum of increments each multiplied by a factor known before its increment occurs — the same predictability that gives the Itô integral zero mean. Taking expectations, $\mathbb{E}[Q_n(t)] = \mathbb{E}[M_t^2]$ at every resolution, so the sums are not merely converging to the right object, they have the right expectation all the way along.

That also identifies what Q_n is: by (1.3) it is M_t^2 minus a discrete stochastic integral of M against itself, so its convergence is the convergence of that integral, which is the construction of the Itô integral and nothing new. Refining the partition changes the integral by an integral of the difference between M and its piecewise-constant approximation, the isometry bounds that in L^2 by the size of the approximation error, and continuity of the path drives the error to zero uniformly on $[0,t]$. So $\{Q_n(t)\}$ is Cauchy in L^2; L^2 is complete, so the limit exists; and the same bound shows it is independent of the sequence of partitions. Passing to the limit in (1.3) leaves $M^2 - \langle M\rangle$ as an L^2 limit of martingales, hence a martingale.

The path properties come along with it. Each Q_n is non-decreasing in t, and a limit of non-decreasing functions is non-decreasing; adaptedness is inherited term by term; and continuity follows from the same L^2 bound, which controls the increase of $\langle M\rangle$ across an interval by the oscillation of M on it.

Uniqueness needs no new machinery. If A and A' both qualify then $A-A' = (M^2-A')-(M^2-A)$ is a difference of martingales, hence a martingale; it is also continuous, zero at the origin, and of finite variation — the total variation of a function over $[0,t]$ is the supremum of $\sum_i |f(t_{i+1})-f(t_i)|$ over partitions $0 = t_0 < \cdots < t_k = t$; for a non-decreasing f the absolute values are redundant and the sum telescopes to $f(t)-f(0)$ whatever the partition, so A and A' each have finite variation, and $A-A'$ has total variation at most the sum of theirs. By the lemma below such a process has zero quadratic variation, and a continuous martingale starting at zero whose bracket vanishes is identically zero. So $A = A'$. □

Two things about the scope of that. Localisation extends it from square integrable martingales to local martingales in the usual way — stop at each τ_n, take the bracket of the stopped process, and observe that the brackets agree where they overlap because each is a limit of the same sums. And it does not require M to be an Itô integral, which matters, because the Cantor-function example below is a martingale it covers and an integral it does not. What we do still need to check is which parts of a process contribute.

Lemma 1.28 (Only the martingale part contributes)**.** *Let A be continuous and of finite variation on $[0,t]$, and let X be continuous with quadratic variation $\langle X\rangle_t$. Then $\langle A\rangle_t = 0$ and $\langle X+A\rangle_t = \langle X\rangle_t$.*

Proof. Write V for the total variation of A on $[0,t]$ and $\omega(\delta)$ for its modulus of continuity. Then

$$\sum_i (A_{t_{i+1}^n} - A_{t_i^n})^2 \le \max_i |A_{t_{i+1}^n} - A_{t_i^n}| \sum_i |A_{t_{i+1}^n} - A_{t_i^n}| \le \omega(\operatorname{mesh}\Pi_n)\, V,$$

and A is uniformly continuous on the compact interval, so $\omega(\operatorname{mesh}\Pi_n)\to 0$. For the second claim expand the square and bound the cross term by Cauchy–Schwarz,

$$\Big|2\sum_i (X_{t_{i+1}^n} - X_{t_i^n})(A_{t_{i+1}^n} - A_{t_i^n})\Big| \le 2\Big(\sum_i (\Delta X_i)^2\Big)^{1/2}\Big(\sum_i (\Delta A_i)^2\Big)^{1/2},$$

whose second factor tends to zero while the first stays bounded. □

Definition 1.29 (Semimartingale)**.** A càdlàg adapted process is a *semimartingale* if it is the sum of a local martingale and a finite-variation process. This chapter works entirely with the continuous case: a stochastic integral against W plus a drift of the form $\int_0^t \mu_s\,ds$. The lemma just proved is the statement that the quadratic variation sees only the local martingale summand. That a price must be a semimartingale — on pain of arbitrage — is a theorem of chapter 4.

So the drift of a diffusion is invisible to $\langle X\rangle$ — which is the content of $(dt)^2 = 0$ and $dt\,dW = 0$ — and the quadratic variation measures the martingale part alone. For the martingale part itself the two calculations above have already done the work: if $X_t = \int_0^t H_s\,dW_s$ then the increments are

$\Delta X_i \approx H_{t_i}\Delta W_i$, and summing their squares is summing $H_{t_i}^2(\Delta W_i)^2$, which we showed converges to $\int_0^t H_s^2\,ds$. Hence

$$X_t = \int_0^t H_s\,dW_s \quad\Longrightarrow\quad \langle X\rangle_t = \int_0^t H_s^2\,ds. \tag{1.4}$$

One consequence recurs whenever two candidate martingales have to be shown equal: a continuous local martingale M with $M_0 = 0$ and $\langle M\rangle = 0$ vanishes identically. Indeed $M^2 - \langle M\rangle = M^2$ is then a non-negative local martingale, hence a supermartingale, so $\mathbb{E}[M_t^2] \le M_0^2 = 0$.

Definition 1.30 (The differential of the quadratic variation)**.** Since almost every path of $\langle X\rangle$ is a continuous non-decreasing function starting at zero, it defines a measure on $[0,\infty)$ by $d\langle X\rangle\big((s,u]\big) = \langle X\rangle_u - \langle X\rangle_s$, and $\int_0^t H_s\,d\langle X\rangle_s$ means the Lebesgue–Stieltjes integral of H against that measure, computed path by path.

Remark (What $d\langle X\rangle_t$ is and is not)**.** It is tempting to read $d\langle X\rangle_t$ as "some multiple of dt", on the grounds that the only thing a differential in t can be proportional to is dt. Two things hide inside that reading.

First, the multiple is random. For $X = \int H\,dW$, equation (1.4) gives $d\langle X\rangle_t = H_t^2\,dt$, and H_t is an adapted process, not a number: the density is a random variable that the path itself determines, and it can depend on the same noise that drives X. That density is exactly what a stochastic volatility model models.

Second, there need not be a multiple at all. A measure carries no obligation to be absolutely continuous with respect to Lebesgue measure. Let A be the Cantor function, continuous and non-decreasing with $A' = 0$ almost everywhere, and set $M_t = B_{A_t}$ for a Brownian motion B. Then M is a continuous martingale in the time-changed filtration with $\langle M\rangle = A$, so its quadratic variation increases while having no density: all of the randomness arrives on a Lebesgue-null set of times. Writing an equation $dX_t = \mu_t\,dt + \sigma_t\,dW_t$ rules such a process out. The ansatz is not only a choice of drift and volatility; it is already the assumption that the quadratic variation accumulates at a rate.

What the bracket *means* has to wait for the theorem that establishes it, which is chapter 2's: every continuous local martingale turns out to be a Brownian motion watched on a clock, and $\langle M\rangle$ is the clock's reading. Until then it is enough that it measures accumulated randomness, and that a diffusion's drift contributes none of it.

Two processes have a covariation for the same reason one has a variation.

Definition 1.31 (Covariation)**.** For continuous adapted X and Y,

$$\langle X,Y\rangle_t = \lim_{n\to\infty} \sum_{i=0}^{k_n-1} \big(X_{t_{i+1}^n} - X_{t_i^n}\big)\big(Y_{t_{i+1}^n} - Y_{t_i^n}\big) = \tfrac14\big(\langle X+Y\rangle_t - \langle X-Y\rangle_t\big),$$

the second equality being the polarisation identity. It is symmetric and bilinear, and $\langle X,X\rangle = \langle X\rangle$.

In differential notation the product of two differentials *is* the covariation, $dX\,dY = d\langle X,Y\rangle$, and the algebra of the remark above is the special case

$$dt\,dt = 0, \qquad dW\,dt = 0, \qquad dW^i\,dW^j = \rho_{ij}\,dt,$$

for correlated Brownian motions W^i, since $\langle W^i, W^j\rangle_t = \rho_{ij}t$.

From the ordinary calculus we have the Taylor expansion of a function as

$$f((x) + \boldsymbol{\delta x}) = f(\boldsymbol{x}) + f'(\boldsymbol{x}) \cdot \boldsymbol{\delta x} + \frac{1}{2}(\boldsymbol{\delta x})^T \cdot f'' \cdot \boldsymbol{\delta x} + \dots$$

where f' is the gradient and f'' is the Hessian. When taking the limit $\delta x \to dx$ we often ignore the second and higher order differentials $(dx)^2, (dx)^3, \dots$ giving the usual chain rule. In stochastic calculus, however, as we have seen second order differentials need not be 0. And we have:

Remark (Itô's Lemma)**.**

$$df(\boldsymbol{x}) = f'(\boldsymbol{x}) \cdot \boldsymbol{dx} + (\boldsymbol{dx})^T \cdot f'' \cdot \boldsymbol{dx} \tag{1.5}$$

and in the special case when f is a function of t and a stochastic process X_t.

$$df(t, X_t) = \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial X_t}dX_t + \frac{1}{2}\frac{\partial^2 f}{\partial X_t^2}(dX_t)^2$$

where the second order term is $(dX_t)^2 = d\langle X\rangle_t$, so the lemma applies to any continuous process whose quadratic variation exists, and not only to functions of W. When X_t is a Brownian motion W_t we have

$$df(t, X_t) = \big(\frac{\partial f}{\partial t} + \frac{1}{2}\frac{\partial^2 f}{\partial W_t^2}\big)dt + \frac{\partial f}{\partial W_t}dW_t$$

Definition 1.32 (Itô Diffusion)**.** An Itô diffusion is a stochastic process satisfying a stochastic differential equation of the form

$$d\boldsymbol{X}_t = \underbrace{\boldsymbol{\mu}(t, \boldsymbol{X}_t)dt}_{\text{drift}} + \underbrace{\boldsymbol{\sigma}(t, \boldsymbol{X}_t) \cdot d\boldsymbol{W}_t}_{\text{diffusion}} \tag{1.6}$$

The diffusion coefficient in the equation above is also known as volatility.

A stochastic integral against a Brownian motion always has zero drift, but its expectation can still fail to be constant if the integrand is wild enough, which is why the conclusion is local.

Remark. An Itô diffusion with zero drift,

$$d\boldsymbol{X}_t = \boldsymbol{\sigma}(t, \boldsymbol{X}_t) \cdot d\boldsymbol{W}_t,$$

is a local martingale. The construction above made $\int H\,dW$ a true martingale when $\mathbb{E}\int_0^T H_s^2\,ds < \infty$. When only $\int_0^t H_s^2\,ds < \infty$ almost surely, stop at $\tau_n = \inf\{\, t : \int_0^t H_s^2\,ds \geq n \,\}$: the integrand of $\int_0^{t\wedge\tau_n} H\,dW$ then has $\mathbb{E}\int_0^T H_s^2 \mathbf{1}_{\{s\leq\tau_n\}}\,ds \leq n$, so each stopped integral is a martingale, and $\tau_n \uparrow \infty$. So a stochastic integral is a local martingale under the weak condition and a true martingale under the strong one.

Example 1.2 (Geometric Brownian Motion)**.**

$$\frac{dX_t}{X_t} = \mu dt + \sigma dW_t \tag{1.7}$$

$$
\begin{aligned}
d\log(X_t) &= \frac{dX_t}{X_t} - \frac{1}{2}\frac{1}{X_t^2}(dX_t)^2 \qquad \text{(by Itô's Lemma)} \\
&= \mu dt + \sigma dW_t - \frac{1}{2}\sigma^2 dt \\
&= \big(\mu - \frac{1}{2}\sigma^2\big)dt + \sigma dW_t \\
\log(X_t) - \log(X_0) &= \int_0^t \big(\mu - \frac{1}{2}\sigma^2\big)ds + \int_0^t \sigma\, dW_s \\
&= \big(\mu - \frac{1}{2}\sigma^2\big)t + \sigma W_t \\
X_t &= X_0 e^{(\mu - \frac{1}{2}\sigma^2)t + \sigma W_t}
\end{aligned}
$$

Because of its simplicity Geometric Brownian Motion has been widely used in finance as a model for evolution of stock prices.

References

- Protter, P. (2005). *Stochastic Integration and Differential Equations.* Springer.

Chapter 2

Price Process Characterization

We ask which stochastic processes are available to model a price, and find the answer far more restrictive than it looks. Brownian motion turns out to be the only continuous noise there is, which forces the diffusion term of the previous chapter's equation rather than merely permitting it — the licence to write the equation at all comes from chapter 4, and the two halves are worth keeping distinct. Dropping continuity admits jumps, and we work the Poisson process through in enough detail to see how a default is modelled and why that case is so much more tractable than the general one.

2.1 Why This Form and No Other

Everything from chapter 5 onwards models a price as an Itô diffusion. Before that becomes a habit, ask how much of a restriction that is. The equation has a specific shape — a drift proportional to dt, a random part proportional to dW, and nothing else. Why should any price be of that shape?

The answer is that most of the shape is forced, but by two different things, and they should be kept apart from the start.

That the random part is proportional to dW — that there is no other continuous noise to choose from — is what this chapter proves, and it needs continuity together with the martingale property. That a price may be written in this shape *at all*, with a drift accumulating at a rate alongside the noise, is a separate question and continuity does not answer it: it needs the price to be a semimartingale, which is a consequence of no-arbitrage rather than of smooth paths. Chapter 4 supplies that half and states the combined result.

So what follows pins the noise, and not the drift.

Write $\langle M \rangle_t$ for the quadratic variation defined in chapter 1, the limit of the sums of squared increments over refining partitions, which for Brownian motion is t.

Theorem 2.1 (Lévy's characterization)**.** *Let M be a continuous local martingale with $M_0 = 0$. Then M is a standard Brownian motion if and only if $\langle M \rangle_t = t$ for all t.*

Sketch. Fix $\lambda \in \mathbb{R}$ and consider the exponential

$$Z_t = e^{\lambda M_t - \frac{1}{2}\lambda^2 t}.$$

Applying Itô's lemma, and remembering that the second order term carries $d\langle M\rangle_t$ rather than dt,

$$dZ_t = Z_t\Big(\lambda\, dM_t - \tfrac{1}{2}\lambda^2\, dt + \tfrac{1}{2}\lambda^2\, d\langle M\rangle_t\Big).$$

Now impose the hypothesis $\langle M\rangle_t = t$. The last two terms cancel exactly, leaving

$$dZ_t = Z_t \lambda\, dM_t,$$

which has no drift, so Z is a martingale. Therefore $\mathbb{E}[Z_t \mid \mathcal{F}_s] = Z_s$, and rearranging,

$$\mathbb{E}\big[e^{\lambda(M_t - M_s)} \,\big|\, \mathcal{F}_s\big] = e^{\frac{1}{2}\lambda^2(t-s)}. \tag{2.1}$$

The right hand side is the moment generating function of a $N(0, t-s)$ variable, so the increment is Gaussian with variance $t-s$. And the right hand side is a *number*: it does not depend on $\mathcal{F}_s$ at all, so the increment is independent of everything that has happened. Gaussian, independent increments, variance $t-s$ — that is Brownian motion. □

Remark (Where the independence came from)**.** Look again at which step produced it. Equation (2.1) has a deterministic right hand side for exactly one reason: $d\langle M\rangle_t$ cancelled against dt. Without the hypothesis there is nothing to cancel against, and the exponential that stays a martingale is the one that carries its own quadratic variation in the exponent, $\tilde{Z}_t = e^{\lambda M_t - \frac{1}{2}\lambda^2\langle M\rangle_t}$, which gives

$$\mathbb{E}\Big[e^{\lambda(M_t - M_s) - \frac{1}{2}\lambda^2(\langle M\rangle_t - \langle M\rangle_s)} \,\Big|\, \mathcal{F}_s\Big] = 1.$$

This is all that survives, and the two factors cannot be separated: the increment and the variance it accumulates are in general dependent, and in chapter 10 that dependence is the leverage correlation which produces the skew. Only when $\langle M\rangle_t - \langle M\rangle_s$ is deterministic can the second factor be taken out of the expectation, and it is that step, and no other, which turns the identity into a statement about the law of the increment alone. Independence of increments is not a general feature of continuous martingales; it is what a deterministic quadratic variation buys.

One honest caveat on the sketch: using the real exponential requires an integrability argument to promote Z from a local martingale to a true one. The standard proof uses $e^{i\lambda M_t + \frac{1}{2}\lambda^2 t}$ instead, whose modulus is bounded so that the promotion is free, and reads off the characteristic function rather than the moment generating function. Nothing else changes.

What the theorem says is: Brownian motion is not one choice of noise among many. It is the *only* continuous local martingale whose quadratic variation accumulates at a constant rate, and any other candidate is either not continuous, not a martingale, or a rescaling of this one in disguise. The next proposition makes "in disguise" precise.

Theorem 2.2 (A continuous local martingale is an integral against Brownian motion)**.** *Let M be a continuous local martingale with $M_0 = 0$ whose quadratic variation is absolutely continuous,*

$$\langle M\rangle_t = \int_0^t \sigma_s^2\, ds,$$

with $\sigma_s > 0$. Then there is a Brownian motion W such that

$$M_t = \int_0^t \sigma_s \, dW_s.$$

Proof. Define W by integrating the increments of M scaled down by σ:

$$W_t = \int_0^t \frac{1}{\sigma_s} \, dM_s.$$

It is a stochastic integral against a local martingale, so it is itself a continuous local martingale starting from zero. Its quadratic variation is

$$\langle W \rangle_t = \int_0^t \frac{1}{\sigma_s^2} \, d\langle M \rangle_s = \int_0^t \frac{1}{\sigma_s^2} \, \sigma_s^2 \, ds = t,$$

using that the quadratic variation of $\int H \, dM$ is $\int H^2 \, d\langle M \rangle$. So W satisfies the hypotheses of Theorem 2.1 and is a Brownian motion. And integrating it back,

$$\int_0^t \sigma_s \, dW_s = \int_0^t \sigma_s \frac{1}{\sigma_s} \, dM_s = M_t. \qquad \square$$

Remark. The trick recurs in chapter 6. A process whose quadratic variation grows at a varying rate is a Brownian motion running on a distorted clock, and dividing by σ resets the clock. Lévy's theorem is what certifies that the reset produces a genuine Brownian motion rather than merely something with the right variance.

Two conditions in the statement are doing work. If σ vanishes on a set of times of positive measure, M is simply not moving there and the construction needs an independent Brownian motion added on that set to fill the gap; the conclusion survives, on a slightly enlarged probability space. And if $\langle M \rangle$ is not absolutely continuous — if the process accumulates variance in bursts concentrated on a set of measure zero — then M is still a time-changed Brownian motion, by the Dambis-Dubins-Schwarz theorem, but not an integral against one. Such processes are unusual in practice and this is the only place they will be mentioned.

The general case is proved next because the argument is the same idea as the one above with the integral removed, and because the picture it produces — the clock of chapter 1 — is used repeatedly later.

Theorem 2.3 (Dambis, Dubins and Schwarz)**.** *Let M be a continuous local martingale with $M_0 = 0$ and $\langle M \rangle_\infty = \infty$. Set*

$$\tau_u = \inf\{t : \langle M \rangle_t > u\}.$$

Then $B_u = M_{\tau_u}$ is a Brownian motion with respect to the filtration $\mathcal{G}_u = \mathcal{F}_{\tau_u}$, and

$$M_t = B_{\langle M \rangle_t}.$$

Proof. One lemma first. *If $\langle M \rangle$ is constant on an interval then so is M.* For if $\langle M \rangle_b = \langle M \rangle_a$ with $a < b$, then the local martingale $t \mapsto M_{a+t} - M_a$ has vanishing quadratic variation on $[0, b-a]$, and a continuous local martingale starting from zero with vanishing bracket is identically zero — the

supermartingale argument of chapter 1. So M does not move where its clock does not run, which is what one would want: the bracket measures the accumulated randomness, and no accumulation means nothing happened.

Now the properties of τ. Because $\langle M\rangle$ is continuous, non-decreasing and increases to infinity, τ_u is finite for every u, non-decreasing, right-continuous, and satisfies

$$\langle M\rangle_{\tau_u} = u.$$

It fails to be continuous exactly where $\langle M\rangle$ is flat: an interval on which the clock does not advance is crossed by τ in a single jump. By the lemma M is constant across such an interval, so $B_u = M_{\tau_u}$ does not jump there, and B is continuous. This is the step the theorem's proof exists to handle, and it is why the hypothesis is about $\langle M\rangle_\infty$ rather than about M.

B is adapted to $\mathcal{G}$ by construction. It is a local martingale because optional stopping says so: for $v \geq u$, applying it to M between the stopping times τ_u and τ_v gives $\mathbb{E}[M_{\tau_v} \mid \mathcal{F}_{\tau_u}] = M_{\tau_u}$, which is $\mathbb{E}[B_v \mid \mathcal{G}_u] = B_u$. The integrability that optional stopping wants is supplied by localising, exactly as elsewhere: stop at $\tau_u \wedge n$ and let n grow.

Its bracket is the time change of the bracket, $\langle B\rangle_u = \langle M\rangle_{\tau_u} = u$, since the sums of squared increments defining $\langle B\rangle$ over a partition of $[0,u]$ are the sums defining $\langle M\rangle$ over the image partition of $[0,\tau_u]$ — the same increments of the same path, relabelled. So $\langle B\rangle_u = u$, and Theorem 2.1 applies: B is a Brownian motion.

Finally the identity. Fix t and put $u = \langle M\rangle_t$. Then $\tau_u \geq t$, with strict inequality only when $\langle M\rangle$ is flat on $[t,\tau_u]$ — and by the lemma M is then constant there, so $M_{\tau_u} = M_t$ either way. That is $B_{\langle M\rangle_t} = M_t$, which is the square. □

Structure (Quadratic variation is a clock)**.** Every continuous local martingale is a Brownian motion watched on a clock of its own making, and $\langle M\rangle_t$ is the reading of that clock at calendar time t. Volatility is the speed of the clock: high volatility is not a different noise but the same noise consumed faster.

Several later statements are this one in disguise. Lévy's characterization is the case where the clock runs at calendar speed, and Theorem 2.2 the case where it has a rate. A stochastic volatility model (chapter 10) is a model of the clock, with v_t the rate at which it turns. A skew is what appears when the clock is correlated with the motion: if time speeds up on the way down, the terminal distribution cannot be symmetric. And chapter 4's doubling strategy is a clock deliberately built to reach infinity before the calendar reaches one.

Structure (The theorems above are a classification)**.** Read together, Lévy's characterization and the representation theorem do something stronger than supply two facts. They classify.

Up to a time change, there is only one continuous local martingale. Dambis-Dubins-Schwarz says every one of them is $B_{\langle M\rangle_t}$ for a Brownian motion B; Lévy says B is unique in law; and the representation theorem is the special case where the clock is absolutely continuous, so the time change can be written as an integral.

A classification beats a list because it tells you when to stop looking. There is no exotic continuous martingale waiting to be discovered and used to model a price differently — the class is exhausted by one object and a choice of clock. All the modelling freedom that remains lives in

the clock, which is to say in the volatility, and that is why every chapter after this one is about volatility rather than about finding a better noise.

Remark (Where the deterministic quadratic variation went)**.** Lévy's theorem required $\langle M\rangle_t = t$, which is about as deterministic as a quadratic variation gets. A stochastic volatility model has $\langle M\rangle_t = \int_0^t \sigma_s^2\,ds$ with σ random, which is not deterministic at all. How is the theory available to such a model?

The answer is that Lévy's theorem was never applied to the price. Read the proof again and watch which process it is applied to. The hypothesis of Theorem 2.2 allows σ_s to be any adapted process — as random as one likes, driven by its own noise, depending on M's whole history. Nothing deterministic is assumed. What the proof then does is compute

$$\langle W\rangle_t = \int_0^t \frac{1}{\sigma_s^2}\,d\langle M\rangle_s = \int_0^t \frac{1}{\sigma_s^2}\,\sigma_s^2\,ds = t,$$

and the randomness cancels identically. Not approximately, and not on average: the σ_s^2 that arrived from $d\langle M\rangle_s$ is the same σ_s^2 that was divided by, whatever it happened to be on that path.

So the deterministic quadratic variation Lévy needs is *manufactured* rather than assumed. Dividing by σ is precisely the operation that removes the randomness from the variation, and W is the process it is removed from. The price keeps its random variation and its dependent increments; W takes the deterministic variation and the independence.

In the language of chapter 1 the price is B read on a random clock, and reading a Brownian motion at random times is what destroys the independence: knowing the past tells you how fast the clock is currently running, and therefore how large the next increment is likely to be. Lévy's hypothesis is the statement that the clock cannot surprise you.

Remark (Checking which process is which)**.** Both claims are measurable, on the same simulated paths, in `variation_spread`. Take the realised variance over $[0, 1]$ and record how much it varies from path to path, as the sampling grid is refined:

	400 **steps**	1600 **steps**
Price, constant σ	0.071	0.035
Price, 100% vol-of-vol	2.57	2.31
Recovered W, either case	0.071	0.035

The figures are standard deviations across paths as a fraction of the mean. Read the rows against each other. With constant volatility the price's own variation concentrates as the grid is refined, halving as the step count quadruples — it is converging to a constant. With random volatility it does not: concentration would have taken 2.57 to 1.29 and it went to 2.31, which is the estimate of a heavy-tailed quantity wobbling rather than a distribution collapsing. Its limit is the random variable $\int_0^1 \sigma_s^2\,ds$, and refining the grid measures that random number more precisely rather than making it less random.

And the recovered W concentrates in both cases, at exactly the rate $\sqrt{2/n}$ that a chi-square average must. Its variation is converging to the constant 1 however wildly the volatility behaved, which is Lévy's hypothesis being satisfied by construction. That is the whole content of the previous two remarks, in numbers.

So the shape of the Itô diffusion is not chosen. Any continuous local martingale, given only that it accumulates variance at some rate, *must* be $\int \sigma\, dW$; and adding a finite-variation drift then gives the general form. Chapter 4 completes the argument by explaining why a price has to be a local martingale in the first place.

Remark (The representation does not hand down independent increments)**.** Theorem 2.2 is easy to over-read. It says every continuous local martingale is an integral against a Brownian motion, and Brownian motion has independent increments, so it is tempting to conclude that M inherits them. It does not, and the reason is that σ_s is an *adapted* process — it is allowed to depend on everything that has happened up to time s, including M itself.

The independence lives in W, not in M. What the theorem provides is a Brownian motion out of which M can be rebuilt; all of M's memory has been pushed into the integrand. A local volatility model $\sigma_s = \sigma(M_s, s)$ is the plainest case: its increments obviously depend on the past, since they depend on where the process currently is, and it is nonetheless exactly of the form $\int \sigma\, dW$.

The previous remark already gives the sharp version. For a continuous local martingale,

$$M \text{ has independent (and Gaussian) increments} \iff \langle M \rangle_t \text{ is deterministic}$$

which for $M = \int \sigma\, dW$ means σ deterministic. Anything else — any model in which volatility is itself random — has dependent increments by construction. That is the whole of chapters 10 to 12.

Remark (How restrictive independent increments would be)**.** Very little would survive the requirement. A continuous process with stationary independent increments must be $\mu t + \sigma W_t$ for constants μ and σ — Brownian motion with drift, and nothing else. Relaxing stationarity but keeping independence buys only deterministic $\mu(t)$ and $\sigma(t)$.

Remark (Checking it)**.** The claim is concrete enough to measure. Take $M = \int \sigma\, dW$ on $[0, 1]$, split the interval in half, and correlate the realised variance of the first half against the second across many paths. Increments independent of the past would force this to be zero, since the two windows are disjoint. Running it in `variance_clustering`:

Volatility of volatility	Correlation
0 (constant σ)	-0.00
50%	$+0.67$
100%	$+0.54$

Zero when the volatility is constant, as it must be, and emphatically not zero otherwise — and M is a perfectly good continuous local martingale in every row, since the volatility is driven by its own independent Brownian motion. Under its market name the second row is volatility clustering, which is among the least controversial facts about returns.

The non-monotonicity in the table is real and is not the dependence weakening. It is the linear correlation losing its ability to see the dependence, because the realised variances become more extremely lognormal — the same effect chapter 18 derives as a hard bound on how correlated two lognormals can be. It is an early warning that a correlation is a poor summary of a dependence.

2.2 Dropping Continuity

Every statement in the previous section carried the word *continuous*, and it is fair to ask how much it was doing. Prices in fact jump: a bond that defaults, a stock through an earnings announcement, a currency whose peg breaks. If continuity were a technical convenience the theory would extend and nobody would mind. It is not, and the cheapest way to see that is a counterexample.

Definition 2.4 (Poisson process)**.** A Poisson process N_t with intensity λ starts at zero, has independent increments, and $N_t - N_s$ is Poisson distributed with mean $\lambda(t-s)$. Its paths are constant except for jumps of size 1, arriving at rate λ.

Writing the Poisson law into the definition makes it look like a choice among many, and it is not. The distribution is forced by the same two hypotheses that force everything else in this chapter.

Theorem 2.5 (Counting with stationary independent increments leaves no choice)**.** *Let N be a counting process — starting at zero, non-decreasing, moving only by integer jumps — with stationary independent increments, and suppose two jumps do not arrive at once, in the sense that*

$$\mathbb{P}(N_h \geq 2) = o(h) \qquad \text{as } h \downarrow 0. \tag{2.2}$$

Then N is a Poisson process: there is a $\lambda \geq 0$ with N_t Poisson distributed of mean λt.

Proof. Everything comes from the probability of *nothing happening*. Put $p(t) = \mathbb{P}(N_t = 0)$. The event that nothing happens on $[0, t+s]$ is the intersection of nothing happening on $[0,t]$ and nothing happening on $(t, t+s]$, and those are independent by hypothesis and identically distributed to $[0,s]$ by stationarity. So

$$p(t+s) = p(t)\,p(s),$$

which is Cauchy's exponential equation. A monotone solution — and p is non-increasing, since nothing happening on a longer interval is a smaller event — must be $p(t) = e^{-\lambda t}$ for some $\lambda \in [0,\infty]$. This is the same functional equation, and the same conclusion, that chapter 12 meets when asking which volatility structures admit a finite-dimensional state; there it forces an exponential in maturity, here in time.

That fixes the law of the first arrival: $\mathbb{P}(\tau_1 > t) = e^{-\lambda t}$, so the waiting time is exponential, and by stationarity and independence each subsequent wait is an independent copy. A sum of n independent exponentials is Gamma distributed, and $\{N_t \geq n\} = \{\tau_1 + \cdots + \tau_n \leq t\}$, so

$$\mathbb{P}(N_t \geq n) = \int_0^t \frac{\lambda^n u^{n-1} e^{-\lambda u}}{(n-1)!}\, du,$$

and differencing consecutive values of n gives $\mathbb{P}(N_t = n) = e^{-\lambda t}(\lambda t)^n / n!$.

Condition (2.2) is what rules out the alternatives. Without it the jumps need not have size one — a process that jumps by two at Poisson times has stationary independent increments and is not a Poisson process — so (2.2) is precisely the statement that the counting is of single events. The degenerate endpoints are $\lambda = 0$, where nothing ever happens, and $\lambda = \infty$, where $p(t) = 0$ for every t and infinitely many arrivals occur immediately, which the right-continuity of a counting path excludes. □

Remark (What the theorem is for)**.** Read alongside Lévy's characterization it completes a pair. There, continuity plus a constant rate of accumulated variance forced Brownian motion and left no freedom. Here, counting plus stationary independent increments forces the Poisson process and leaves one number. So each of the two building blocks of (2.10) below is not a modelling choice but the unique object satisfying its description, which is why a general process with stationary independent increments can be decomposed into them and nothing else.

It is also the step the Lévy-Itô proof leans on hardest, and now it need not be waved at: applying the theorem to the events "a jump of size at least ϵ occurred" gives that those arrivals are Poisson, with a rate that must be finite because a right-continuous path cannot have infinitely many jumps of a fixed size in a bounded interval.

It is not a martingale — it only ever goes up — but subtracting its mean makes one.

Calculation 2.6 (The compensator)**.** Look for a deterministic function $a(t)$ making $N_t - a(t)$ a martingale. Splitting the increment and using independence,

$$\begin{aligned}\mathbb{E}\big[N_t - a(t) \,\big|\, \mathcal{F}_s\big] &= \mathbb{E}\big[N_s + (N_t - N_s) \,\big|\, \mathcal{F}_s\big] - a(t) \\ &= N_s + \mathbb{E}\big[N_t - N_s\big] - a(t) && \text{(independent increments)} \\ &= N_s + \lambda(t-s) - a(t).\end{aligned}$$

For this to equal $N_s - a(s)$ we need $a(t) - a(s) = \lambda(t-s)$ for all $s < t$, so $a(t) = \lambda t$ up to a constant.

Definition 2.7 (Compensated Poisson process)**.** $M_t = N_t - \lambda t$, which by the calculation above is a martingale.

The compensator λt is doing the same job as the drift correction elsewhere: it removes the part of the motion that was predictable and leaves only the surprise. Note how little was used — only that the increments are independent with mean $\lambda(t-s)$ — which is why the same subtraction works for any process with independent increments.

Now its two quadratic variations, which is where the jump case starts to diverge from everything in the previous chapter.

Calculation 2.8 (Realised quadratic variation)**.** Take any partition and refine it. Between jumps the path is flat, so those increments contribute nothing. Once the partition is fine enough that no two jumps share an interval, each interval containing a jump contributes $(\Delta M)^2 = (\pm 1)^2 = 1$ — the $-\lambda\, dt$ part of the increment contributes at order dt^2 and vanishes in the limit. Counting,

$$[M]_t = \sum_{s \le t} (\Delta M_s)^2 = N_t.$$

So the realised quadratic variation is the number of jumps: an integer, a staircase, and *random*.

Calculation 2.9 (Predictable quadratic variation)**.** Now find the deterministic $b(t)$ making $M_t^2 - b(t)$ a martingale. Split the increment again, using that M is a martingale so the cross term drops:

$$\begin{aligned}\mathbb{E}\big[M_t^2 \,\big|\, \mathcal{F}_s\big] &= \mathbb{E}\big[\big(M_s + (M_t - M_s)\big)^2 \,\big|\, \mathcal{F}_s\big] \\ &= M_s^2 + 2M_s \underbrace{\mathbb{E}\big[M_t - M_s \,\big|\, \mathcal{F}_s\big]}_{=\,0} + \mathbb{E}\big[(M_t - M_s)^2 \,\big|\, \mathcal{F}_s\big] \\ &= M_s^2 + \mathrm{Var}\big(N_t - N_s\big) && \text{(the increment has mean zero)} \\ &= M_s^2 + \lambda(t-s),\end{aligned}$$

using that a Poisson variable has variance equal to its mean. So $b(t) = \lambda t$ works, and

$$\langle M \rangle_t = \lambda t.$$

Set $\lambda = 1$ and compare the two answers: $\langle M \rangle_t = t$, exactly Brownian motion's, while $[M]_t = N_t$, which is not t and is not even deterministic. That is the counterexample below, already assembled.

Example 2.1 (Why Lévy's theorem needs continuity)**.** Take $\lambda = 1$, so $M_t = N_t - t$. Then:

- M is a martingale, hence a local martingale, and $M_0 = 0$;
- $\mathrm{Var}(N_t) = \lambda t = t$, so M accumulates variance at exactly the rate Brownian motion does, $\langle M \rangle_t = t$;
- M is emphatically not a Brownian motion. Its paths are flat except at a countable set of times where they jump by one, and its increments are Poisson rather than normal.

Every hypothesis of Lévy's characterization holds except continuity, and the conclusion fails completely. So continuity is not a regularity condition attached to make a proof go through; it is the entire content of the theorem.

Remark (Two quadratic variations, which agree only in the continuous case)**.** The two calculations above computed different things. The *realised* quadratic variation $[M]_t$ is the limit of $\sum(\Delta M)^2$ along a refining partition, the quantity chapter 1's figure measured; the *predictable* quadratic variation $\langle M \rangle_t$ is the compensator, the process subtracted from M^2 to leave a martingale. For a continuous local martingale the two coincide, which is why we have been writing one symbol without comment — and the coincidence is not an accident of notation but something chapter 1 already proved without saying so.

Its theorem on the bracket of a continuous square integrable martingale did three things: it showed the sums of squares converge, which constructs $[M]$; it showed $M^2 - [M]$ is a martingale; and it showed $[M]$ is the *only* continuous increasing adapted process starting at zero with that property. The last clause is the one that matters here, because the predictable bracket is by definition the unique *predictable* increasing process compensating M^2 — and a continuous adapted process is automatically predictable, its value at any time being the limit of its values strictly before. So $[M]$ is a candidate for the compensator, uniqueness makes it the compensator, and $[M] = \langle M \rangle$. Localisation carries it from square integrable martingales to local ones.

Continuity is doing the whole of the work in that argument, and the Poisson case shows exactly where it fails. $[M]_t = N_t$, since every jump contributes $(\Delta M)^2 = 1$, and N is adapted and increasing — but it is not predictable: its jumps arrive at the totally inaccessible times of §2.3, so no sequence announcing them exists and its value cannot be known an instant beforehand. So $[M]$ is disqualified as a compensator, and the compensator is something else: $\langle M \rangle_t = t$. The realised variance is itself random, and equal to the number of jumps.

That is a genuine difference in kind. In the continuous world, variance accumulates at a rate you can know in advance. In the jump world it arrives in lumps, and how much arrived is not known until it has.

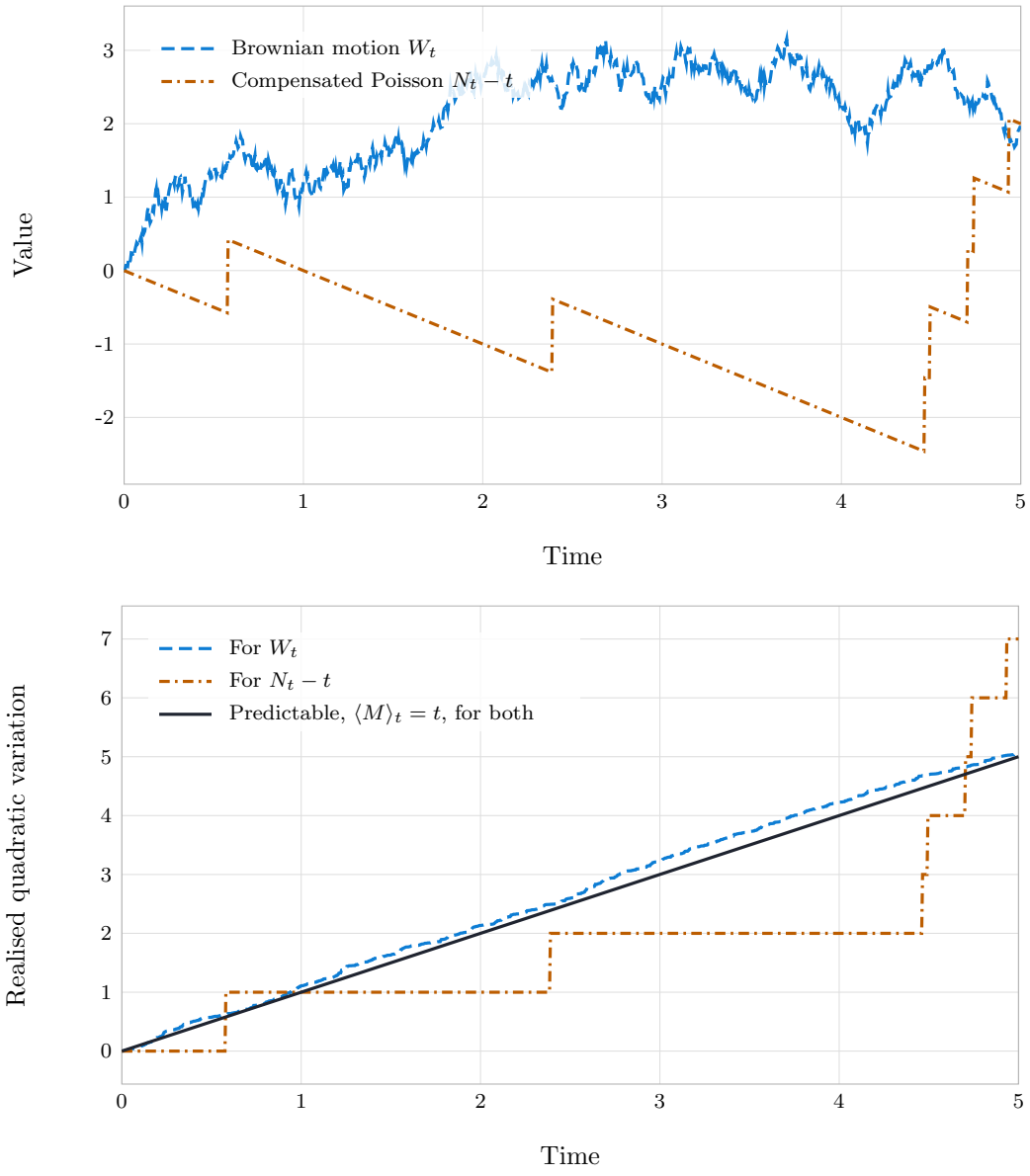

Figure 2.1: Above: a Brownian motion and a compensated Poisson process of intensity one, over five years — both martingales from zero, both accumulating variance at rate one. Below: their realised quadratic variations against the straight line that is the predictable variation of both. The Brownian motion's realised variation follows the line; the jump process's is a staircase counting the jumps, and where it sits relative to the line is not known in advance.
Drawn by `poisson_arrivals` and `compensated_poisson`.

Once continuity is abandoned the classification changes shape rather than breaking down. The general statement is that a process with stationary independent increments is a drift, plus a Brownian motion, plus jumps, and nothing else:

Theorem 2.10 (Lévy-Itô decomposition)**.** *Any Lévy process X can be written as*

$$X_t = b\,t + \sigma W_t + \underbrace{\int_{|x|<1} x\,\tilde{N}(t,dx)}_{\text{small jumps, compensated}} + \underbrace{\int_{|x|\geq 1} x\,N(t,dx)}_{\text{large jumps}},$$

where N counts the jumps of each size and $\tilde{N}$ is that count less its mean.

Structure of the proof. The pieces are not found by inspiration; each is peeled off by a step that is forced.

The large jumps. Fix $\epsilon > 0$ and count the jumps of size at least ϵ. Stationary independent increments make the arrival times of these a Poisson process, by Theorem 2.5 applied to the events "a jump of size at least ϵ occurred", and its intensity $\nu(\{|x| \geq \epsilon\})$ must be finite, because a path with infinitely many jumps of size ϵ in a bounded interval is not right-continuous. So for each ϵ the large jumps form a compound Poisson process, which can be subtracted. What is left is a Lévy process with jumps smaller than ϵ.

The small jumps. Now let $\epsilon \downarrow 0$. The measure ν need not be finite near the origin — infinitely many tiny jumps per unit time is allowed — so the sum of them may diverge. What does not diverge is the sum with its mean removed. The compensated sums over successive shells $\epsilon_{k+1} \leq |x| < \epsilon_k$ are martingales, and they are orthogonal, because disjoint jump sizes are independent. So their partial sums are a martingale whose variance is the sum of the variances, $\int_{|x|<1} x^2\,\nu(dx)$, and that integral is finite for every Lévy measure. An orthogonal series with summable variances converges in L^2; the limit is the compensated integral, and the convergence is the whole content of the small-jump term.

What remains. Subtracting both jump parts leaves a Lévy process with continuous paths. It has stationary independent increments and no jumps, so its increments are Gaussian — this is the same rigidity that Lévy's characterization exploits — and a continuous Lévy process is therefore $bt + \sigma W_t$.

Two things are glossed here and are where the real work lies: the interchange of limits each step needs, and the proof that the three pieces are independent of one another. The L^2 orthogonality behind the small-jump step is the argument that built the Itô integral in chapter 1, which is why $\int_{|x|<1} x^2\nu(dx) < \infty$ is written into the definition of a Lévy measure rather than derived here. □

Structure (Finite quadratic variation, infinite length)**.** The small jump term is a *purely discontinuous martingale*, and when ν is infinite near the origin — which is allowed, and is the interesting case — its path jumps infinitely often in every interval, with no first jump and no way to list them. "The sum of the jumps" is then not a sum one can take: $\sum |\Delta X_s|$ diverges whenever $\int_{|x|<1} |x|\,\nu(dx) = \infty$. Only the compensated sum converges, because subtracting the mean leaves a fluctuation whose size is the variance $\int_{|x|<1} x^2\,\nu(dx)$, finite for every Lévy measure — the manoeuvre that turned N into a martingale, applied to every jump size at once. Without the compensation there is nothing to converge to.

So there are two regimes, separated by a single integral:

- *Finite variation*, when $\int_{|x|<1} |x|\,\nu(dx) < \infty$. The jumps can be added up directly, the compensator is a finite drift that may be absorbed into b, and the process is a difference of two increasing pure-jump processes.

- *Infinite variation*, when that integral diverges while $\int_{|x|<1} x^2\nu(dx)$ does not. The paths wander an infinite distance in any interval, yet the quadratic variation is finite,

$$[X]_t = \sum_{s\le t} (\Delta X_s)^2, \qquad \mathbb{E}[X]_t = t\int_{|x|<1} x^2\,\nu(dx) < \infty,$$

  the squares summable exactly where the absolute values are not.

The infinite-variation case reproduces Brownian motion's signature — infinite length, finite quadratic variation — with no continuous part at all, and chapter 1 built the whole of Itô calculus on that second property. So "rough enough to need Itô calculus" does not mean continuous. The α-stable processes make the boundary explicit: with $\nu(dx) \propto |x|^{-1-\alpha}dx$ the small-jump integral $\int_0^1 x^{1-\alpha}dx$ converges exactly when $\alpha < 1$, so those are the finite variation cases and $\alpha \in [1, 2)$ are not. As $\alpha \to 2$ the jump activity concentrates ever more tightly near zero and the process approaches Brownian motion, the $\alpha = 2$ member of the family and the only one with no jumps at all.

Two consequences follow, and the second is the one that costs money. The Brownian part and the jump part do not interact: a purely discontinuous martingale has zero covariation with a continuous one, which is why the decomposition splits into pieces that can be treated separately rather than merely into terms that happen to add up. And a process made of small jumps can be made to look very like a diffusion — same finite quadratic variation, same infinite variation, paths that are visually indistinguishable at any resolution a data set offers. Whether a price is a diffusion or an infinite-activity jump process is therefore not settled by looking at it, and chapter 20 is where that becomes an identification problem rather than a philosophical one.

Remark (The cutoff at one is arbitrary, and b is not intrinsic)**.** One piece of bookkeeping, because it explains an otherwise puzzling feature of how Lévy processes are quoted.

Nothing distinguishes $|x| = 1$ as the place to split. Moving the cutoff moves jumps between the compensated term and the uncompensated one, and since the two are centred differently, the drift b changes to absorb the difference. So b is not a property of the process; it is a property of the process *and* the truncation chosen to state it. The diffusion coefficient σ^2 and the Lévy measure ν are intrinsic, and the triplet (b, σ^2, ν) is only meaningful once the truncation is named — which is why the literature carries a truncation function through statements that would otherwise be cleaner without it.

In the finite variation case the awkwardness disappears, since the jumps need no compensation at all and there is a genuine drift to quote.

Read against the previous section, the decomposition places the earlier result as the continuous corner of a larger classification: what dropping continuity buys is the jump terms, and nothing else.

Remark (Itô's lemma acquires a term that is not a derivative)**.** The chain rule changes in an instructive way. For a process with both parts,

$$f(X_t) = f(X_0) + \int_0^t f'(X_{s-})\,dX_s + \frac{1}{2}\int_0^t f''(X_{s-})\,d\langle X^c\rangle_s + \sum_{s\le t}\Big[f(X_s) - f(X_{s-}) - f'(X_{s-})\Delta X_s\Big],$$

where X^c is the continuous part and X_{s-} the value just before any jump.

The first three terms are the Itô formula already derived. The sum is new, and the important thing about it is that it is not a Taylor expansion. Across a jump the process does not move a little, it moves a lot, so no expansion in ΔX is legitimate; the bracket is the *exact* change $f(X_s)-f(X_{s-})$, less what the first-order term already booked at that instant, $f'(X_{s-})\,\Delta X_s$ — the first-order integral runs against dX_s, which jumps, so it over-counts there and the sum makes up the difference. The second-order term needs no such correction: it integrates $d\langle X^c\rangle_s$, the bracket of the continuous part, which does not move across a jump. For a continuous process every bracket is zero and the formula collapses to the one we have.

The split is a choice of grouping. Carrying the full quadratic variation in the second-order term instead would put a $-\frac{1}{2}f''(X_{s-})(\Delta X_s)^2$ into each bracket as well; keeping that term continuous is what leaves it as the Itô correction alone.

This is why the second-order term in Itô's lemma is a statement about continuous paths specifically. It came from $(\Delta W)^2$ surviving the limit while higher powers vanished, and that balance is a property of a process whose increments are of order $\sqrt{dt}$. A jump has no such scaling and has to be handled whole.

Remark (A third direction)**.** Jumps are one way out of the continuous world. There is another that gets asked about often, and the answer is not the obvious one.

Fractional Brownian motion B^H is the natural generalisation of Brownian motion in which increments are no longer independent: a Hurst parameter $H > 1/2$ gives a process that tends to continue in the direction it was going, and $H < 1/2$ one that tends to reverse. It is continuous. So one might expect it to slot into the classification above as a continuous alternative to W.

It does not, and the quadratic variation says why immediately.

Calculation 2.11 (The quadratic variation of fractional Brownian motion)**.** The defining property is that increments have variance

$$\mathbb{E}\big[(B_t^H - B_s^H)^2\big] = |t-s|^{2H},$$

which at $H = 1/2$ is the $|t-s|$ of ordinary Brownian motion. Chop $[0,t]$ into n equal intervals of width t/n and take expectations of the sum of squared increments:

$$\mathbb{E}\left[\sum_{i=0}^{n-1}\big(\Delta_i B^H\big)^2\right] = n\cdot\Big(\frac{t}{n}\Big)^{2H} = t^{2H}\,n^{\,1-2H}.$$

Everything now turns on the sign of the exponent $1-2H$ as $n\to\infty$:

$$t^{2H}n^{1-2H} \longrightarrow \begin{cases} 0 & \text{if } H > 1/2,\\ t & \text{if } H = 1/2,\\ \infty & \text{if } H < 1/2. \end{cases}$$

Read the middle line: the two powers of n cancel exactly, and only at $H = 1/2$. The whole of Itô calculus lives on that cancellation. A rougher process ($H < 1/2$) accumulates squared increments faster than the partition refines them away, and a smoother one ($H > 1/2$) more slowly; in neither case is there a finite non-zero limit for a second-order term to be built from.

Neither of the other cases can be a semimartingale. A semimartingale has finite quadratic variation, which rules out $H < 1/2$ outright; and a continuous process with *zero* quadratic variation has no martingale part at all, so it would have to be of finite variation, which fractional Brownian motion is not. So for every $H \neq 1/2$ it fails to be a semimartingale.

By the theorem quoted in chapter 4, that is fatal for a *price*: a non-semimartingale price admits a free lunch with vanishing risk, and for fractional Brownian motion the arbitrage can be written down explicitly. It is instructive that these arbitrages have the same character as the doubling strategy — they require trading continuously and arbitrarily fast, and they disappear once a minimum delay between trades, or any transaction cost, is imposed.

None of which rules out roughness in finance, because prices are not the only processes in a model. The volatility σ_t is not a traded asset, and nothing in this chapter's representation theorem asked it to be a semimartingale — only to be adapted, and to make $\int \sigma^2 ds$ finite. So one may take σ to be as rough as the data demands while the price it drives remains an ordinary continuous semimartingale, and every result here still applies. That is exactly what rough volatility models do, and chapter 10 takes up why anyone would want to.

2.3 Two Kinds of Jump

Before leaving the subject, two things the word "jump" runs together need separating: they are different mathematical objects, modelled by different means, and confusing them is a live source of error.

A jump happens at a random time τ, and stopping times come in two kinds.

Definition 2.12 (Predictable and totally inaccessible times)**.** A stopping time τ is *predictable* if there is a sequence $\tau_n \uparrow \tau$ with $\tau_n < \tau$ — an announcing sequence, so the time can be seen coming. It is *totally inaccessible* if it coincides with no predictable time: $\mathbb{P}(\tau = \sigma) = 0$ for every predictable σ.

The distinction is exactly the one between a scheduled event and a surprise, and both occur in markets.

A rate moves when the central bank meets, and the meeting dates are published a year ahead. A stock gaps on its earnings date. These are jumps at times that are not merely predictable but *deterministic*: what is random is the size, not the timing.

A company defaults. Nobody knows when, and no announcing sequence exists — the whole character of the event is that it is not visible in advance. This is a totally inaccessible time.

Remark (Only one of the two has an intensity)**.** The consequence that matters is what the compensator looks like. For a totally inaccessible time the compensator is continuous, and can be written as $\int_0^t \lambda_s \, ds$ for an *intensity* λ — a probability of the event per unit time, exactly as for

the Poisson process above. For a jump at a scheduled date the compensator has an atom at that date, and there is no intensity: a rate of occurrence per unit time is meaningless for an event that happens at a known moment.

So the two are modelled by entirely different means. A default is given an intensity. A meeting date is given a distribution for the size of the move on that date, and no intensity at all. Practitioners do the second by putting a discrete lump of variance on the event date rather than spreading variance evenly through calendar time, which is why an implied volatility term structure has visible kinks around known event dates and why interpolating a surface straight through them is wrong.

The inaccessible case deserves more attention: it is far more tractable than the general theory of jumps would suggest, and it is how the entire credit market is modelled.

Definition 2.13 (Doubly stochastic, or Cox, default)**.** Let $\lambda_t \geq 0$ be an adapted intensity. Conditional on the whole path of λ, let N be an inhomogeneous Poisson process with that intensity, and let the default time τ be its first jump.

Calculation 2.14 (Survival is a discount factor)**.** Conditional on the path of λ, the first jump of an inhomogeneous Poisson process arrives after T with probability $\exp(-\int_t^T \lambda_s ds)$. Averaging over the paths,

$$\mathbb{P}\big(\tau > T \,\big|\, \mathcal{F}_t,\ \tau > t\big) = \mathbb{E}\Big[e^{-\int_t^T \lambda_s\, ds} \,\Big|\, \mathcal{F}_t\Big]. \tag{2.3}$$

Compare that with the bond price of chapter 7, $P(t,T) = \mathbb{E}[e^{-\int_t^T r_s ds} \mid \mathcal{F}_t]$. They are the same formula. The survival probability is a discount factor in which the intensity plays the part of the short rate, and everything that was ever proved about one transfers immediately to the other:

- a defaultable zero coupon bond paying nothing on default is worth $\mathbb{E}[e^{-\int_t^T (r_s+\lambda_s)ds}]$, so default risk simply adds the intensity to the discount rate;
- a credit curve is bootstrapped from credit default swaps exactly as chapter 7 bootstraps a rate curve from par swaps;
- and if λ is given affine dynamics, chapter 12's Riccati equations deliver survival probabilities in closed form, because (2.3) is the object those equations were derived for.

So the most important jump in finance is modelled by machinery these notes develop anyway, wearing different labels — which rules out concluding that continuous models leave credit untouched.

Remark (Incompleteness is not the end of the argument)**.** Chapter 4 will observe that jumps make a market incomplete, because a move of random size at a random time cannot be hedged by a position taken beforehand. What do markets do about it: they trade the jump.

A credit default swap pays precisely on the default event, so holding one spans exactly the risk that the underlying bond could not hedge. Adding it to the traded set restores completeness for that risk, and the price stops being a range. The same pattern recurs — variance swaps span variance risk, and an option written on an event date spans the event.

The honest statement is therefore not that incompleteness makes jump models unusable, but that it tells you which instrument you are missing. An incomplete model is a shopping list.

These notes stay with continuous models from here on, and chapter 4 gives the reason: continuity is what lets the underlying and the money market replicate everything, which is what makes the price unique rather than a range, and that uniqueness is what the rest of the theory is built on.

Remark (So should a trade be modelled with jumps?)**.** The question the previous paragraph invites, and it does not have a single answer. It has a criterion: does the payoff care whether the path is continuous?

For a great many trades it does not. A payoff that depends on variance accumulated over months — a vanilla option held to expiry, a variance swap — is insensitive to whether that variance arrived smoothly or in a few lumps, because a sum of many small contributions and a sum containing a few large ones look alike once they are added up. There a diffusion calibrated to the right prices will also give the right hedge, and the jumps are absorbed into the volatility with no harm done.

For others it is the whole trade:

- anything with a *barrier*, since a diffusion cannot cross a level without touching it and a jump can gap straight over — the two models disagree about the one event the payoff is defined by;
- short-dated options far from the money, whose entire value is the probability of a large move over a short horizon, which is precisely what a diffusion says is negligible and a jump model says is not;
- anything on an asset that can default, where the jump is the risk;
- and the *hedge* of any of the above, since a delta computed from a model that does not believe in gaps cannot protect against one.

So the criterion is not the asset, it is the payoff. The same stock may be modelled as a diffusion for a one year at-the-money option and must not be for a one week barrier struck nearby.

There is a cost on the other side, and it is the running theme of these notes rather than a new objection. A jump model has more parameters — an intensity, a distribution of jump sizes — and those parameters are poorly identified by liquid vanilla quotes, in exactly the way chapter 10's β and chapter 11's mixing weight are. Fitting the surface does not pin them down, so adding jumps removes one model risk and introduces another, to be marked by judgement and reserved against as chapter 24 describes.

Which is why common practice is neither extreme. Most things are priced with a continuous model, with a lump of variance added on known event dates and an explicit reserve held for gap risk on the products where continuity is doing real work — reaching for a full jump model when the payoff makes the jump unavoidable, rather than as a general upgrade.

References

- Rogers, L. C. G. (1997). Arbitrage with fractional Brownian motion. *Mathematical Finance*, 7(1), 95–105.

- Cheridito, P. (2003). Arbitrage in fractional Brownian motion models. *Finance and Stochastics*, 7(4), 533–553.

- Kunita, H., & Watanabe, S. (1967). On square integrable martingales. *Nagoya Mathematical Journal*, 30, 209–245.

- Follmer, H., & Sondermann, D. (1986). Hedging of non-redundant contingent claims. In *Contributions to Mathematical Economics*, North-Holland, 205–223.

- Schweizer, M. (2001). A guided tour through quadratic hedging approaches. In *Option Pricing, Interest Rates and Risk Management*, Cambridge University Press, 538–574.

- Dubins, L. E., & Schwarz, G. (1965). On continuous martingales. *Proceedings of the National Academy of Sciences*, 53(5), 913–916.

- Mandelbrot, B. B., & Van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10(4), 422–437.

Chapter 3

The Generator

One operator runs through most of what follows, and it repays meeting on its own before it appears in disguise four times. For a diffusion it is $\mathcal{L}f = \mu f' + \frac{1}{2}\sigma^2 f''$, and Itô's lemma says precisely that $\mathcal{L}$ is the drift of $f(X)$. Everything else in this chapter is a consequence: pricing runs $\mathcal{L}$ backwards on functions of the state, densities run its adjoint forwards, and the two are related by integration by parts and nothing more. Later chapters derive both equations separately and at length. They are the same object, and knowing that in advance turns four derivations into one.

3.1 Itô's Lemma, Read as a Definition

Chapter 2 established that continuous martingale noise has only one source: a continuous local martingale is an integral against Brownian motion, and there is nothing else on offer. What it did not establish is that a *price* may be written that way at all — that needs the price to be a semimartingale, which follows from no-arbitrage rather than from continuity, and chapter 4 is where it is obtained. Taking both, together with a drift that accumulates at a rate, a continuous arbitrage-free price is

$$dX_t = \mu(X_t, t)\,dt + \sigma(X_t, t)\,dW_t,$$

and this chapter takes that form as given. Itô's lemma then says that for smooth f,

$$df(X_t) = \Big(\mu\frac{\partial f}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2 f}{\partial x^2}\Big)dt + \sigma\frac{\partial f}{\partial x}\,dW_t. \tag{3.1}$$

The bracket has a name.

Definition 3.1 (Infinitesimal generator)**.** The generator of the diffusion is the operator

$$\mathcal{L}f = \mu(x,t)\frac{\partial f}{\partial x} + \frac{1}{2}\sigma^2(x,t)\frac{\partial^2 f}{\partial x^2}. \tag{3.2}$$

With that, (3.1) reads $df(X_t) = \mathcal{L}f\,dt + \sigma f'\,dW_t$.

Remark (The one sentence to remember)**.** $\mathcal{L}f$ is the drift of $f(X)$.

Equivalently, and this is the definition that does not presuppose Itô's lemma,

$$\mathcal{L}f(x) = \lim_{t \downarrow 0} \frac{\mathbb{E}^x[f(X_t)] - f(x)}{t}, \tag{3.3}$$

the rate at which the expectation of f starts to move from a state x. Itô's lemma is then the statement that for a diffusion this limit is (3.2), which is a computation rather than a definition, and once done it need not be done again.

Remark (The limit is a definition only where it exists)**.** Equation (3.3) is written as though it always makes sense, and it does not. The set of f for which the limit exists — uniformly, in the sense the theory requires — is called the *domain* of the generator, written $\mathcal{D}(\mathcal{L})$, and it is a genuine restriction rather than a technicality to be waved past. Two examples, failing in different ways.

Roughness. Take standard Brownian motion and $f(x) = |x|$, and evaluate (3.3) at the origin. Since $\mathbb{E}|W_t| = \sqrt{2t/\pi}$,

$$\frac{\mathbb{E}^0[f(W_t)] - f(0)}{t} = \frac{1}{t}\sqrt{\frac{2t}{\pi}} = \sqrt{\frac{2}{\pi t}} \;\longrightarrow\; \infty.$$

The limit does not exist, so $|x|$ is not in the domain, and $\mathcal{L}|x|$ is not a hard quantity to evaluate at the origin — it does not exist.[1] This is not a curiosity, because a call payoff $(S-K)^+$ has exactly this kink at the strike. What saves the pricing equation of chapter 5 is that it never asks for $\mathcal{L}$ applied to the payoff. The payoff enters as the terminal condition $V(T,\cdot) = (S-K)^+$, and the equation $\partial_t V + \mathcal{L}V = 0$ is solved backwards from it. For every $t < T$ the solution is $V(t,\cdot) = P_{T-t}(S-K)^+$, a bounded function convolved with a Gaussian density, and that is infinitely differentiable however rough the function was: the semigroup smooths instantly, and the kink exists only at the single instant T. So $\mathcal{L}$ is applied to $V(t,\cdot)$, which is in its domain, and never to the data that started it off.

The one place the kink is still visible is numerical, where the payoff is evaluated on a finite set of points rather than integrated against a density. Chapter 19's grid-versus-simulation table wobbles for exactly this reason: down the grid-error column the value jumps from row to row because the strike falls in a different place between nodes each time, with no smoothing left to hide it.

Growth. Now take $f(x) = e^{x^2}$, which is smooth everywhere, so the formal expression (3.2) is perfectly well defined at every point. What fails is the expectation:

$$\mathbb{E}^0\big[e^{W_t^2}\big] = \frac{1}{\sqrt{1-2t}} \quad \text{for } t < \tfrac{1}{2}, \qquad \text{and } +\infty \text{ beyond.}$$

So $P_t f$ does not exist for $t \geq \frac{1}{2}$, and no amount of care with the limit at zero repairs it.[2] The domain therefore controls growth as well as smoothness — but only relative to the process's tails. Linear and polynomial growth clear that bar for every diffusion here: a call payoff is integrable against a lognormal, and the unbounded $f(x) = x$ of the counterexample below is a legitimate test function. What fails is growth that outpaces the tails, as e^{x^2} outpaces a Gaussian. The ambient

[1] both the closed form and the divergence, checked.

[2] checked, with the exponents combined --- evaluating e^{x^2} and the density separately overflows one and underflows the other.

space is then a growth-weighted class matched to the process — chapter 14's $L^2(\pi)$, say — not all of C^2, and not merely the bounded functions.

The domain is also where boundary behaviour lives.

Definition 3.2 (Absorbing and reflecting barriers)**.** Let X be a diffusion on $[0,\infty)$ and $\tau = \inf\{t : X_t = 0\}$. The barrier at 0 is *absorbing* if the path stops on arrival, $X_t = 0$ for all $t \geq \tau$, and *reflecting* if the path is turned back instantly, spending zero time at the origin and immediately re-entering $(0,\infty)$.

Both occur. An equity price that is wiped out has an absorbing zero, a defaulted name does not recover, and a knocked-out barrier option is worth zero from the moment the barrier is touched. A short rate floored at zero which is meant to keep trading has a reflecting one, as does any process constrained to a range for reasons that do not stop it.

Remark (The same $\mathcal{L}$, two domains)**.** The coefficients cannot tell these apart: μ and σ are the same function of x in both cases, and the difference is entirely a condition at the boundary — which is to say, a domain.

For a reflecting barrier the condition is $f'(0) = 0$. A one-sided derivative at the origin would be multiplied by a drift the path is not allowed to follow, and requiring it to vanish is what makes the two sides of the reflection agree.

For an absorbing barrier the process at 0 does not move, so $\mathbb{E}^0[f(X_t)] = f(0)$ for every t and (3.3) gives $\mathcal{L}f(0) = 0$: the condition is on $\mathcal{L}f$, not on f. If instead the process is *killed* at the boundary — sent to a cemetery state where every payoff is worth zero, which is what a knock-out does — the natural space is functions with $f(0) = 0$.

Every statement below is a statement about $f \in \mathcal{D}(\mathcal{L})$, and we will stop saying so.

Example 3.1 (Two generators worth recognising)**.** Geometric Brownian motion, $dS = rS\,dt + \sigma S\,dW$, has

$$\mathcal{L}f = rS\frac{\partial f}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 f}{\partial S^2},$$

which is every term of the Black-Scholes equation except the $\partial f/\partial t$ and the discounting. Chapter 5 will assemble it.

Theorem 3.3 (Dynkin's formula)**.** *Let $f \in \mathcal{D}(\mathcal{L})$ and let τ be a stopping time with $\mathbb{E}^x[\tau] < \infty$. Then*

$$\mathbb{E}^x\big[f(X_\tau)\big] = f(x) + \mathbb{E}^x\left[\int_0^\tau \mathcal{L}f(X_s)\,ds\right]. \tag{3.4}$$

Proof. Integrate $df(X_t) = \mathcal{L}f\,dt + \sigma f'\,dW_t$ from 0 to τ:

$$f(X_\tau) = f(x) + \int_0^\tau \mathcal{L}f(X_s)\,ds + \int_0^\tau \sigma(X_s) f'(X_s)\,dW_s.$$

The last term has zero expectation, and that is the only step needing an argument. By chapter 1, a stochastic integral is always a local martingale, and a true martingale when its integrand is square integrable over the interval; with f in the domain, $\sigma f'$ is bounded, so

$$\mathbb{E}^x \int_0^\tau \sigma^2 (f')^2\,ds \;\leq\; \|\sigma f'\|_\infty^2\, \mathbb{E}^x[\tau] \;<\; \infty$$

by the hypothesis on τ, and the Itô isometry of chapter 1 turns that into a finite second moment. So the integral is a true martingale, stopped at τ it still has expectation zero, and (3.4) follows. □

Remark (Where the hypothesis on τ earns its keep)**.** Dropping $\mathbb{E}[\tau] < \infty$ does not merely weaken (3.4); it makes it false. Let X be standard Brownian motion, $f(x) = x$, so $\mathcal{L}f = 0$, and let τ be the first time X reaches 1. Then τ is finite almost surely, $f(X_\tau) = 1$ always, and (3.4) would read $1 = 0 + 0$.

What has gone wrong is exactly the step above: $\mathbb{E}[\tau] = \infty$ for that stopping time, the local martingale is not a martingale, and its expectation is not preserved. This is the same failure that appears in chapter 14 as a Riccati solution exploding and in chapter 4 as a doubling strategy.

Almost every expectation computed in these notes is (3.4) with a particular f chosen to make $\mathcal{L}f$ simple — often zero, in which case $f(X)$ is a martingale and the expectation is just $f(x)$.

Remark (Finding a martingale is choosing f with $\mathcal{L}f = 0$)**.** This converts a search into an equation. A function of the state is a martingale exactly when it is in the kernel of the generator, so hunting for martingales — which is what pricing does — is solving $\mathcal{L}f = 0$.

Chapter 4's risk neutral measure will be the statement that discounted prices are martingales, which is the statement that they satisfy $\mathcal{L}f = 0$ for the appropriately adjusted $\mathcal{L}$. That is the whole of the pricing equation, and it is already visible here.

3.2 The Semigroup

The semigroup is the Markov property written as operators, so the property comes first.

Definition 3.4 (Markov property)**.** A process X adapted to a filtration $(\mathcal{F}_t)$ is *Markov* if for every bounded measurable f and every $s \le t$, the conditional expectation $\mathbb{E}\big[f(X_t)\,\big|\,\mathcal{F}_s\big]$ depends on $\mathcal{F}_s$ only through X_s: the present state carries everything the past says about the future. It is *time homogeneous* if the dependence is through $t-s$ alone, $\mathbb{E}\big[f(X_t)\,\big|\,\mathcal{F}_s\big] = (P_{t-s}f)(X_s)$, with P defined below.

Remark (Markov is a statement about a choice of state)**.** Whether a process is Markov depends on what is counted as its state. A function of a Markov process need not be Markov, and a process that is not Markov can be made so by enlarging the state. An Ornstein-Uhlenbeck process X is Markov, but its running integral $A_t = \int_0^t X_s\,ds$ is not, since where A goes next depends on X as well as on A; the pair (X, A) is. Likewise Brownian motion is Markov and its running maximum alone is not, while the pair is. A value function is a function of the state, which is why a finite state is what makes pricing a low dimensional problem, and why the state of a path dependent claim has to carry a summary of the path.

Fix f and let the starting point vary. Define

$$(P_t f)(x) = \mathbb{E}^x\big[f(X_t)\big]. \tag{3.5}$$

Proposition 3.5 (The semigroup property is the Markov property)**.** *For bounded measurable f and $s, t \ge 0$, $P_{t+s} = P_s P_t$.*

Proof. By the tower property, conditioning on $\mathcal{F}_s$,

$$(P_{t+s}f)(x) = \mathbb{E}^x\big[f(X_{t+s})\big] = \mathbb{E}^x\Big[\mathbb{E}\big[f(X_{t+s})\,\big|\,\mathcal{F}_s\big]\Big].$$

By the Markov property, definition 3.4 in its time homogeneous form, the inner conditional expectation depends on $\mathcal{F}_s$ only through X_s, and equals what the same expectation would be for a process started afresh at X_s and run for time t — that is, $\mathbb{E}[f(X_{t+s}) \mid \mathcal{F}_s] = (P_t f)(X_s)$. Substituting,

$$(P_{t+s}f)(x) = \mathbb{E}^x\big[(P_t f)(X_s)\big] = \big(P_s(P_t f)\big)(x). \qquad \square$$

The middle step is the whole content. The semigroup identity is not a consequence of being Markov that one derives; it is what being Markov *says*, rewritten with the conditioning replaced by composition of operators. Everything that follows in this chapter is available for Markov processes and unavailable otherwise, and chapter 10's rough volatility models are outside it for precisely this reason.

Proposition 3.6 (The backward equation)**.** *For $f \in \mathcal{D}(\mathcal{L})$,*

$$\frac{\partial}{\partial t}P_t f = \mathcal{L}P_t f = P_t\mathcal{L}f. \tag{3.6}$$

Proof. Both equalities come from proposition 3.5 and (3.3), differing only in which side the increment is taken on. Writing the difference quotient with the extra time at the front,

$$\frac{P_{t+h}f - P_t f}{h} = \frac{P_h(P_t f) - P_t f}{h} \;\longrightarrow\; \mathcal{L}\,P_t f$$

as $h \downarrow 0$, by the definition of $\mathcal{L}$ applied to the function $P_t f$. Taking it at the back instead,

$$\frac{P_{t+h}f - P_t f}{h} = P_t\left(\frac{P_h f - f}{h}\right) \;\longrightarrow\; P_t\mathcal{L}f,$$

the limit passing through P_t because P_t is a contraction on bounded functions,

$$\|P_t g\|_\infty = \sup_x \big|\mathbb{E}^x[g(X_t)]\big| \le \sup_x \mathbb{E}^x\big[|g(X_t)|\big] \le \sup_y |g(y)| = \|g\|_\infty,$$

since an average of values of g cannot exceed the largest value g takes anywhere. So uniform convergence of the argument gives uniform convergence of the image. $\square$

The two readings of (3.6) are the two ways of computing an expectation, and the chapter's last section is about the difference between them. The first evolves the *function*; the second evolves the *starting point*.

Structure (The generator is a velocity and the semigroup is a flow)**.** Equation (3.6) is an ordinary differential equation for an operator-valued function of time, with initial condition $P_0 = I$. Its solution is written

$$P_t = e^{t\mathcal{L}},$$

and the notation is more than decoration even though it is formal.

The family $\{P_t\}_{t\geq 0}$ is a one-parameter semigroup of operators, and $\mathcal{L}$ is its derivative at the identity. That is exactly the relationship between a Lie group and its algebra: a rotation by angle θ is $e^{\theta A}$ for an angular velocity A, and knowing A — one matrix, at one instant — determines the whole family of rotations. Here knowing $\mathcal{L}$ — coefficients, at one instant — determines the whole evolution. This is why a chapter on an operator is a chapter about a process, and it is the same identification chapter 12 needs when the criterion for a finite dimensional realisation turns out to be a condition on a Lie algebra.

Where the analogy strains is that $\mathcal{L}$ is *unbounded*: it involves two derivatives, and differentiation makes functions larger without limit. So the power series for $e^{t\mathcal{L}}$ does not converge in any straightforward sense, and the identification cannot be established by writing it down.

Remark (What makes the exponential legitimate)**.** The theorem repairing that is the Hille-Yosida theorem.

The difficulty is that $e^{t\mathcal{L}}$ cannot be defined by its series. Hille-Yosida instead characterises which unbounded operators generate a semigroup, and does it through the *resolvent* $(\lambda I - \mathcal{L})^{-1}$ — an object that is bounded where $\mathcal{L}$ is not, because inverting a differential operator is integrating, and integrating smooths. The statement is that $\mathcal{L}$ generates a contraction semigroup precisely when it is densely defined, closed, and its resolvent satisfies $\|(\lambda I - \mathcal{L})^{-1}\| \leq 1/\lambda$ for every $\lambda > 0$.

Two things follow. The exponential notation is a theorem rather than an abbreviation — something has to be checked, and for the diffusions in these notes it holds. And the resolvent, which appears here as a technical device, is the Laplace transform of the semigroup,

$$(\lambda I - \mathcal{L})^{-1} f = \int_0^\infty e^{-\lambda t}\, P_t f \, dt,$$

which is to say it is the price of a perpetual cashflow discounted at rate λ. Perpetuities are the objects on which an unbounded generator becomes bounded, and that is not a coincidence: discounting is what makes an infinite horizon finite.

Remark (What this buys)**.** Two things immediately. The first equality in (3.6) is the backward equation and the second is a statement that the generator commutes with its own semigroup, which is why one may evolve either the function or the starting point and get the same answer — the subject of the next section.

The second is a way of reading the model classes to come. If $\mathcal{L}$ maps some finite-dimensional family of functions into itself, then $e^{t\mathcal{L}}$ does too, and computing $P_t f$ for f in that family reduces to a finite system of ordinary differential equations. That single observation is why affine models are solvable, and chapter 14 is devoted to it.

3.3 The Adjoint, and Why There Are Two Equations

So far $\mathcal{L}$ has acted on functions of the state. Densities are not functions of the state in the same sense — they are the things one integrates functions against — and the operator that moves them is not $\mathcal{L}$.

Definition 3.7 (Adjoint)**.** Under the pairing $\langle f, p\rangle = \int f(x)p(x)\,dx$, the adjoint $\mathcal{L}^*$ is the operator satisfying

$$\langle \mathcal{L}f, p\rangle = \langle f, \mathcal{L}^* p\rangle \tag{3.7}$$

for every f in the domain of $\mathcal{L}$ and every p in the domain of $\mathcal{L}^*$.

Calculation 3.8 (Computing it, boundary terms included)**.** Integrate by parts on an interval $[a, b]$, keeping every term. Once for the first order piece,

$$\int_a^b \mu p\,\frac{\partial f}{\partial x}\,dx = \Big[\mu p f\Big]_a^b - \int_a^b f\,\frac{\partial}{\partial x}\big[\mu p\big]\,dx,$$

and twice for the second, which leaves two boundary terms because each integration produces one:

$$\int_a^b \tfrac{1}{2}\sigma^2 p\,\frac{\partial^2 f}{\partial x^2}\,dx = \Big[\tfrac{1}{2}\sigma^2 p\,\frac{\partial f}{\partial x} - f\,\frac{\partial}{\partial x}\big[\tfrac{1}{2}\sigma^2 p\big]\Big]_a^b + \int_a^b f\,\frac{\partial^2}{\partial x^2}\Big[\tfrac{1}{2}\sigma^2 p\Big]\,dx.$$

Collecting, (3.7) holds with

$$\mathcal{L}^* p = -\frac{\partial}{\partial x}\big[\mu p\big] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\big[\sigma^2 p\big] \tag{3.8}$$

provided the boundary terms cancel:

$$\left[\underbrace{\mu p f - f\frac{\partial}{\partial x}\Big[\tfrac{1}{2}\sigma^2 p\Big]}_{-f\cdot J[p]} + \tfrac{1}{2}\sigma^2 p\,\frac{\partial f}{\partial x}\right]_a^b = 0. \tag{3.9}$$

The bracket in (3.9) is not clutter, and the quantity named under it is the reason.

Definition 3.9 (Probability flux)**.** The flux of a density p is

$$J[p] = \mu p - \frac{\partial}{\partial x}\Big[\tfrac{1}{2}\sigma^2 p\Big], \tag{3.10}$$

so that $\mathcal{L}^* p = -\partial_x J[p]$.

Remark (What the boundary terms are for)**.** Writing $\mathcal{L}^*$ as the derivative of a flux makes the forward equation a conservation law: $\partial_t p = -\partial_x J$ says probability is neither created nor destroyed, only transported, and J is the rate at which it moves past a point. Chapter 9 makes use of exactly this form.

The boundary terms then have a reading rather than being a nuisance to assume away. On the whole line with p decaying and f bounded, both terms vanish and (3.8) holds as stated. On a bounded interval they do not vanish by themselves, and requiring them to cancel is precisely a boundary condition — $J = 0$ at the ends for a reflecting barrier, $p = 0$ for an absorbing one. So the pair $(\mathcal{L}, \mathcal{L}^*)$ carries the boundary behaviour in the condition that makes them adjoint, which is the concrete form of the remark above that a domain, not a formula, distinguishes two processes with the same coefficients.

The flux says what the two barriers of the previous section are doing, in one quantity. Integrating the forward equation over $[0,\infty)$,

$$\frac{d}{dt}\int_0^\infty p(t,x)\,dx = -\big[J\big]_0^\infty = J(t,0)$$

when J vanishes at infinity. A reflecting barrier has $J(t,0)=0$ and the total is conserved: nothing crosses. An absorbing barrier has $J(t,0)<0$, probability leaving through the origin and not returning, so p is the density of the paths that have survived and integrates to less than one. And since the mass lost by time t is exactly $\mathbb{P}(\tau\le t)$, the outward flux at the barrier *is* the density of the first passage time — which is how a structural model of default prices, though chapter 16 builds its default density from an intensity instead.

Chapter 14 needs the zero-flux case specifically. Setting $J[p]=0$ pointwise and solving is what produces the stationary density there, and the reason it can be done in one dimension is visible in (3.10): a flux with zero divergence on a line must be constant, and a constant flux that vanishes at either end vanishes everywhere. In two dimensions a divergence-free flux may circulate instead, which is the whole of that chapter's obstruction to reversibility.

Structure (It is the same flux as in physics)**.** The word is not an analogy. $\partial_t p+\partial_x J=0$ is the continuity equation, the same statement that carries charge in electromagnetism and mass in fluid dynamics, and it is derived the same way: whatever leaves a region has to cross its boundary, which in higher dimensions is the divergence theorem and in one dimension is the $[\,\cdot\,]_a^b$ that integration by parts produced above. Quantum mechanics calls the corresponding object for $|\psi|^2$ the probability current, and means by it exactly what (3.10) means.

The two terms are the two classical fluxes. The drift term μp is convective transport, density carried along by a velocity field, the ρv of a fluid. The second term is diffusive transport down a gradient, which is Fick's law for a solute and Fourier's law for heat, with $\frac12\sigma^2$ in the role of the diffusion constant. A Fokker-Planck equation is an advection-diffusion equation, and a Feynman-Kac problem with a discount rate is one with a sink term.

The boundaries transfer too. A reflecting barrier is an insulated wall, no heat crossing it, a Neumann condition. An absorbing barrier is a perfect sink held at zero concentration, a Dirichlet condition, and the flux at it is the rate of arrival — the same computation that gives a first passage density gives the heat flowing into a cold wall.

Two places in the dictionary need care. The first is that $\frac12\sigma^2$ sits *inside* the derivative, where the physical Fick's law puts the diffusion constant outside. For constant σ there is no difference; for state-dependent σ there is. The next remark is about that difference from the operator side.

The second is the two-dimensional obstruction just mentioned. In stationarity $\partial_t p=0$, so the flux is divergence-free — and a divergence-free field need not vanish. It may circulate, exactly as a magnetostatic field does around a current. Reversibility is the condition that the stationary flux is not merely divergence-free but zero; what stands between the two is the circulating part that the Helmholtz decomposition isolates, and on a line there is nowhere for it to live. Above one dimension there is, and an irreversible stationary diffusion is one with a vortex in it.

Remark (Inside or outside is the whole difference)**.** Put (3.2) and (3.8) side by side. The coefficients are the same, the derivatives are the same, and the only difference is that in $\mathcal{L}$ they sit outside the

derivatives and in $\mathcal{L}^*$ they sit inside. Integration by parts is the operation that moves them, and that is the entire content of the relationship.

The difference is not cosmetic, because μ and σ depend on x: $\partial_x[\mu p] = \mu\,\partial_x p + p\,\partial_x \mu$, and the second term is real. Getting it wrong gives an equation that does not conserve probability, and `a test` makes the failure explicit rather than leaving it as a warning.

That is all the adjoint is. And it immediately gives the equation the density satisfies, at no further cost.

Calculation 3.10 (The forward equation, in two lines)**.** Let p_t be the density of X_t. Differentiating the pairing two ways and using (3.7),

$$\Big\langle f, \frac{\partial p_t}{\partial t} \Big\rangle = \frac{d}{dt}\langle f, p_t\rangle = \frac{d}{dt}\,\mathbb{E}[f(X_t)] = \langle \mathcal{L}f, p_t\rangle = \langle f, \mathcal{L}^* p_t\rangle,$$

the third equality being the definition of the generator. This holds for every test function f, so the two right hand arguments agree:

$$\frac{\partial p}{\partial t} = \mathcal{L}^* p. \tag{3.11}$$

Under its own name this is the Fokker-Planck, or forward Kolmogorov, equation. Chapter 9 puts it to work.

The two equations that occupy later chapters are now both visible, and say the same thing:

	Acts on	**Equation**
Backward, $\mathcal{L}$	functions of the state	$\partial u/\partial t + \mathcal{L}u = 0$
Forward, $\mathcal{L}^*$	densities	$\partial p/\partial t = \mathcal{L}^* p$

The first is Feynman-Kac, derived in chapter 5 and used to price. The second is (3.11), which chapter 9 uses to obtain Dupire's formula. Neither derivation needs to be read as new once this table is understood; each is (3.7) with the specific coefficients substituted.

Remark (The duality, in words that a desk uses)**.** An expectation $\mathbb{E}[f(X_T)]$ can be computed in two ways. Evolve the payoff backwards to today and pair it with today's density, or evolve the density forwards to expiry and pair it with the payoff. By (3.7) these agree exactly, and the choice is one of convenience:

$$\langle f, p_T\rangle = \langle P_T f, p_0\rangle. \tag{3.12}$$

Which is more convenient depends on what is being varied. One payoff and many starting points — an option, revalued as the market moves — wants the backward equation. Many payoffs and one starting point — a whole strike surface, today — wants the forward one, since a single forward solve prices every strike at once. That is why chapter 9 reaches for the forward equation and chapter 5 for the backward.

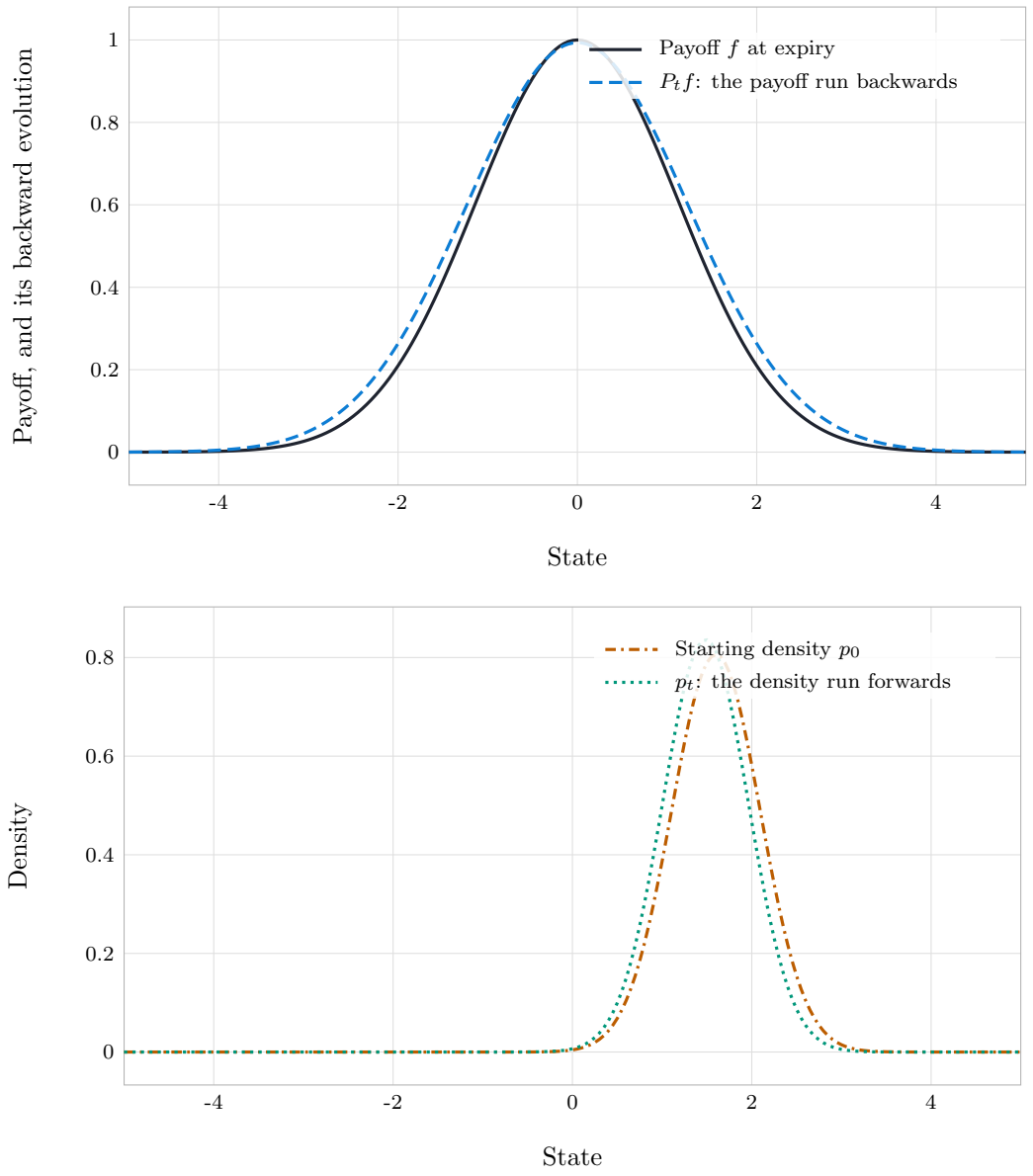

Figure 3.1: The same expectation computed from both ends, for a mean reverting process. The upper panel shows a payoff and what it becomes when run backwards under $\mathcal{L}$; the lower shows a starting density and what it becomes when run forwards under $\mathcal{L}^*$. Pairing the original payoff with the evolved density gives 0.4437028414; pairing the evolved payoff with the original density gives 0.4437028414. The two routes have no numerical step in common beyond the grid, and (3.12) says they must agree.

Drawn by `Diffusion::generator` and `Diffusion::adjoint`.

3.4 What to Carry Forward

Three things, and they are the reason this chapter comes before the pricing rather than after it.

First, $\mathcal{L}f$ is the drift of $f(X)$, so a martingale is a solution of $\mathcal{L}f = 0$. Every pricing equation in these notes is that sentence with discounting attached.

Second, there are not two theories, one for prices and one for densities. There is one operator and its adjoint, and integration by parts is the only thing separating them. When chapter 9 derives the Fokker-Planck equation from scratch, notice that the derivation is three lines of integration by parts and nothing else.

Third, and least obvious now, tractability is a property of $\mathcal{L}$. A model is solvable when its generator preserves a small family of functions, so that an infinite-dimensional evolution collapses onto a finite one. Affine models, quadratic models and polynomial models are three answers to the same question about $\mathcal{L}$, and chapter 14 asks it properly.

References

- Oksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications*, 6th ed. Springer.

- Ethier, S. N., & Kurtz, T. G. (1986). *Markov Processes: Characterization and Convergence*. Wiley.

- Pavliotis, G. A. (2014). *Stochastic Processes and Applications*. Springer.

Chapter 4

No-Arbitrage Pricing

We try to explain what arbitrage is and how to price financial derivatives in order to avoid arbitrage. The argument runs from a two-state model to the fundamental theorem in general form, and then asks what the theorem needs: completeness, which is what makes a price a number rather than a range. Jumps break it, so we work out what a jump model hands over instead — a split of the risk into the part that can be hedged continuously and the part that cannot — and close by deriving, rather than assuming, why every model in these notes is a diffusion.

4.1 Binomial Model

Arbitrage and risk-neutral pricing are easiest to see away from the continuous-time model, in the simplest possible one: a single time step in which, given today's value, the process takes one of two possible values at the end, each with some probability.

Definition 4.1 (Forward Contract)**.** A forward contract is an agreement between two parties to buy or sell a financial tradable at a decided date in the future at a price determined at the inception of the contract. The set price is known as the forward price, and the decided date in future is known as the expiration date. Usually there is no initial exchange of cashflows before the expiration.

A stock trades at \$100. Historical probabilities put it at \$120 in a year with probability 0.5 and at \$90 with probability 0.5, and the bank borrows and lends at 2%. A forward contract on the stock, settling in a year, is also quoted. With the historic probabilities in hand, one might be tempted to price the forward at

$$\mathbb{E}^{\mathbb{P}^{\text{historic}}}[S_{1\text{year}}] = 0.5 \cdot \$120 + 0.5 \cdot \$90 = \$105$$

Suppose the bank is actually quoting the forward at \$105. That price admits a riskless profit.

Sell the forward for \$105, borrow \$100 from the bank, and use the loan to buy the stock now. A year later, deliver the stock against the forward for \$105 and repay the bank \$102. The strategy starts with no money and ends with a profit of \$3 with certainty: arbitrage.

Definition 4.2 (Arbitrage)**.** A portfolio is an arbitrage if we have

$$\pi_0 = 0 \qquad \mathbb{P}^{\text{historic}}(\pi_T \geq 0) = 1 \qquad \mathbb{P}^{\text{historic}}(\pi_T > 0) > 0 \tag{4.1}$$

where π_t is the value of the portfolio at time t. The portfolio must be self-financing: no cash is added during the strategy.

The two probability conditions say the strategy loses money almost surely never, and makes money with positive probability.

Definition 4.3 (Option contract)**.** An option contract is an agreement between two parties where the buyer of the contract has the option, but not the obligation, to buy or sell a financial tradable at a decided date in the future at a price determined at the inception of the contract. An option contract with the option to buy (sell) is called a call (put) option, the set price is known as the strike, and the decided date in future is known as the option expiration date.

Definition 4.4 (Bond)**.** A bond is a contract between two parties that pays \$1 to the buyer of the bond at the expiration date.

The riskless bond price is governed primarily by the interest rates.

Example 4.1 (Binomial Option Pricing)**.** The same stock and market, now pricing a call option with strike \$105 and expiration in a year rather than a forward. If the stock rises to \$120, the client exercises and buys the stock at \$105; if it falls to \$90, the client lets the option lapse and buys in the market instead. Buying the stock now to hand over later protects against the client exercising, but leaves an unhedged stock position if the price falls and the option lapses — so the stock alone does not replicate the payoff. What does is a portfolio of stock and bonds matching the option's payoff in both states exactly; bought today, it carries zero risk whichever way the stock moves. The final payoff

$$\begin{aligned}
&= \begin{cases} S_{1\text{year}} - \$105 & \text{if} \quad S_{1\text{year}} = \$120 \\ 0 & \text{if} \quad S_{1\text{year}} = \$90 \end{cases} \\
&= \begin{cases} \frac{1}{2}S_{1\text{year}} - \$45 & \text{if} \quad S_{1\text{year}} = \$120 \\ \frac{1}{2}S_{1\text{year}} - \$45 & \text{if} \quad S_{1\text{year}} = \$90 \end{cases} \\
&= \frac{1}{2}S_{1\text{year}} - \$45 \\
&= \frac{1}{2}S_{1\text{year}} - 45B_{1\text{year}}
\end{aligned}$$

The option payoff is replicated by holding 1/2 a stock and selling 45 bonds, which hedges the position completely. The price of that replicating portfolio, and hence of the call option, is

$$\frac{1}{2} \cdot S_{\text{now}} - 45 \cdot B_{\text{now}} = \frac{1}{2} \cdot \$100 - 45 \cdot \$\frac{1}{1.02} \sim \$6$$

Any other price for the option admits unlimited arbitrage profit, trading it against the replicating portfolio.

Exercise (Binomial Forward Price)**.** Show that the forward price of the forward contract in the first example should be equal to $S_{\text{now}}/B_{\text{now}}$ to not have arbitrage.

The binomial model's branching is binary, so a replicating portfolio of stock and bond always exists: solve the linear system

$$\begin{aligned} a \cdot S_u + b \cdot B &= D_u \\ a \cdot S_d + b \cdot B &= D_d \end{aligned}$$

where S_u and S_d are the two possible stock prices a year from now, and D_u and D_d are the corresponding derivative payoffs. B is a bond paying 1 a year from now. The current price of the derivative is then

$$D_{\text{now}} = a \cdot S_{\text{now}} + b \cdot B_{\text{now}} \tag{4.2}$$

Solving for a and b,

$$\begin{aligned} a &= \frac{D_u - D_d}{S_u - S_d} \\ b &= \frac{D_d \cdot S_u - D_u \cdot S_d}{S_u - S_d} \\ D_{\text{now}} &= \frac{D_u - D_d}{S_u - S_d} \cdot S_{\text{now}} + \frac{D_d \cdot S_u - D_u \cdot S_d}{S_u - S_d} \cdot B_{\text{now}} \\ &= \frac{S_{\text{now}} - S_d B_{\text{now}}}{S_u - S_d} \cdot D_u + \frac{-S_{\text{now}} + S_u B_{\text{now}}}{S_u - S_d} \cdot D_d \\ &\equiv q \cdot B_{\text{now}} D_u + (1-q) \cdot B_{\text{now}} D_d \end{aligned}$$

Definition 4.5 (Risk Neutral Probability)**.** In the representation above, q is itself a probability measure — the risk-neutral probability — and the pricing formula is an expectation under it.

The historic probabilities do not appear in the pricing formula at all. One more identity is worth recording before returning to continuous time, since it translates directly.

Remark (Return on Stock in Risk Neutral Measure)**.**

$$\mathbb{E}^{\mathbb{P}-\text{risk-neutral}}[S_{1\text{year}}] = q \cdot S_u + (1-q) \cdot S_d = S_{\text{now}}/B_{\text{now}} \tag{4.3}$$

The expected return on the stock under the risk-neutral measure equals the return on the riskless bond.

4.2 Continuous Time Itô Diffusion Model

Return now to continuous time.

Consider a portfolio of stock and a money market account. Its value at any time is

$$\begin{aligned} \pi_t &= \Delta_t S_t + M_t \\ \pi_0 &= 0 \end{aligned}$$

where Δ_t is the number of shares of price S_t held at time t, and M_t is the amount held in the money market account, earning the instantaneous riskless rate r.

$$d_{\text{market}}\pi_t = \Delta_t dS_t + rM_t dt$$

the change in portfolio value from market movement alone. Self-financing means no cash is added or withdrawn at any rebalancing, so

$$\begin{aligned} d\pi_t &= d_{\text{market}}\pi_t \\ &= \Delta_t dS_t + rM_t dt \\ (d\Delta_t)S_t + \Delta_t dS_t + (d\Delta_t)(dS_t) + dM_t &= \Delta_t dS_t + rM_t dt \qquad \text{(using Itô's Lemma)} \\ dM_t &= rM_t dt - (d\Delta_t)(S_t + dS_t) \end{aligned}$$

Take the ansatz $M_t = e^{\int_0^t r ds}\tilde{M}_t$ and $S_t = e^{\int_0^t r ds}\tilde{S}_t$. We then have

$$\begin{aligned} e^{\int_0^t r ds}d\tilde{M}_t + re^{\int_0^t r ds}\tilde{M}_t dt &= re^{\int_0^t r ds}\tilde{M}_t dt \\ &\quad - (d\Delta_t)\, e^{\int_0^t r ds}\big(\tilde{S}_t + d\tilde{S}_t + r\tilde{S}_t dt\big) \\ d\tilde{M}_t &= -(d\Delta_t)((1+rdt)\tilde{S}_t + d\tilde{S}_t) \\ &= -(d\Delta_t)(\tilde{S}_t + d\tilde{S}_t) \qquad (\because (d\Delta_t)(dt) = 0) \end{aligned}$$

Therefore, with the ansatz $\pi_t = e^{\int_0^t r ds}\tilde{\pi}_t$, we have

$$\begin{aligned} \tilde{\pi}_t &= \Delta_t\tilde{S}_t + \tilde{M}_t \\ d\tilde{\pi}_t &= d(\Delta_t\tilde{S}_t) + d\tilde{M}_t \\ &= d\Delta_t d\tilde{S}_t + (d\Delta_t)\tilde{S}_t + \Delta_t d\tilde{S}_t + d\tilde{M}_t \\ &= \Delta_t d\tilde{S}_t \end{aligned}$$

Denoting $e^{\int_0^t r ds}$ by A_t, also known as the accumulation factor, we have

$$\frac{\pi_T}{A_T} = \int_0^T \Delta_t d\frac{S_t}{A_t}$$

The argument holds for the price of any tradable asset, including a derivative, in place of S_t. Write $D_t = 1/A_t$ for the discount factor and $D_t S_t = S_t/A_t$ for the discounted asset price.

Definition 4.6 (Equivalent Measures)**.** Two probability measures $\mathbb{P}^a$ and $\mathbb{P}^b$ are equivalent if for any event A of the σ-algebra $\mathbb{P}^a(A) = 0$ if and only if $\mathbb{P}^b(A) = 0$.

Definition 4.7 (Admissible strategy)**.** A self-financing strategy is admissible if there is a constant $a \geq 0$ such that its discounted wealth satisfies $\pi_t/A_t \geq -a$ for all t, almost surely. The restriction is not a technicality: §4.5 exhibits a strategy whose wealth is unbounded below along the way, and would otherwise have to be called risk-free.

Lemma 4.8 (Admissibility buys a supermartingale)**.** *Let S/A be a continuous local martingale and Δ a strategy whose discounted wealth $\pi_t/A_t = \int_0^t \Delta\, d(S/A)$ satisfies $\pi_t/A_t \geq -a$ for a constant a. Then π/A is a supermartingale, and in particular $\mathbb{E}[\pi_T/A_T] \leq 0$.*

Proof. A stochastic integral against a continuous local martingale is a continuous local martingale — chapter 1's localisation, with the bracket $\langle S/A \rangle$ in place of ds — so $\pi/A + a$ is one too, and it is non-negative. A non-negative local martingale is a supermartingale — the lemma of chapter 1, which is conditional Fatou and nothing more — so $\pi/A + a$ is a supermartingale, and subtracting the constant leaves one. Its expectation is therefore at most its initial value, which is zero. □

Lemma 4.9 (Sufficient condition for no-arbitrage)**.** *If there is a measure $\mathbb{P}$ equivalent to the real world historic measure $\mathbb{P}^{historic}$ such that the discounted tradable asset process S_t/A_t is a local martingale under $\mathbb{P}$, then the market $\{S_t, M_t\}$ has no arbitrage.*

Proof. We show the two conditions defining an arbitrage,

$$\mathbb{P}^{\text{historic}}(\pi_T \geq 0) = 1 \qquad \mathbb{P}^{\text{historic}}(\pi_T > 0) > 0,$$

cannot hold together. The first step is the only place equivalence of the measures is used, and it is used twice over: equivalent measures agree on which events have probability zero, so both conditions may be rewritten under $\mathbb{P}$,

$$\mathbb{P}(\pi_T \geq 0) = 1 \qquad \mathbb{P}(\pi_T > 0) > 0.$$

This is why the theorem asks for an equivalent measure rather than merely some measure. A measure that assigned zero probability to a state could price away an arbitrage that occurs only in that state.

Now the discounted wealth is $\pi_T/A_T = \int_0^T \Delta_t \, d(S_t/A_t)$, a stochastic integral against a local martingale, and if the strategy is admissible then by Lemma 4.8 it is a supermartingale, so

$$\mathbb{E}^{\mathbb{P}}[\pi_T/A_T] \leq \pi_0/A_0 = 0.$$

An inequality is all that is available — a local martingale need not have constant expectation, and the doubling strategy later in this chapter is the case where it does not — and an inequality is all that is needed. A non-negative variable with non-positive expectation is zero almost surely, and $A_T > 0$, so $\mathbb{P}(\pi_T = 0) = 1$. That contradicts the second condition. □

Two things the proof did not need should be named, because their absence is what makes the direction easy. It did not need the market to be complete, and it did not need the measure to be unique: any one equivalent local martingale measure rules out arbitrage. Completeness is what decides whether the measure is unique, and §4.6 is about the case where it is not.

Together with its converse this gives the Fundamental Theorem of Asset Pricing.

Theorem 4.10 (First Fundamental Theorem of Asset Pricing)**.** *The market $\{S_t, M_t\}$ has no arbitrage if and only if there is a measure $\mathbb{P}$ equivalent to the real world historic measure $\mathbb{P}^{historic}$ such that the discounted tradable asset process S_t/A_t is a local martingale under $\mathbb{P}$.*

Where the two directions are proved. The statement as written is true in discrete settings and *false* in continuous time without repair, so it pays to be exact about what is proved where before proceeding.

The direction just established — measure implies no arbitrage — holds as stated, in every setting, for admissible strategies. It is the half that gets used, since a modelling exercise always starts by writing the measure down.

The converse is proved completely below in two finite settings, one period in Theorem 4.12 and a finite tree in Theorem 4.14, and the geometry there is the whole idea. In continuous time it fails: §4.5 exhibits a market with no arbitrage in the sense defined above and no equivalent martingale measure. The repair is to strengthen the hypothesis from no arbitrage to no free lunch with vanishing risk, and the resulting statement is the Delbaen-Schachermayer theorem of that section. So the honest reading of this theorem is that it is exactly right in finite dimensions and is the finite-dimensional shadow of a harder statement in infinite ones. □

4.3 From No Arbitrage to a Martingale Measure

What we proved above is the half of the theorem we actually use: hand me a martingale measure and I will tell you the market is safe. The other half says that the measure is not a lucky accident of the models we happen to write down — if a market admits no arbitrage at all, such a measure must exist. Without it, "find an equivalent martingale measure" would be a modelling convention. With it, it is the only thing one can do.

This direction is genuinely harder. So we will first prove it completely in a simpler setting — one period, finitely many states — then extend it to many periods, and only then look at what breaks in continuous time. The finite case is not a toy. It contains the whole idea, and the idea is geometric: the strategies available to you sweep out a flat subspace of payoffs, arbitrage means that subspace touches the positive orthant, and a subspace that misses the positive orthant can be separated from it by a plane. The coefficients of that plane, normalised, are the risk neutral probabilities. That is the entire proof.

Now the market. Take one period, from time 0 to time 1, and a sample space $\Omega = \{\omega_1, \dots, \omega_n\}$ with finitely many states, each of which the historic measure gives positive probability — a state of probability zero can simply be deleted from the list. There are d risky tradables and a money market account. As before we work with discounted prices, writing $\tilde{S}_0 \in \mathbb{R}^d$ for today's discounted prices and $\tilde{S}_1(\omega) \in \mathbb{R}^d$ for tomorrow's; discounting makes the money market account worth 1 in every state, at both dates.

A portfolio is a vector $\theta \in \mathbb{R}^d$ of holdings in the risky assets, with whatever cash is needed to fund it borrowed or lent in the money market. Its discounted value at time 1, having started from $\pi_0 = 0$, is

$$\tilde{\pi}_1 = \theta \cdot (\tilde{S}_1 - \tilde{S}_0),$$

which is the one-period version of the identity $\pi_T/A_T = \int_0^T \Delta_t \, d(S_t/A_t)$ we derived for the continuous time model: a self-financing portfolio starting from nothing ends up holding exactly its gains from trading, measured in the numeraire.

Read a payoff as a vector in $\mathbb{R}^n$ by listing its value in each state, and let

$$K = \left\{ \big(\theta \cdot (\tilde{S}_1(\omega_1) - \tilde{S}_0), \dots, \theta \cdot (\tilde{S}_1(\omega_n) - \tilde{S}_0)\big) : \theta \in \mathbb{R}^d \right\} \subseteq \mathbb{R}^n$$

be the set of discounted payoffs reachable from zero capital. It is the image of $\mathbb{R}^d$ under a linear map, so a linear subspace of $\mathbb{R}^n$ — the only thing the argument needs about it, and what makes the geometry simple.

An arbitrage is precisely an element of K that is non-negative in every coordinate and non-zero: a portfolio costing nothing to set up that cannot lose and might gain. If $x \in K$ is such an element then $\sum_i x_i > 0$, and since K is a subspace it also contains $x/\sum_i x_i$, which lies in the unit simplex

$$C = \Big\{ x \in \mathbb{R}^n : x_i \geq 0 \text{ for all } i, \sum_{i=1}^{n} x_i = 1 \Big\}.$$

So an arbitrage exists if and only if $K \cap C \neq \emptyset$, and *no* arbitrage means the subspace K misses the simplex. What remains is then a statement about a subspace and the simplex alone, with the finance stripped out.

Lemma 4.11 (Separating a subspace from the simplex)**.** *Let $K \subseteq \mathbb{R}^n$ be a linear subspace and C the unit simplex above. If $K \cap C = \emptyset$, then there exists $\lambda \in \mathbb{R}^n$ with $\lambda_i > 0$ for every i, such that $\lambda \cdot k = 0$ for every $k \in K$.*

Proof. Consider the set of differences

$$D = C - K = \{c - k : c \in C, k \in K\}.$$

D is convex, being a sum of convex sets. It is also closed: if $c_m - k_m \to z$ with $c_m \in C$ and $k_m \in K$, then since C is compact we may pass to a subsequence along which $c_m \to c \in C$, whence $k_m \to c - z$, and K is closed (every linear subspace of $\mathbb{R}^n$ is), so $c - z \in K$ and $z = c - (c - z) \in D$. Finally $0 \notin D$, since $0 = c - k$ would mean $c = k \in K \cap C$.

Let λ be the point of D closest to the origin; it exists because D is closed and non-empty, and $\lambda \neq 0$ because $0 \notin D$. We claim $\lambda \cdot z \geq |\lambda|^2$ for every $z \in D$. Indeed, by convexity $\lambda + t(z - \lambda) \in D$ for $t \in [0, 1]$, so by the choice of λ

$$\begin{aligned} |\lambda + t(z - \lambda)|^2 &\geq |\lambda|^2 \\ 2t\lambda \cdot (z - \lambda) + t^2 |z - \lambda|^2 &\geq 0 \\ 2\lambda \cdot (z - \lambda) + t|z - \lambda|^2 &\geq 0 \end{aligned}$$

for every $t \in (0, 1]$, and letting $t \downarrow 0$ gives $\lambda \cdot z \geq \lambda \cdot \lambda = |\lambda|^2$.

Written out, this says

$$\lambda \cdot c - \lambda \cdot k \geq |\lambda|^2 > 0 \qquad \text{for all } c \in C,\ k \in K.$$

Now fix any $k \in K$. Since K is a subspace, $sk \in K$ for every real s, so $\lambda \cdot c - s(\lambda \cdot k) \geq |\lambda|^2$ for every s. A linear function of s that is bounded below must be constant, so $\lambda \cdot k = 0$. This holds for every $k \in K$, which is the second claim.

We are left with $\lambda \cdot c \geq |\lambda|^2 > 0$ for every $c \in C$. Taking c to be the i-th vertex of the simplex, the vector with a 1 in position i and zeros elsewhere, gives $\lambda_i > 0$. □

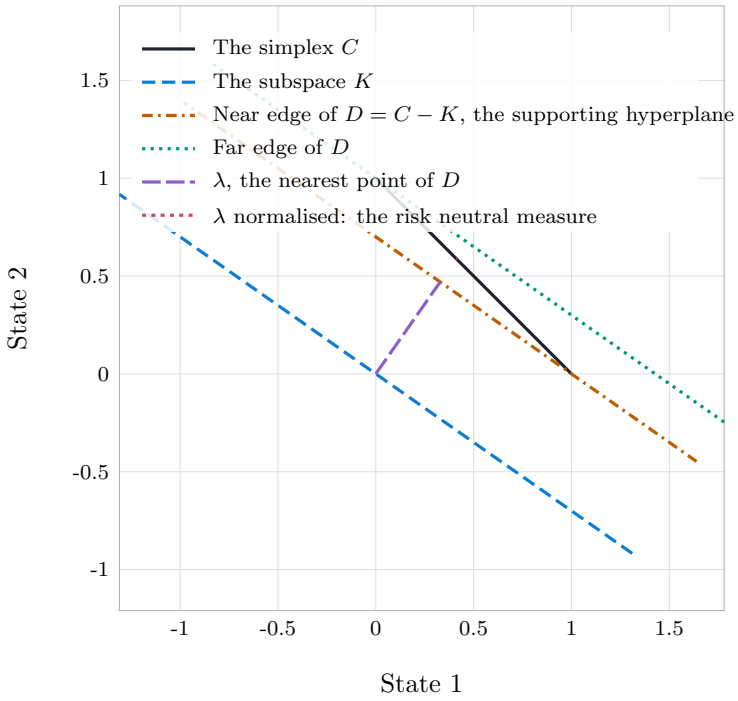

Figure 4.1: The lemma's proof in two dimensions, where K is a line and C is the segment joining the two axis vectors. The difference set $D = C - K$ is that segment swept along the line, which makes it an infinite strip parallel to K, and λ is its nearest point to the origin. Three of the proof's steps become things to look at rather than verify. λ is perpendicular to K because the nearest point of a strip is reached by walking straight at it. The supporting hyperplane is not a new object at all — it is the near edge of the strip, which is why the proof never has to build one. And λ falls in the positive quadrant, which is what makes it a measure rather than merely a separating direction. Rotate K into the first quadrant and the hypothesis fails in front of you: the line cuts the simplex, the strip swallows the origin, and the nearest point of D to the origin becomes the origin itself, leaving nothing to normalise.

Drawn by `separation`.

Remark (Why the proof is arranged this way)**.** Forming $D = C - K$ looks like a trick until one asks what the alternative would be. Separating two sets requires the pair to be disjoint and at least one of them to be compact — K is a subspace and is neither bounded nor, on its own, easy to argue about. Subtracting collapses the two-set problem to a one-set problem: the single question "is the origin in D?", which is the same question as $K \cap C = \emptyset$ and much easier to attack, since the nearest point of a closed convex set to a point outside it always exists and is unique.

That is the entire idea. Everything after it in the proof is checking that D really is closed and convex, and then reading off consequences of λ being nearest.

Theorem 4.12 (First Fundamental Theorem, one period and finitely many states)**.** *The market $\{\tilde{S}_0, \tilde{S}_1\}$ admits no arbitrage if and only if there is a probability measure $\mathbb{Q}$ on Ω giving every state positive probability, such that*

$$\mathbb{E}^{\mathbb{Q}}[\tilde{S}_1] = \tilde{S}_0.$$

Proof. One direction is the lemma we already proved, and in this setting it is a single line: if such a $\mathbb{Q}$ exists then $\mathbb{E}^{\mathbb{Q}}[\tilde{\pi}_1] = \theta \cdot (\mathbb{E}^{\mathbb{Q}}[\tilde{S}_1] - \tilde{S}_0) = 0$ for every θ, and a random variable that is non-negative everywhere with mean zero under a measure that charges every state must be zero in every state.

For the converse, no arbitrage means $K \cap C = \emptyset$, as set up above. Let λ be the vector supplied by Lemma 4.11, and set

$$q_i = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j},$$

which is well defined and strictly positive because every $\lambda_i > 0$, and sums to 1 by construction, so it defines a probability measure $\mathbb{Q}$ charging every state. That λ annihilates K says that for every $\theta \in \mathbb{R}^d$

$$\sum_{i=1}^{n} q_i \, \theta \cdot \big(\tilde{S}_1(\omega_i) - \tilde{S}_0\big) = 0$$
$$\theta \cdot \big(\mathbb{E}^{\mathbb{Q}}[\tilde{S}_1] - \tilde{S}_0\big) = 0,$$

and a vector orthogonal to every $\theta \in \mathbb{R}^d$ is the zero vector. □

With two states the whole argument fits on a page of graph paper.

Let the axes be the payoff in each of the two states. The attainable payoffs K are the multiples of a single vector, so they are a line through the origin. The simplex C is the segment joining $(1, 0)$ to $(0, 1)$. Arbitrage is a point of K with both coordinates non-negative and not both zero — so, after scaling, a point where the line crosses the segment.

The separating vector is not an abstraction. In two dimensions the orthogonal complement of a line is a line, so there is exactly one candidate for λ, and the only question the lemma settles is whether it points into the positive quadrant. It does exactly when the price sits inside the range, and normalising it to sum to one gives, in the discounted coordinates $\tilde{S}_0, \tilde{S}_u, \tilde{S}_d$ this section has used throughout,

$$q = \left(\frac{\tilde{S}_0 - \tilde{S}_d}{\tilde{S}_u - \tilde{S}_d}, \; \frac{\tilde{S}_u - \tilde{S}_0}{\tilde{S}_u - \tilde{S}_d} \right).$$

With $\tilde{S}_0 = S_{\text{now}}$ and $\tilde{S}_u = S_u B_{\text{now}}$, $\tilde{S}_d = S_d B_{\text{now}}$, this is $\big((S_{\text{now}}/B_{\text{now}} - S_d)/(S_u - S_d),\ (S_u - S_{\text{now}}/B_{\text{now}})/(S_u - S_d)\big)$, which is exactly the q the binomial calculation produced at the start of the chapter — the same object, written in undiscounted terms. The geometry and the algebra agree once the discounting is restored; dropping the tildes and reading S_{now}, S_u, S_d as the plain prices, as opposed to $\tilde{S}_0, \tilde{S}_u, \tilde{S}_d$, gives a different number.

Remark. Return to the binomial example at the start of this chapter with this in hand. There $n = 2$ and $d = 1$, K is a line in the plane, and we found q by solving two linear equations for the replicating portfolio. Nothing in that calculation explained why q came out between 0 and 1; we simply observed that it had. The separation argument is what that observation was: q is positive exactly because the line K misses the simplex, and it misses the simplex exactly because there is no arbitrage. Had we chosen a forward price outside the range the market allows, K would have cut the positive quadrant, the "probability" would have come out negative, and the strategy realising the arbitrage would have been the point where it did so.

One question is left: when is $\mathbb{Q}$ the *only* martingale measure? Call the market *complete* if every payoff is replicable — if for each $X \in \mathbb{R}^n$ some portfolio, funded with whatever capital it costs, has time-1 value X in every state. Uniqueness of the measure and completeness of the market turn out to be one condition.

Theorem 4.13 (Second Fundamental Theorem, one period and finitely many states)**.** *Suppose the market of Theorem 4.12 admits no arbitrage. Then the martingale measure is unique if and only if the market is complete — which happens exactly when K together with the constant payoffs is all of $\mathbb{R}^n$.*

Proof. Write $\mathbf{1} = (1, \dots, 1)$ for the payoff worth one in every state — discounting made the money market account worth exactly that. Recall that K collects the payoffs reachable from *zero* capital. Starting instead with capital c puts c in the money market and adds the constant $c\,\mathbf{1}$, so the payoffs replicable from any capital are exactly $K + \mathbb{R}\mathbf{1}$; the market is complete when that is all of $\mathbb{R}^n$.

A probability vector q is a martingale measure iff $\sum_i q_i\big(\tilde{S}_1(\omega_i) - \tilde{S}_0\big) = 0$, which says $q \cdot k = 0$ for every $k \in K$, i.e. $q \in K^\perp$. With the normalisation $q \cdot \mathbf{1} = 1$, the martingale measures are

$$\mathcal{M} = \big\{ q \in K^\perp : q \cdot \mathbf{1} = 1,\ q_i > 0 \text{ for every } i \big\}.$$

Set the positivity aside: $A = \{q \in K^\perp : q \cdot \mathbf{1} = 1\}$ is an affine subspace, parallel to the linear space $K^\perp \cap \mathbf{1}^\perp = (K + \mathbb{R}\mathbf{1})^\perp$. No arbitrage makes $\mathcal{M}$ non-empty, by Theorem 4.12, and $\mathcal{M}$ is A intersected with the open positive orthant, hence a non-empty relatively open subset of A. Such a subset is a single point exactly when A is, and A is a single point exactly when $(K + \mathbb{R}\mathbf{1})^\perp = \{0\}$, i.e. when $K + \mathbb{R}\mathbf{1} = \mathbb{R}^n$. So the martingale measure is unique precisely when the market is complete. □

4.4 Many Periods

Extending this to a finite tree costs surprisingly little, because a multi-period market is nothing more than a collection of one-period markets glued together, one at each node.

Let the market trade at times $t = 0, 1, \dots, T$, let Ω still be finite, and let $\mathcal{F}_t$ be the information available at time t. Since Ω is finite, $\mathcal{F}_t$ is generated by a partition of Ω into finitely many cells, and

we call each cell a node at time t. Sitting at a node ν at time t, the market you face over the next period is exactly a one-period market: today's discounted prices $\tilde{S}_t(\nu)$ are known, and tomorrow's take one of finitely many values, one for each of the successor nodes of ν.

Theorem 4.14 (First Fundamental Theorem, finite tree)**.** *A finite multi-period market admits no arbitrage if and only if there is a measure $\mathbb{Q}$ equivalent to $\mathbb{P}^{historic}$ under which the discounted price process $\tilde{S}_t$ is a martingale.*

Proof. Again one direction is the lemma of the previous section. For the converse, suppose the market admits no arbitrage.

First, no one-period market at any node admits an arbitrage either. For suppose the market at node ν at time t did, realised by holdings θ. Consider the strategy: hold nothing until time t; if the state at time t is not in ν, continue to hold nothing; if it is, take the position θ for one period and then put whatever it yields into the money market and leave it there. This is self-financing, costs nothing to set up, and its discounted value at T is the payoff of the node arbitrage — non-negative always, and strictly positive on a set of states which has positive probability, since ν itself is reached with positive probability. That is an arbitrage in the multi-period market, which we assumed there is none of.

So Theorem 4.12 applies at every node, giving for each node ν at each time t a strictly positive probability $q(\mu \mid \nu)$ on the successors μ of ν, with

$$\sum_{\mu} q(\mu \mid \nu)\, \tilde{S}_{t+1}(\mu) = \tilde{S}_t(\nu).$$

Define $\mathbb{Q}$ by multiplying these together along each path: for $\omega \in \Omega$, let $\nu_0, \nu_1, \ldots, \nu_T$ be the nodes it passes through and set

$$\mathbb{Q}(\omega) = \prod_{t=0}^{T-1} q(\nu_{t+1} \mid \nu_t).$$

This is a probability measure, being a product of conditional probabilities down a tree, and it is strictly positive on every state because each factor is, so it is equivalent to $\mathbb{P}^{\text{historic}}$. And by construction the conditional distribution of ν_{t+1} given ν_t under $\mathbb{Q}$ is $q(\cdot \mid \nu_t)$, so the displayed identity above is exactly

$$\mathbb{E}^{\mathbb{Q}}[\tilde{S}_{t+1} \mid \mathcal{F}_t] = \tilde{S}_t,$$

which is the martingale property. □

Remark. The restriction to finitely many states is not needed. With finitely many trading dates but arbitrarily many states, no arbitrage still implies the existence of an equivalent martingale measure; this is the Dalang-Morton-Willinger theorem. The geometry is the same and the work is in replacing "closed because it is a subspace of $\mathbb{R}^n$" with a genuine closedness argument, which is the whole difficulty. That should be a warning about what is coming next.

4.5 What Changes in Continuous Time

The proof above rested on two facts, and in continuous time both of them fail.

The first is that the set of attainable payoffs was closed. In $\mathbb{R}^n$ it was closed for free, being a subspace. With continuously many trading dates the attainable payoffs form a subspace of an infinite dimensional space, and such a subspace need not be closed at all. The separating hyperplane theorem needs closedness, so the argument stops applying.

The second, and the more interesting one, is that the notion of arbitrage we have been using is too weak once trading is continuous. To see this, forget the market for a moment and consider a single Brownian motion W_t on $[0,1)$, which is a martingale, so on the face of it nothing should be extractable from it.

Example 4.2 (A doubling strategy)**.** Fix a target of \$1 and bet on W until the winnings first reach it. The obstacle is that the first passage

$$\tau = \inf\{t : W_t = 1\}$$

is finite almost surely but has no bound — its mean is infinite — so holding one unit does not finish by time 1 with certainty. The repair is to compress the clock. Take the deterministic position

$$\Delta_t = \frac{1}{1-t}, \qquad 0 \le t < 1, \tag{4.4}$$

and let $\pi_t = \int_0^t \Delta_s \, dW_s$. This is a continuous local martingale, and the quadratic variation of a stochastic integral is the integral of the squared integrand, so

$$\langle \pi \rangle_t = \int_0^t \frac{ds}{(1-s)^2} = \frac{t}{1-t},$$

which runs to infinity as $t \uparrow 1$. By the Dambis-Dubins-Schwarz theorem of chapter 2 there is then a Brownian motion B with

$$\pi_t = B_{t/(1-t)}.$$

The wealth is a Brownian motion on a clock that reaches infinity before calendar time 1. Nothing has been assumed about the size of the position beyond (4.4): the compression is a consequence of the quadratic variation, and the position is the square root of the clock's speed.

Now B attains the level 1 at some almost surely finite clock time τ_B, and the calendar time corresponding to it, $t^* = \tau_B/(1+\tau_B)$, is almost surely strictly less than 1. So stop there:

$$\Delta_t = \frac{1}{1-t}\,\mathbf{1}_{\{t<t^*\}}.$$

This is adapted, since t^* is a stopping time, and the integral is defined on the whole of $[0,1]$ because the integrand is zero after t^*. The wealth satisfies $\pi_0 = 0$ and $\pi_1 = 1$ almost surely.[1] It is an arbitrage, in the exact sense of our definition, in a market whose only asset is a martingale.

Nothing is wrong with the mathematics; what is wrong is the definition. The strategy works by being prepared to lose an unbounded amount along the way, which no counterparty would fund and no exchange would clear — and "unbounded" can be made precise.

[1] `simulated`: the fraction reaching the target matches the reflection principle's $2(1-\Phi(1/\sqrt{U}))$ at clock horizon U, which is what confirms the wealth really is a Brownian motion on the stretched clock, and the calendar times at which it does are inside the interval.

Calculation 4.15 (What the strategy has to be willing to lose)**.** Write M for the lowest the wealth reaches before the target is struck, and let τ be the first time it leaves the interval $(-a, 1)$. A Brownian path leaves any bounded interval in finite time, so $\tau < \infty$ almost surely, and until τ the wealth stays in $[-a, 1]$ — so the stopped process is bounded, hence uniformly integrable, and optional stopping applies in the form chapter 1 gives it, by uniform integrability rather than by a bounded τ. Writing $p = \mathbb{P}(M \le -a)$ for the chance of touching $-a$ first, $\mathbb{E}[W_\tau] = 0$ reads $(1)(1-p) - a\,p = 0$, so

$$\mathbb{P}\big(M \le -a\big) = \frac{1}{1+a}, \tag{4.5}$$

a tail decaying like $1/a$.[2] The drawdown $-M$ is non-negative, and for such a variable the mean is the area under the tail. For $X \ge 0$ the identity $X = \int_0^X da = \int_0^\infty \mathbf{1}_{\{a<X\}}\, da$ holds outcome by outcome, so

$$\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{a<X\}}\, da\right] = \int_0^\infty \mathbb{P}(X > a)\, da,$$

exchanging $\mathbb{E}$ and $\int da$ by Tonelli, which is available because the integrand is non-negative. This is the convenient form here because the barrier law (4.5) is a tail. So

$$\mathbb{E}\big[-M\big] = \int_0^\infty \mathbb{P}\big(-M > a\big)\, da = \int_0^\infty \frac{da}{1+a} = \infty.$$

The divergence is logarithmic: a trader who would cut the position at a loss of a still faces an expected drawdown of $\mathbb{E}[\min(-M, a)] = \int_0^a \frac{dt}{1+t} = \ln(1+a)$ — the same tail, integrated only that far. It has no finite limit, which is why the uncapped mean is infinite, but it grows so slowly that ten times the loss tolerance is only a little over twice the expected drawdown.[3]

Equation (4.5) is why the repair is a bound rather than a moment condition. Requiring finite expected loss would not exclude the strategy, since one could stop it early and lose only a little on average; requiring the wealth to stay above a fixed level does exclude it, and excludes it for every level, however large. It is the continuous time version of doubling your stake at roulette until you win, and it is why the wealth process must be required to stay above some fixed level: Definition 4.7 already named the restriction, for exactly this reason, and Lemma 4.8 supplied the one fact the sufficiency argument needed from it: an admissible strategy's discounted wealth is a supermartingale, so its expectation can only fall from zero.

That is all the sufficiency argument ever needed, and it is exactly what the doubling strategy does not have: its wealth is bounded below by no constant, so there is no a to add and Fatou has nothing to work on. When the discounted price has jumps the first sentence of the proof needs more care. For a continuous local martingale the integral is again one, which is what the argument above used; but against a jump martingale an integrand that grows without bound can produce an integral that is not even a local martingale. The statement that repairs it, that a bounded-below

[2] `checked by simulation`. The grid has to be fine: a discretised path overshoots the barrier it crosses, which biases the probability upwards by a few per cent at these depths.

[3] `the capped mean` $\mathbb{E}[\min(-M, a)] = \ln(1+a)$`, simulated`. The uncapped mean cannot be simulated at all: every run stops somewhere, stopping censors the tail that carries the divergence, and the sample mean then converges to a finite number and converges to the wrong one.

integral against a local martingale is a local martingale, is the Ansel-Stricker theorem, and the conclusion is the same.

Even after restricting to admissible strategies, no arbitrage is still strictly weaker than the existence of an equivalent martingale measure. One can construct markets with no arbitrage in which no such measure exists: the reason is again closedness, and the failure shows up as a sequence of strategies whose payoffs are not arbitrages but converge to one, the losses required shrinking to zero along the way. Ruling this out is the right condition.

Definition 4.16 (No Free Lunch with Vanishing Risk)**.** A market satisfies NFLVR if there is no sequence of discounted payoffs f_n of admissible strategies starting from zero, and no sequence of constants $\epsilon_n \to 0$ with $f_n \geq -\epsilon_n$, such that f_n converges uniformly to some f with $f \geq 0$ and $\mathbb{P}(f > 0) > 0$.

Read it as: not only can you not make something out of nothing, you cannot get arbitrarily close to doing so with an arbitrarily small stake at risk. It implies no arbitrage — take $f_n = f$ and $\epsilon_n = 0$ — and is strictly stronger.

Theorem 4.17 (Delbaen and Schachermayer)**.** *Let S be a locally bounded semimartingale. The market satisfies NFLVR if and only if there is a measure equivalent to $\mathbb{P}^{historic}$ under which the discounted price process is a local martingale.*

How the proof is organised. It splits into two steps of very unequal difficulty, and only one of them is deep. Neither is carried out in full here; the aim is to show that the shape is the finite case's, not a new theory.

Step one: the attainable claims form a closed set. Write C for the discounted claims super-replicable from nothing — payoffs of admissible strategies, less any non-negative amount. In $\mathbb{R}^n$ this was free, a subspace being closed automatically. In infinite dimensions it is a theorem, and the deep one; it is where NFLVR is spent, and the remark below says what its proof costs.

Step two: closedness gives the measure. Given step one, this is the separation argument of Lemma 4.11 in an infinite-dimensional space, and it is due to Kreps and Yan. A non-negative claim $\mathbb{I}_B$ lies outside the closed convex set C; separating the two by Hahn-Banach — in the pairing of bounded claims with densities that plays the part $\mathbb{R}^n$'s dot product played before — gives a density g_B with $\mathbb{E}[g_B f] \leq 0$ for every attainable f and $\mathbb{E}[g_B \mathbb{I}_B] > 0$. The shape of C finishes it: it contains every non-positive claim, which forces $g_B \geq 0$, and it is closed under positive scaling, which puts the bound at zero. Each event B yields such a g_B, non-negative everywhere and positive somewhere on B; adding countably many with weights 2^{-n}, chosen to cover as much of Ω as they can, gives one density positive *everywhere* — any set left uncovered would supply, through its own g_B, the correction that covers more. Normalised, that density is the equivalent local martingale measure.

Every move in step two is one from the finite case carried out in a larger space. Only step one is a new theorem. □

Remark (What the hard step actually is, and why)**.** Step two above assumed the closedness. Supplying it is the whole difficulty.

The attainable payoffs are *stochastic integrals*, indexed by the strategies producing them. To show their cone C is closed one takes a sequence $f_n = \int \Delta_n \, dS$ converging to some f, and must

produce a single strategy Δ with $\int \Delta\, dS = f$. There is no reason for the integrands Δ_n to converge to anything — they live in a space with no useful compactness — so the limit strategy has to be constructed rather than extracted, and constructing it is what Delbaen and Schachermayer's proof does. It uses NFLVR to get a uniform bound on the strategies, then a compactness theorem for the resulting integrals.

The finite dimensional case had none of this because a linear subspace of $\mathbb{R}^n$ is closed automatically. That single sentence is the entire difference in difficulty between the two halves of this chapter.

It helps to know roughly what the construction costs, both to see why it is not reproduced here and to see that nothing conceptually new enters. Three ingredients do the work. NFLVR is converted into a bound: the terminal values of admissible strategies are bounded in probability, which is where the vanishing-risk sequences are ruled out and where the strengthened hypothesis earns its place. That bound feeds a compactness argument of Komlós type, which extracts from a bounded sequence not a convergent subsequence — there is none — but a sequence of convex combinations whose Cesàro averages converge almost surely; this is the step with no finite-dimensional analogue, because in $\mathbb{R}^n$ boundedness already gives convergent subsequences. The limit is then identified as a stochastic integral, which needs a result on when a limit of integrals is an integral, and Ansel and Stricker's theorem on bounded-below integrals against local martingales — the same one that repaired the jump case in Lemma 4.8 — is what closes it.

One further honesty. If S is not locally bounded the conclusion weakens: what NFLVR delivers is an equivalent *sigma-martingale* measure, and a sigma-martingale is a stochastic integral against a martingale which need not itself be a local martingale. For the diffusions in these notes the distinction never arises, and it is mentioned only so that the statement above is not read as more general than it is.

Structure (The same three moves, twice)**.** The continuous case is not a different argument but the same one carried out in a harder space.

	Finite	**Continuous**
the space	$\mathbb{R}^n$	L^∞, paired with L^1
closedness	free: a subspace	the theorem, and it is hard
separation	separating hyperplane	Hahn-Banach, in that pairing
positivity of the normal	from the cone $\mathbb{R}^n_+$	from C containing every non-positive claim
strict positivity	automatic in finite dimension	the exhaustion argument

Every row but the second is a translation. The second is a theorem. The summary is that the fundamental theorem is a separation argument wearing a hard closedness result as a hat, and a reader who has understood the finite case has understood the shape of the general one — while a reader who wants the general one in full needs a monograph.

Two features of the statement explain the shape of everything we do afterwards.

The first is that the conclusion is a *local* martingale, not a martingale. This is not a technicality to be waved away: local martingales that are not martingales appear in perfectly reasonable models,

and when one does, prices computed as expectations can fail to be the prices of the tradables they are supposed to be. Dropping local boundedness weakens the conclusion further, to a σ-martingale.

The second is that this is why the notes are organised the way they are. From here on we will always work in the direction the easy half licenses: we will posit a model, find the measure in which discounted tradables are martingales, and price by taking expectations. The theorem is what tells us that in doing so we are not choosing one pricing convention among many, but following the only one a market without arbitrage permits.

4.6 Splitting the Risk in Two

A jump model does something more useful than price the jump, and it answers the question a desk actually asks: of the risk in this trade, how much can I hedge by trading continuously, and how much must I hedge with an instrument that pays on the event itself?

The answer is precise. Under the pricing measure the claim's value V_t is a martingale, and in a market driven by a Brownian motion and a jump measure every martingale splits into a piece driven by each:

$$dV_t = \xi_t\, dW_t + \int \psi_t(x)\, \tilde{N}(dt, dx). \tag{4.6}$$

Here ξ is the claim's sensitivity to continuous movement and $\psi(x)$ its sensitivity to a jump of size x — the amount the value changes if that jump happens now.

Suppose the only thing we can trade is the underlying S, which has both kinds of movement itself. That is one instrument against two sources of risk, so we cannot expect to remove both. What we can do is remove as much as possible, and the statement of how much is the Galtchouk-Kunita-Watanabe decomposition:

$$V_T = V_0 + \int_0^T \phi_t\, dS_t + L_T, \qquad \langle L, S\rangle = 0. \tag{4.7}$$

The integral is everything trading S can deliver. The residual L is orthogonal to S — no position in S has any effect on it — and is exactly the part that continuous trading cannot touch. Its size $\mathbb{E}[L_T^2]$ is the variance of the best possible hedging error, a number that can be computed and reserved against as chapter 24 describes.

Calculation 4.18 (The hedge ratio, derived)**.** Take a price with both kinds of motion,

$$\frac{dS}{S} = \mu\, dt + \sigma\, dW + \int \left(e^x - 1\right) N(dt, dx),$$

and hold ϕ units of it against a claim worth $V(S)$. Over an instant the hedged position moves by

$$d\Pi = \underbrace{\Big(\frac{\partial V}{\partial S} - \phi\Big)\sigma S\, dW}_{\text{diffusive}} + \underbrace{\int \Big[V(Se^x) - V(S) - \phi S(e^x - 1)\Big] N(dt, dx)}_{\text{on a jump}} + (\ldots)\, dt. \tag{4.8}$$

Two features of (4.8) decide everything that follows. The diffusive term can be set to zero exactly, by one choice of ϕ, and that choice is the delta. The jump term cannot be set to zero by any choice,

because it must vanish for *every* x at once and ϕ is one number. So perfect hedging is unavailable and the question changes from eliminating the error to minimising it.

Minimise its variance. The two sources are independent, so the variance per unit time adds:

$$\operatorname{Var}\big[d\Pi\big]/dt = \Big(\frac{\partial V}{\partial S} - \phi\Big)^2\sigma^2S^2 + \int \Big[V(Se^x) - V(S) - \phi S(e^x-1)\Big]^2\nu(dx),$$

using that a Poisson random measure of intensity ν contributes $\int f(x)^2\nu(dx)$ to the variance rate of $\int f\,dN$. This is a quadratic in ϕ with a positive coefficient, so it has one minimum, found by differentiating:

$$-2\Big(\frac{\partial V}{\partial S} - \phi\Big)\sigma^2S^2 - 2\int \Big[V(Se^x) - V(S) - \phi S(e^x-1)\Big]S(e^x-1)\,\nu(dx) = 0.$$

Collecting the terms in ϕ on one side,

$$\phi\left[\sigma^2S^2 + \int S^2(e^x-1)^2\nu(dx)\right] = \sigma^2S^2\frac{\partial V}{\partial S} + \int S(e^x-1)\big[V(Se^x)-V(S)\big]\nu(dx).$$

The bracket on the left is the total variance rate of S, and the right hand side is the covariance rate of V with S — so the minimising ϕ is $d\langle V,S\rangle/d\langle S\rangle$, and no separate argument is needed to identify it with the projection of (4.7).

Structure (A hedge ratio is a regression coefficient)**.** Written as $\phi = d\langle V,S\rangle/d\langle S\rangle$ the answer is a familiar object in unfamiliar clothing: it is the slope of a least squares regression of the claim's moves on the hedge instrument's moves. Covariance over variance is what a regression coefficient is.

That is the right way to hold the whole subject. In a complete market the regression happens to be exact — the residual is zero, the coefficient is a derivative, and calling it "delta" hides the fact that it was ever a projection. Once the market is incomplete the residual is not zero and the coefficient stops being a derivative, but it is the same construction throughout. Nothing new appears with jumps; what disappears is the coincidence that made the projection look like differentiation.

The minimising ϕ is a variance-weighted compromise between the delta and a jump-specific ratio: it equals the delta only for a linear product, and for anything with curvature the gap grows with the size of the jump and with the gamma — so holding $\partial V/\partial S$ where the underlying jumps under-hedges the jump rather than approximating it conservatively.

Reading the ratio as a regression also says immediately what it optimises and what it does not. It minimises variance, which is a symmetric penalty: it treats an unexpected gain as costly as an unexpected loss, and a desk short a jump does not feel that way. A hedge chosen to minimise a one-sided measure — chapter 24's expected shortfall, say — is a different number, and the difference is largest exactly where the payoff is most asymmetric across the jump.

Now add an instrument that pays on the event. Whether the residual disappears depends on something we already have a name for: whether the jump is one size or many.

Default. With a known recovery there is a single jump, of a single size, at an unpredictable time. So (4.6) has one ψ rather than a function of x, and there are two risks and — once a credit

default swap on the same name is available — two instruments. The system is square. Solving it gives an exact hedge: a position in the underlying for the diffusive part and a determined notional of protection for the default. The residual L vanishes and the market is complete again.

This is why a credit desk reports *jump to default* as a risk in its own right. It is ψ for the default jump: the profit and loss that would arrive if the name defaulted this instant, after the continuous hedge has done what it can. It sits beside the delta on the risk report rather than inside it, because the decomposition says they are different risks requiring different instruments.

A scheduled event, in general. The date is known but the move is not, and in general it can be any size. Then $\psi(\cdot)$ is a genuine function on a continuum and spanning it would need an instrument for every size. Two or three cannot do it.

What they can do is span the first few moments. A future dated across the event hedges the expected move; a straddle expiring just after it hedges the convexity of the outcome; a strangle reaches further into the tails. Each instrument removes another moment of ψ and shrinks L without eliminating it. That is why hedging an event requires instruments expiring on the far side of it, and why each additional strike buys one more moment of the distribution.

And what a policy meeting actually is. The general case overstates the difficulty for the most important scheduled event in rates, and the reason repays attention: it moves the problem into territory we have already solved.

A central bank does not move its policy rate by an arbitrary amount. It moves in multiples of twenty-five basis points, and at the great majority of meetings the outcome is one of

$$\{\,-25\text{bp}, \quad 0, \quad +25\text{bp}\,\},$$

with fifty and occasionally seventy-five appearing only at turning points. So ψ is not a function on a continuum at all. It is a short list of numbers, and the state space of the event is finite and small.

That is this chapter's setting exactly. With n outcomes the market is complete precisely when the traded payoffs span $\mathbb{R}^n$, and n here is three or four rather than infinite — so a handful of instruments expiring across the meeting spans the event exactly rather than approximately. A fed funds future is linear in the outcome and pins only the mean, which settles a two-outcome meeting; three or four outcomes need the future plus options struck between them — a call above the target pays only on a hike, a put below only on a cut — and that strip is what the quoted hike and cut probabilities are read from. The separating-hyperplane picture of the one-period case is the right one here, in three dimensions rather than two.

It also means the risk neutral probabilities are not merely assumed to exist. They are quoted.

Example 4.3 (Reading the hike probability off the futures)**.** A fed funds future settles on the *average* effective rate over its delivery month, which is what makes the arithmetic work. Take a target rate of 4.00%, a thirty day month, and a meeting on the tenth. If the rate is r_1 after the meeting, the average over the month is

$$\frac{10 \times 4.00 + 20 \times r_1}{30},$$

so a future implying an average of 4.15% implies

$$r_1 = \frac{30 \times 4.15 - 10 \times 4.00}{20} = 4.225\%.$$

With only two outcomes on the table — hold at 4.00 or hike to 4.25 — the implied rate is a weighted average of the two, so

$$q_{\text{hike}} = \frac{4.225 - 4.00}{0.25} = 90\%.$$

Compare this with the binomial calculation that opens this chapter. There we solved for q from two possible values and a traded price, and observed that the historic probabilities never appeared. Here the market has done the solve for us and publishes the answer daily. The number quoted as "the probability of a hike" is not anyone's forecast; it is the q derived above, extracted from a price, and it will differ from the true probability by whatever risk premium the market demands.

Where the continuum comes back. None of which means the general case was wasted, because it applies as soon as the exposure is to something other than the policy rate itself.

What moves on a meeting day is not only the target. The statement, the projections and the press conference revise the expected path of *all* future meetings, and a ten year swap rate aggregates those revisions into a number that can take any value. So a position in fed funds futures faces three outcomes and is exactly hedgeable; a Bermudan swaption faces a continuum on the same afternoon and is not.

Which case one is in is decided, as usual, by the product rather than by the event.

Remark (Which hedge, though)**.** One honesty is owed here. In an incomplete market the pricing measure is not unique, and ϕ in (4.7) depends on which one was chosen and on what "best" was taken to mean. Minimising the variance of the terminal error, minimising it period by period, and super-replicating the payoff give three different hedges, and they can differ materially.

So a jump model does not hand over a hedge the way the Black-Scholes model does. It hands over a decomposition — this much is hedgeable, that much is not, and here is the instrument that would span the remainder — together with a choice about the residual that has to be made on other grounds. That is more useful than it sounds, and it is considerably more honest than a model that reports the unhedgeable part as zero because it cannot represent it.

It is a real restriction, and it deserves to be stated clearly. It is not that a jumping asset is beyond reach — credit, as above, is modelled with the same tools — but that the specific chain of reasoning in these notes, in which no arbitrage pins down a single price, needs the continuous case. Where risk arrives suddenly rather than by accumulation, a continuous model can be calibrated to the right prices and will still misstate the hedge, because it is hedging a move it does not believe can happen.

4.7 Why Every Model Here Is a Diffusion

There is one loose end, and it should be tied off before the modelling starts, because otherwise it will look like a habit rather than a result.

Every model in these notes writes the price as

$$dS_t = \mu_t\,dt + \sigma_t\,dW_t,$$

and a reader is entitled to ask what licenses that. It is a very specific shape. Why should a price take it?

The answer assembles from two things we now have, and it turns out that almost nothing is being assumed.

First, the class is forced, not chosen. A price process that is not a semimartingale admits a free lunch with vanishing risk. This is a theorem, not a convention — it follows from the Bichteler-Dellacherie theorem, and is part of what Delbaen and Schachermayer established. So a model that is arbitrage-free has no option: its price must be a semimartingale. Nothing has been assumed yet; this is the market's constraint speaking.

Second, a continuous semimartingale has only one possible shape. Every semimartingale decomposes into a local martingale plus a process of finite variation, and if the price path is continuous both pieces are continuous:

$$S_t = S_0 + \underbrace{M_t}_{\text{local martingale}} + \underbrace{A_t}_{\text{finite variation}}.$$

Chapter 2 dealt with the first piece. By Lévy's characterization and its consequence there, a continuous local martingale accumulating variance at a rate σ_t^2 is *necessarily* $M_t = \int_0^t \sigma_s dW_s$ for some Brownian motion W — there is no other continuous local martingale it could be. And a continuous finite-variation process that accumulates at a rate is $A_t = \int_0^t \mu_s ds$. Adding them,

$$S_t = S_0 + \int_0^t \mu_s \, ds + \int_0^t \sigma_s \, dW_s,$$

which is the Itô diffusion, arrived at rather than posited.

Remark (What is actually being assumed)**.** Reading the argument backwards is the useful exercise, because it isolates the one place a choice was made.

No arbitrage was not a choice — it is the premise of the subject. The decomposition was not a choice — it is a theorem. Lévy's characterization was not a choice. The only assumption in the chain is *continuity of the path*, and the only thing it excludes is a jump.

So the entire content of "we model prices as diffusions" is the claim that the price does not jump. That is an empirical claim about an asset. It is reasonable for a liquid index over a quiet afternoon. It is plainly false for a bond that can default, for a stock through an earnings announcement, for a currency with a peg that may break, and for anything over a horizon long enough to contain a crisis.

Remark (What lies outside)**.** When the assumption fails, the honest response is to enlarge the model rather than to reinterpret the volatility, and the shape of the enlargement is dictated by the same decomposition: what a general semimartingale has and a continuous one does not is a jump term. Adding it gives

$$dS_t = \mu_t \, dt + \sigma_t \, dW_t + dJ_t,$$

with J built from a Poisson random measure. Merton's jump diffusion and the variance gamma model are two such, and the general theory of Lévy processes is the study of what J may be.

One consequence deserves flagging now, because it changes the character of everything that follows. With jumps the market is generically *incomplete*: the underlying and the money market can no longer replicate an arbitrary payoff, because a jump of random size cannot be hedged by a position chosen before it happens. By Theorem 4.13, the martingale measure is then no longer unique — there is an interval of arbitrage-free prices rather than a single one, and choosing within it is a modelling decision that no amount of no-arbitrage reasoning will make for you.

Incompleteness is relative to the instruments available, though, and markets respond to it by trading the missing risk directly — chapter 2's remark on credit default swaps makes the point. So the interval of prices narrows as the traded set grows, and the practical question is never "is this model complete" but "which instrument spans the risk I cannot hedge".

That is the honest reason these notes stay with continuous models. Not that jumps are unimportant — they are how credit is modelled, and how any asset with a scheduled announcement behaves — but that continuity is what buys the uniqueness that makes this particular line of argument as sharp as it is.

References

- Harrison, J. M., & Kreps, D. M. (1979). Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory*, 20(3), 381–408.

- Harrison, J. M., & Pliska, S. R. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications*, 11(3), 215–260.

- Dalang, R. C., Morton, A., & Willinger, W. (1990). Equivalent martingale measures and no-arbitrage in stochastic securities market models. *Stochastics and Stochastic Reports*, 29(2), 185–201.

- Delbaen, F., & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing. *Mathematische Annalen*, 300(1), 463–520.

- Delbaen, F., & Schachermayer, W. (1998). The fundamental theorem of asset pricing for unbounded stochastic processes. *Mathematische Annalen*, 312(2), 215–250.

- Bichteler, K. (1981). Stochastic integration and L^p-theory of semimartingales. *Annals of Probability*, 9(1), 49–89.

Chapter 5

Black Scholes

We price a European option, and pay attention to which parts of the argument are doing work. The asset's expected return disappears, for a reason the lemma below makes precise. The formula's two terms turn out to be the same probability measured under two different numeraires, which is visible in the derivation as a completed square before chapter 6 gives it a name. We then read the equation for what it says about a hedged position — that theta and gamma are one quantity seen twice — and collect the approximations a trader carries in place of the formula.

5.1 Black Scholes Option Pricing

Definition 5.1 (Black-Scholes market)**.** The Black-Scholes market has a constant risk-free rate r, so that the discount factor is $D_t = e^{-rt}$, and one traded asset following a geometric Brownian motion,

$$\frac{dS_t}{S_t} = \mu\,dt + \sigma\,dW_t, \tag{5.1}$$

with σ constant.

Lemma 5.2 (The drift is not a parameter)**.** *Under the risk-neutral measure of chapter 4, $\mu = r$.*

Proof. That measure is characterised by D_tS_t being a martingale, so its drift vanishes. By Itô's product rule, using $dD_t = -rD_t\,dt$ and that D has no diffusion term,

$$d(D_tS_t) = -rD_tS_t\,dt + D_t\big(\mu S_t\,dt + \sigma S_t\,dW_t\big) = (\mu - r)D_tS_t\,dt + \sigma D_tS_t\,dW_t,$$

and the drift is zero for all t only if $\mu = r$. □

So the asset's expected return, the one quantity a reader might expect to matter most, does not appear in any price. Chapter 6 explains what has happened: changing measure moves drifts and cannot touch volatilities, so μ was never an observable of the option in the first place. Under the risk-neutral measure (5.1) integrates to

$$S_T = S_t \exp\left(\left(r - \tfrac{1}{2}\sigma^2\right)\tau + \sigma\sqrt{\tau}\,Z\right), \qquad \tau = T - t, \quad Z \sim N(0,1). \tag{5.2}$$

Theorem 5.3 (Black and Scholes)**.** *In the market of definition 5.1, a European call struck at K and expiring at T is worth*

$$V_t = S_t\, N(d_1) \;-\; K e^{-r\tau}\, N(d_2), \tag{5.3}$$

where N is the standard normal distribution function and

$$d_{1,2} = \frac{1}{\sigma\sqrt{\tau}}\left[\ln\frac{S_t}{K} + \Big(r \pm \tfrac{1}{2}\sigma^2\Big)\tau\right], \qquad \textit{so that} \qquad d_1 = d_2 + \sigma\sqrt{\tau}. \tag{5.4}$$

Proof. Since $D_t V_t$ is a martingale, $V_t = e^{-r\tau}\,\mathbb{E}\big[(S_T - K)^+ \,\big|\, \mathcal{F}_t\big]$. Substituting (5.2), the option finishes in the money exactly when

$$Z > \frac{1}{\sigma\sqrt{\tau}}\left[\ln\frac{K}{S_t} - \Big(r - \tfrac{1}{2}\sigma^2\Big)\tau\right] = -d_2,$$

so with φ the standard normal density the expectation splits into two integrals over the same region:

$$V_t = \underbrace{S_t \int_{-d_2}^{\infty} e^{-\frac{1}{2}\sigma^2\tau + \sigma\sqrt{\tau}\,z}\,\varphi(z)\,dz}_{\text{the asset leg}} \;-\; \underbrace{K e^{-r\tau}\int_{-d_2}^{\infty}\varphi(z)\,dz}_{\text{the cash leg}}. \tag{5.5}$$

The cash leg is immediate: $\int_{-d_2}^{\infty}\varphi = N(d_2)$ by the symmetry of φ.

The asset leg needs one manipulation. Collect the exponentials and complete the square:

$$e^{-\frac{1}{2}\sigma^2\tau + \sigma\sqrt{\tau}z}\,\varphi(z) = \frac{1}{\sqrt{2\pi}}\exp\big(-\tfrac{1}{2}\sigma^2\tau + \sigma\sqrt{\tau}z - \tfrac{1}{2}z^2\big) = \frac{1}{\sqrt{2\pi}}\exp\Big(-\tfrac{1}{2}\big(z - \sigma\sqrt{\tau}\big)^2\Big) = \varphi\big(z - \sigma\sqrt{\tau}\big).$$

The integrand is a normal density again, *shifted* by $\sigma\sqrt{\tau}$. Substituting $y = z - \sigma\sqrt{\tau}$ moves the limit with it,

$$\int_{-d_2}^{\infty}\varphi\big(z - \sigma\sqrt{\tau}\big)\,dz = \int_{-d_2-\sigma\sqrt{\tau}}^{\infty}\varphi(y)\,dy = N\big(d_2 + \sigma\sqrt{\tau}\big) = N(d_1)$$

by symmetry and (5.4), which is (5.3). □

Structure (Completing the square is a change of measure)**.** What happened is that a density multiplied by an exponential in its own variable came back as the *same* density about a different centre. That is exactly a Girsanov shift: multiplying by $e^{\sigma\sqrt{\tau}z - \sigma^2\tau/2}$ — which is a mean-one positive random variable, so a legitimate Radon-Nikodym derivative — is changing measure, and under the new measure Z has mean $\sigma\sqrt{\tau}$ instead of zero.

So (5.5) is not one expectation but two, taken under two *different* measures. The cash leg is priced in units of the money market account and $N(d_2)$ is the probability of exercise there. The asset leg is priced in units of the asset itself, and $N(d_1)$ is the probability of exercise *under that measure*. The two normal distribution functions are the same event — the option finishing in the money — measured by two different numeraires, which is why they differ by precisely the volatility term $\sigma\sqrt{\tau}$ that separates d_1 from d_2. Chapter 6 derives the formula that way from the start and draws the two probabilities against each other; here the shift arrives unbidden, out of completing a square.

Formula (5.3) is a smooth function of the spot at every positive maturity, and the payoff it is a price for is not smooth at all. The two have to meet, since an option a moment before expiry is worth what it is about to pay, and watching them meet is the most direct account of what the extra value is.

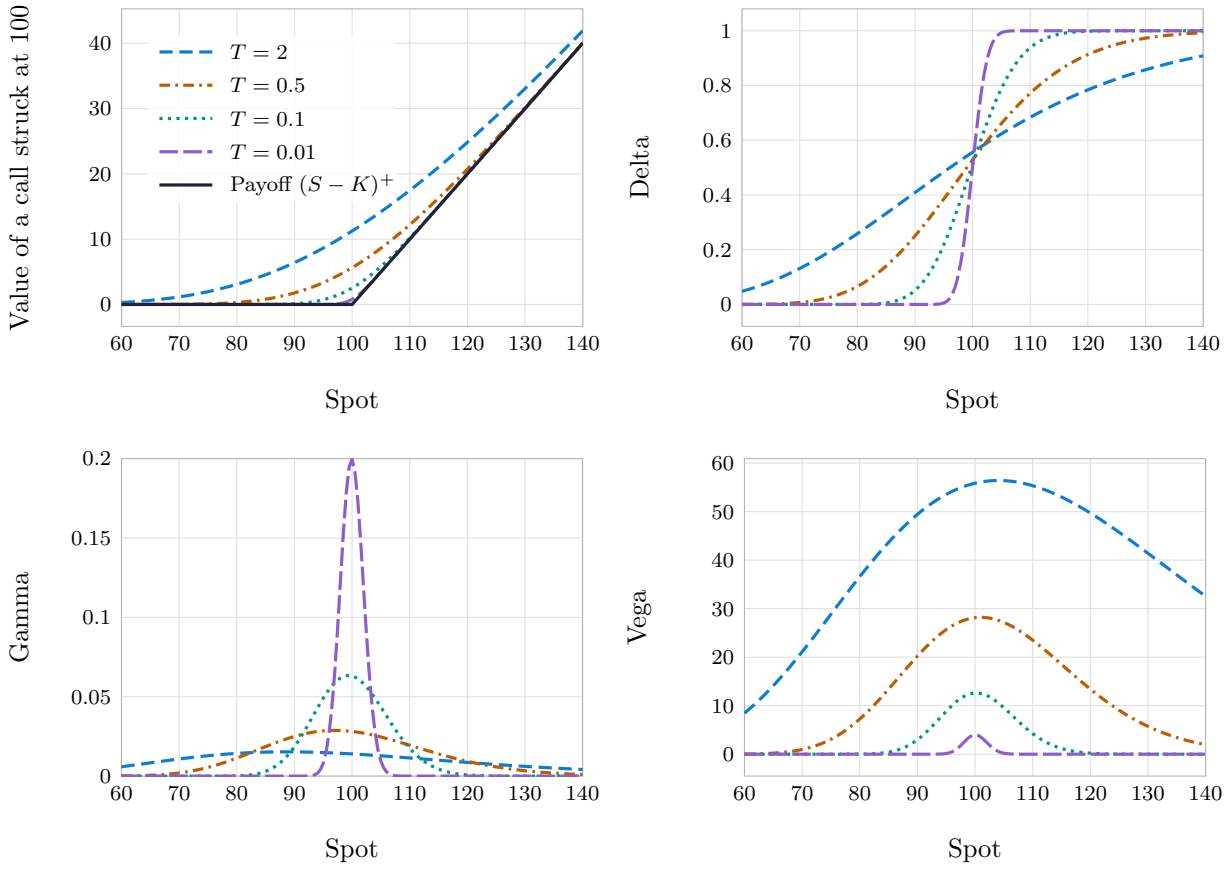

Figure 5.1: A call struck at 100 at twenty percent volatility and zero rates, with its first three sensitivities, all drawn at one maturity at a time. As expiry approaches the value falls onto $(S-K)^+$ and the curve above the payoff — the whole of the option's time value — is squeezed out of it. The derivatives say the same thing more sharply. Delta stiffens from a slope into the step that a forward would have, so a long dated option hedges like a fraction of the stock and a short dated one like all or none of it. Gamma concentrates over the strike and grows without bound, which is why the axis has to be pinned to see anything at all. And vega goes to zero everywhere at once: with no time left there is nothing for volatility to act on, so nothing to be uncertain about and nothing to pay for.

Drawn by `greeks`.

Three things in that picture stand out, each a fact about options rather than about this formula. The kink in the payoff is the same non-smooth point that made the generator of chapter 3 awkward, and the diffusion is what smooths it — time to expiry and volatility enter only through $\sigma\sqrt{\tau}$, so shortening the maturity and lowering the volatility do the identical thing to the picture. Gamma

growing without bound as vega vanishes is not a paradox but the same statement twice, and *Theta Is Gamma* below makes the relation exact. And a delta that approaches a step is a warning: it is the derivative of a function converging to one that has no derivative at the strike, which is where hedging a short dated option near its strike becomes genuinely hard rather than merely expensive.

5.2 What the Model Was Not Needed For

Theorem 5.4 (Put-call parity)**.** *Let C and P be the prices of a European call and put on the same underlying, with the same strike K and the same expiry T. Then in any market without arbitrage,*

$$C - P = P(0,T)\big(F - K\big), \tag{5.6}$$

with F the forward price to T and $P(0,T)$ the discount factor. Under the assumptions of this chapter, where the underlying pays nothing and rates are constant, $F = S_0 e^{rT}$ and (5.6) *reads $C - P = S_0 - Ke^{-rT}$.*

Proof. At expiry, for every value of S_T,

$$(S_T - K)^+ - (K - S_T)^+ = S_T - K,$$

since exactly one of the two options finishes in the money and the other is worth nothing. So a long call and a short put have, in every state, the payoff of a forward struck at K. Two portfolios with identical payoffs in every state have identical prices, by the no-arbitrage argument of chapter 4, and the forward struck at K is worth $P(0,T)(F-K)$. □

Remark (Read what is absent)**.** The proof used no model. It did not ask how S gets to S_T, whether the path is continuous, what the volatility is, or whether one exists. It is an identity between payoffs, and it therefore survives every criticism made of the Black-Scholes model in the chapters that follow — jumps, stochastic volatility, a smile, all of it. (5.6) holds in a market with any of those, because it never had an opinion about them.

This is the same distinction chapter 15 will build a whole chapter on. Some prices are fixed by static replication, and for those a model can only add an assumption to information that was already complete. Others are not, and those are what a model is for. Knowing which is which is most of knowing when to trust one.

Remark (Three things it is used for)**.**

- *A call and a put at the same strike have the same implied volatility.* Since (5.6) holds for market prices and, being a model, also for Black-Scholes prices at any volatility, subtracting the two shows the volatility that reprices the call reprices the put. This is why chapter 9 can speak of *the* smile as a function of strike alone, rather than of strike and option type, and why a desk may quote whichever of the pair is out of the money and let parity supply the other.

- *The forward can be read out of option prices.* Rearranging, $F = K + (C - P)/P(0,T)$. The option market therefore carries its own opinion about the forward, and it need not agree with the one computed from a curve. Where they differ the difference is a borrow cost, a dividend assumption, or a stale quote.

- *It is the first thing to check in an implementation.* A pricer that violates parity is wrong in a way that needs no market data to detect, and the check costs one line. Both `black76` and `Model::price` are tested against it.

Remark (When it appears to fail)**.** Quoted prices routinely violate (5.6) on a screen, and the violation is almost never an arbitrage. It is the forward: an equity with an uncertain dividend, a stock that is expensive to borrow, or a pair of quotes that were not observed at the same instant.

The right reading is that (5.6) has three market inputs and the least reliable of them is F. So rather than concluding the market is mispriced, one solves for the F that makes parity hold and asks whether that number is plausible. Used this way the relation is not a test the market can fail; it is a measuring instrument.

Exercise. Verify that the formula derived above satisfies (5.6), using $N(-x) = 1 - N(x)$ on both terms. Then deduce, without any further calculation, that the put's delta is $N(d_1) - 1$ and that a call and a put at the same strike have the same gamma and the same vega. (Hint: differentiate (5.6); the right hand side has no volatility in it, and its derivative in S_0 is one.)

5.3 The Option's Own Diffusion

The derivation above computed an expectation. There is a second route to the same answer that says more about what the option *is*, and it starts by asking what process the option price itself follows.

The option is worth $V(t, S_t)$, a function of time and of a diffusion, and chapter 3 already says what the drift of such a thing is. Writing $\mathcal{L}$ for the generator of the stock,

$$dV = \left(\frac{\partial V}{\partial t} + \mathcal{L}V \right) dt + \sigma S \frac{\partial V}{\partial S}\, dW_t,$$

the extra $\partial V/\partial t$ appearing because V depends on time directly as well as through S. So the option is itself an Itô diffusion, driven by the same Brownian motion as the stock — which is chapter 4's replication argument appearing as a statement about coefficients rather than about portfolios.

Now impose no arbitrage. The option is a tradable, so by chapter 4 its discounted price is a martingale in the risk neutral measure, which is to say its drift is rV:

$$\frac{\partial V}{\partial t} + \mathcal{L}V = rV. \tag{5.7}$$

That is the Black-Scholes equation, and how little was needed to get here deserves a pause: the drift of a function of a diffusion, and the statement that a traded thing drifts at the riskless rate. Written out, with $\mathcal{L}$ the generator under the risk neutral dynamics,

$$\frac{\partial V}{\partial t} + rS\frac{\partial V}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = rV. \tag{5.8}$$

Structure (The same equation for every model in these notes)**.** Equation (5.7) contains nothing specific to Black-Scholes. Substituting a different generator gives the pricing equation of a different

model, and that is the only step: local volatility in chapter 9, the short rate models of chapter 8, the quasi-Gaussian models of chapter 12 all satisfy (5.7) with their own $\mathcal{L}$.

So none of those chapters derives a pricing equation again. They specify a generator, which is what specifying a model means.

This is the Black-Scholes partial differential equation, and the correspondence just used — that the expectation of a payoff under a diffusion solves a PDE of this shape, and conversely — is the Feynman-Kac theorem. It is the bridge the whole subject walks back and forth across: an expectation is something to simulate, a PDE is something to solve on a grid, and Feynman-Kac says they compute the same object. Chapter 9's Dupire formula is derived from the forward version of the same correspondence.

5.4 Theta Is Gamma

Equation (5.8) looks like a technical detour until it is read as a statement about a trading position, at which point it becomes the most useful thing in this chapter.

Take $r = 0$, which costs nothing and removes the financing terms. Writing $\Theta = \partial V/\partial t$ and $\Gamma = \partial^2 V/\partial S^2$ in the trader's notation, (5.8) reduces to

$$\Theta = -\frac{1}{2}\sigma^2 S^2 \Gamma. \tag{5.9}$$

These are not two independent quantities that happen to be related. They are the same number, with a sign. A position with positive gamma has negative theta, in exact proportion, and the constant of proportionality is the variance.

To see why that matters, hold the option, hedge it, and account for a single day.

Calculation 5.5 (The daily profit and loss of a hedged option)**.** Hold one option and short Δ shares. Over a short interval, expanding the option's value to second order,

$$\begin{aligned} d\Pi &= dV - \Delta\, dS \\ &= \left(\Theta\, dt + \Delta\, dS + \frac{1}{2}\Gamma (dS)^2\right) - \Delta\, dS \\ &= \Theta\, dt + \frac{1}{2}\Gamma (dS)^2, \end{aligned}$$

the first-order terms cancelling, which is what the hedge was for. Substituting (5.9),

$$d\Pi = \frac{1}{2}\Gamma S^2 \left[\left(\frac{dS}{S}\right)^2 - \sigma^2\, dt\right]. \tag{5.10}$$

Read the bracket. The first term is the move the market actually made, squared. The second is the move the option was priced for. A hedged option earns the difference between them, scaled by gamma.

So an option is not a bet on direction — the hedge removed that — and it is not really a bet on volatility either. It is a bet on *realised variance against implied variance*, settled a day at a time. Buy an option, hedge it, and you make money on every day the market moves more than the price assumed and lose on every day it moves less.

Remark (What the premium is)**.** Equation (5.10) also closes a question the replication argument of chapter 4 left open. Rebalancing a hedge buys after a rise and sells after a fall, which loses money on every round trip; that loss is the $\frac{1}{2}\Gamma(dS)^2$ term, and its expected size over the life of the option is

$$\mathbb{E}\left[\int_0^T \frac{1}{2}\Gamma S^2\sigma^2\,dt\right],$$

which by (5.9) is exactly the accumulated theta — that is, the premium. The option premium is not a fee for optionality in the abstract. It is the amount the seller expects to lose rehedging, paid up front.

Which is also why the seller cares about the frequency of large moves rather than only their variance. The premium compensates for the average of $(dS)^2$ and for nothing else, so a distribution with the same average and fatter tails hands over the same premium against a worse experience — and since that is a claim about two distributions rather than about one model, the next calculation shows both.

Calculation 5.6 (The same premium, two worlds)**.** Sell the same one year at-the-money call in two markets whose total variance is identical, so the Black-Scholes price is identical and the seller takes in the same 7.97 either way. In the first all of that variance is diffusive. In the second half of it arrives as compensated jumps at a rate of two a year. Hedge both two hundred and fifty times at the same Black-Scholes delta.[1]

	All diffusion	**Half in jumps**
Mean profit and loss	−0.00	+0.12
Standard deviation	0.44	2.70
Worst one per cent, averaged	−1.50	−12.54
Worst outcome seen	−4.31	−32.86
Fraction of paths that made money	49.9%	63.8%

The premium was fair in both — the mean is zero to within the simulation error, which is what (5.10) guarantees once the variance is matched. Everything else differs. The tail is eight times worse, and the position *wins more often*: a seller in the jump world is right on nearly two thirds of paths and loses thirty-three on the one that goes wrong.

Remark (What that pattern is)**.** The last row is the one to remember, because the shape it describes recurs throughout these notes. A position that makes a little money most of the time and loses a great deal rarely is not a good position that occasionally has bad luck; it is a short option, whether or not anybody wrote one down. Chapter 22 finds the same profile in six trades that have nothing else in common, and chapter 24 is about measuring it, since a mean and a standard deviation describe the left column adequately and the right column not at all.

[1] `pathwise::hedge_under_jumps`.

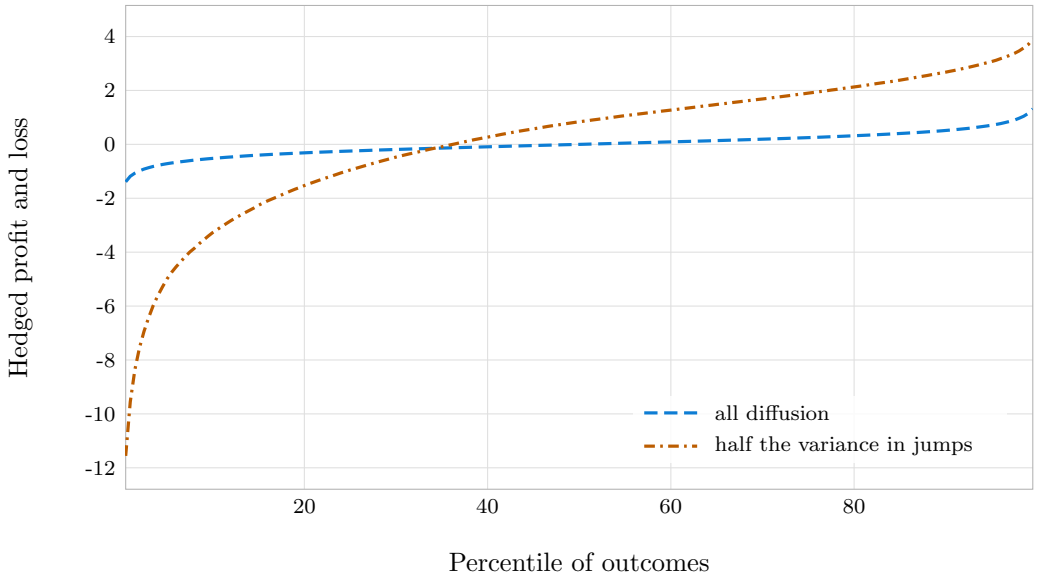

Figure 5.2: Hedged profit and loss at each percentile, selling the same option for the same premium into two markets with the same total variance. The diffusive curve is nearly flat — almost every outcome is close to zero, which is what a working hedge looks like. The jump curve is above it across most of the range and falls off a cliff at the left. Same average of $(dS)^2$, same price, and an entirely different thing to live through, which is why a seller cares about the shape of the distribution and not only its second moment.

Drawn by `hedge_under_jumps`.

5.5 Numbers Carried in the Head

The formula of §5.1 is not something anyone evaluates mentally, and traders do not try. What they carry instead is a handful of approximations, all of which come from the same place: at the money, the normal density is flat near zero, so $N(x) \approx \frac{1}{2} + n(0)x$ with $n(0) = 1/\sqrt{2\pi} \approx 0.4$.

Calculation 5.7 (The at-the-money option price)**.** With $r = q = 0$ and $K = F$, we have $d_{1,2} = \pm\frac{1}{2}\sigma\sqrt{T}$, so

$$\begin{aligned} C &= F\big[N(d_1) - N(d_2)\big] \\ &\approx F\Big[\big(\tfrac{1}{2} + n(0)\tfrac{\sigma\sqrt{T}}{2}\big) - \big(\tfrac{1}{2} - n(0)\tfrac{\sigma\sqrt{T}}{2}\big)\Big] \\ &= F\, n(0)\, \sigma\sqrt{T} \;\approx\; 0.4\, F\, \sigma\sqrt{T}. \end{aligned}$$

The error is under half a percent for ordinary parameters and only reaches one percent at volatilities above fifty. Differentiating the same expression in σ gives the vega, and the rest follow:

at-the-money price	$0.4\, F\sigma\sqrt{T}$
vega	$0.4\, F\sqrt{T}$ per unit of volatility
delta	about $\frac{1}{2}$
gamma	$0.4\,/\,(F\sigma\sqrt{T})$
daily break-even move	$\sigma/\sqrt{252} \approx \sigma/16$

The last line is the one used most. Since $\sqrt{252} = 15.87$, a volatility quoted annually divides by roughly sixteen to give the daily move it implies. Twenty vol is a one and a quarter percent day; thirty-two vol is a two percent day. By (5.10) that is exactly the break-even: a day that moves less than $\sigma/16$ loses money for the holder of a hedged option, and a day that moves more makes it.

Example 5.1 (A trade, decided in one's head)**.** An index is at 4,000 and its three month at-the-money options are quoted at 16 volatility. Then

- the option costs about $0.4 \times 4{,}000 \times 0.16 \times \sqrt{0.25} = 128$ index points, a little over three percent of spot;
- its vega is about $0.4 \times 4{,}000 \times 0.5 = 800$ points per unit of volatility, so eight points per volatility point;
- and it breaks even on a daily move of $16/16 = 1\%$, which is 40 points.

So the decision to buy it is a view that the index will move more than forty points a day, on average, over the next three months — which is a question about the world rather than about the model, and can be answered from a chart. The whole apparatus of this chapter has been reduced to a comparison a trader can make between two sentences of a phone call.

5.6 Why the Formula Is a Price

We wrote down a payoff, took an expectation under a particular measure, and called the answer a price. Why should anyone pay it?

The answer is the argument of chapter 4, carried into continuous time: because the option can be *manufactured*. If holding Δ_t shares and adjusting continuously reproduces the payoff exactly, then the option and that strategy are the same object, and two things that pay identically must cost identically or there is an arbitrage. The formula is not a forecast of what the option will be worth. It is the cost of building one.

Two features of the result follow from this, and both surprise people meeting them for the first time.

Remark (The drift is absent, and that is the point)**.** The real world drift μ appears nowhere in the formula. Two investors who disagree completely about whether the stock will rise must still agree on the option's price — because neither of them is being asked to forecast anything. They are being asked what it costs to build a payoff, and the hedge removes the drift from the answer just as it removes it from the portfolio.

This is the continuous time version of the observation in chapter 4 that the historic probabilities did not appear in the binomial price. It is the same fact, and it has the same cause.

Remark (Continuously is not a word about the world)**.** Replication is exact only if the position is adjusted continuously, and nobody does that. Real hedging happens at discrete times, so real replication is imperfect, and the difference is a genuine risk carried by anyone who sells an option.

How imperfect, and in what way, is a question a simulation can answer.

Read the two panels together, because they say different things.

The upper one is reassuring. The error does go to zero, and at the classical rate: quadruple the number of rebalances and the spread halves. Daily hedging of a one year option leaves a standard deviation of a few tenths of a percent of spot, against a premium of about eight percent. Replication is not a fiction.

The lower one is the warning, and it has to be read carefully, because the asymmetry is a property of the tail and not of the body. The mean is zero — the premium is right — and within one and a half standard deviations the two sides carry very nearly equal mass. The difference sits at the thousandth quantile. Hedging eight times, the worst outcomes reach four and a half standard deviations below the mean while the best stop short of three, against a symmetric reference that would go the same distance either way. That is why the panel is drawn on a log scale: on a linear one the region where the two sides differ is indistinguishable from zero.

So a desk that sells options and hedges them discretely collects a little most of the time and occasionally takes a large loss. But notice what else the panel shows, because it qualifies the warning: the asymmetry *shrinks as the rebalancing gets finer*, from a skewness of -0.42 at eight rebalances to -0.18 at a hundred and twenty-eight. This left tail is the price of trading discretely, and trading less discretely reduces it.

What does not shrink is the left tail of Calculation 5.6, because a jump is a jump however often the hedge is adjusted. The two are different warnings, and should be kept apart: this one is about

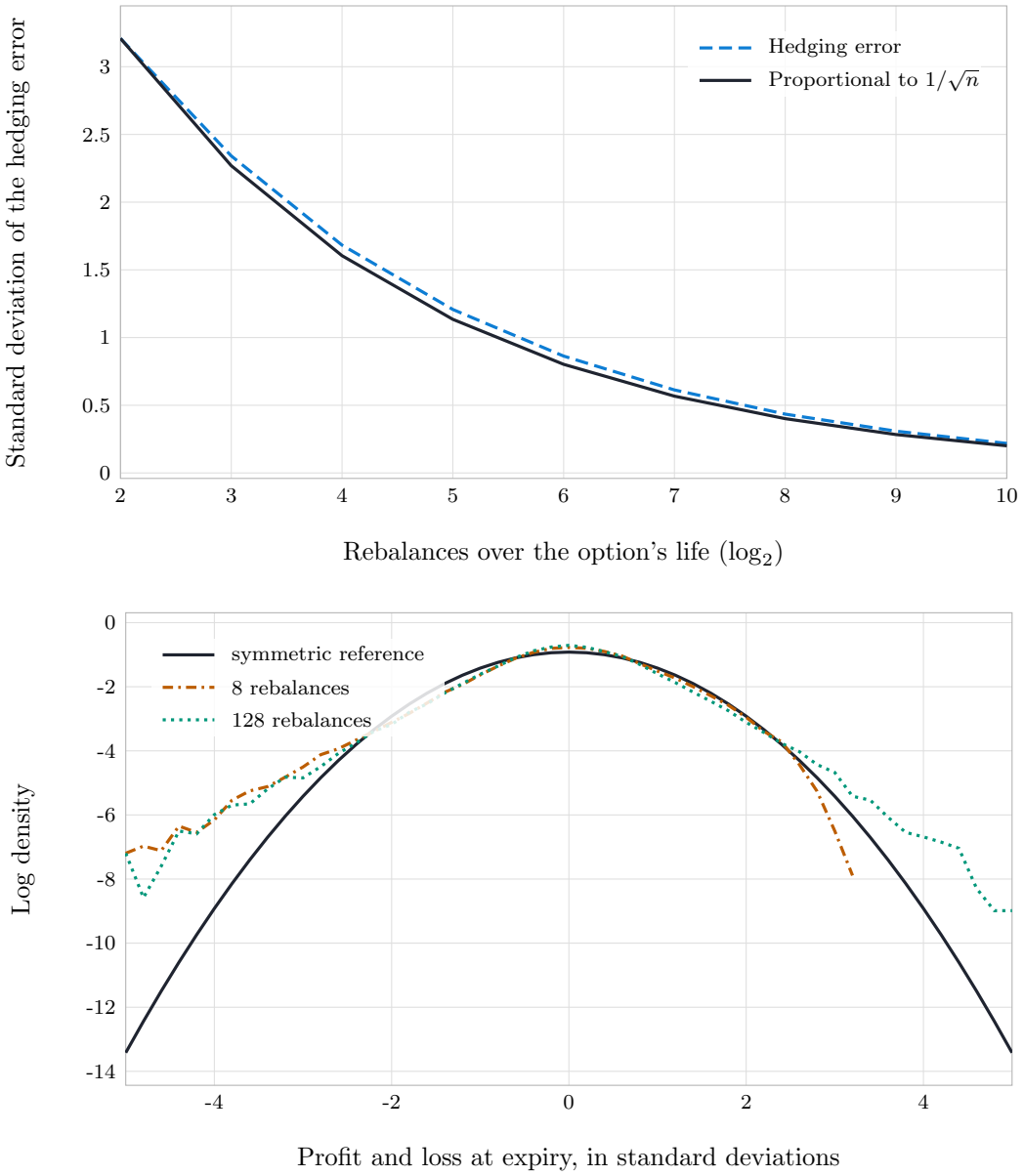

Figure 5.3: Selling a one year at-the-money call for its Black-Scholes price and hedging it n times before expiry. Above: the spread of the outcomes falls in proportion to $1/\sqrt{n}$ — the reference curve is anchored at the coarsest point rather than fitted, so its agreement further along is a prediction. Below: the shape of the profit and loss, standardised so that only the shape is being compared, and drawn against a symmetric reference on a log scale — the asymmetry lives in the last tenth of a per cent, where a linear density is visually zero. At eight rebalances the left arm reaches five standard deviations while the right stops at three; by a hundred and twenty-eight the two are close to level again, which says the skew is the cost of trading discretely rather than a permanent feature of the position.

Drawn by `delta_hedge` and `hedge_statistics`.

how often you trade, that one about what the world does between your trades. Only the second is a reason to care about the frequency of large moves more than a formula containing only σ would suggest.

References

- Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3), 637–654.

- Merton, R. C. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science*, 4(1), 141–183.

- Karatzas, I., & Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer.

Chapter 6

Numeraires

We change the unit of account, which is the most useful single manoeuvre in derivative pricing and the least intuitive. Girsanov's theorem says a measure change moves drifts and cannot touch volatilities, and that asymmetry is what makes the technique work: choosing a numeraire chooses which quantity is a martingale, and the right choice makes a hard expectation trivial. We derive the forward and annuity measures that chapter 7 and chapter 15 need, and find that no-arbitrage assigns one price to each source of risk — the same projection that reappears as a hedge in chapter 23.

6.1 Change of Measure

Definition 6.1 (Radon-Nikodym Derivative)**.** Consider two equivalent probability measures $\mathbb{P}$ and $\hat{\mathbb{P}}$ on a measurable space (Ω, Σ). The Radon-Nikodym derivative $d\hat{\mathbb{P}}/d\mathbb{P} : \Omega \to \mathbb{R}$ is defined such that for any subset $A, \Omega \supseteq A \in \Sigma$,

$$\int_A d\hat{\mathbb{P}} = \int_A (d\hat{\mathbb{P}}/d\mathbb{P}) d\mathbb{P}. \tag{6.1}$$

Suppose $(\Omega, \mathcal{F}, \mathcal{F}_t)$ is a filtered probability space, then note from the above definition we have

$$(d\hat{\mathbb{P}}/d\mathbb{P})|_{\mathcal{F}_t} = \mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})|\mathcal{F}_t] \tag{6.2}$$

$(d\hat{\mathbb{P}}/d\mathbb{P})|_{\mathcal{F}_t}$ is also written as $(d\hat{\mathbb{P}}/d\mathbb{P})_t$ and is thus a martingale stochastic process (by iterated conditioning) in $\mathbb{P}$.

Example 6.1. A die roll makes the Radon-Nikodym derivative concrete: multiple probability distributions can be assigned to the same six outcomes.

The probability in $\hat{\mathbb{P}}$ of getting an odd number is $\int_{\omega\in\{1,3,5\}} d\hat{\mathbb{P}} = 1/2 + 1/8 + 1/32 = 21/32 = 3 * 1/6 + 3/4 * 1/6 + 3/16 * 1/6 = \int_{\omega\in\{1,3,5\}} (d\hat{\mathbb{P}}/d\mathbb{P}) d\mathbb{P}$

ω	$\mathbb{P}$	$\hat{\mathbb{P}}$	$d\hat{\mathbb{P}}/d\mathbb{P}$
1	1/6	1/2	3
2	1/6	1/4	3/2
3	1/6	1/8	3/4
4	1/6	1/16	3/8
5	1/6	1/32	3/16
6	1/6	1/32	3/16

Theorem 6.2 (Abstract Bayes' Theorem)**.** *Let $\mathbb{P}$ and $\hat{\mathbb{P}}$ be two measures on measurable space $(\Omega, \mathcal{F})$. Let $\mathcal{G} \subset \mathcal{F}$ be another sigma algebra on Ω. Then for any $A \in \mathcal{G}$ and random variable X*

$$\mathbb{E}^{\hat{\mathbb{P}}}[X|\mathcal{G}] = \frac{\mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})X|\mathcal{G}]}{\mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})|\mathcal{G}]}. \tag{6.3}$$

Proof. We show that for any $A \in \mathcal{G}$

$$\mathbb{E}^{\hat{\mathbb{P}}}[X|\mathcal{G}]\mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})|\mathcal{G}] = \mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})X|\mathcal{G}]$$

Since the random variables involved are constant over A, we can check equality on integrals over A.

$$\begin{aligned}
\int_A \mathbb{E}^{\hat{\mathbb{P}}}[X|\mathcal{G}]\mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})|\mathcal{G}]d\mathbb{P} &= \int_A \mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})\mathbb{E}^{\hat{\mathbb{P}}}[X|\mathcal{G}]|\mathcal{G}]d\mathbb{P} && (\mathbb{E}^{\hat{\mathbb{P}}}[X|\mathcal{G}] \text{ is } \mathcal{G} \text{ measurable}) \\
&= \int_A (d\hat{\mathbb{P}}/d\mathbb{P})\mathbb{E}^{\hat{\mathbb{P}}}[X|\mathcal{G}]d\mathbb{P} && (\text{conditional expectation}) \\
&= \int_A \mathbb{E}^{\hat{\mathbb{P}}}[X|\mathcal{G}]d\hat{\mathbb{P}} && (\text{Radon-Nikodym}) \\
&= \int_A X d\hat{\mathbb{P}} && (\text{conditional expectation}) \\
&= \int_A (d\hat{\mathbb{P}}/d\mathbb{P})X d\mathbb{P} && (\text{Radon-Nikodym}) \\
&= \int_A \mathbb{E}^{\mathbb{P}}\left[(d\hat{\mathbb{P}}/d\mathbb{P})X \,\middle|\, \mathcal{G}\right] d\mathbb{P}.
\end{aligned}$$

The last step is the definition of conditional expectation. □

Remark. Taking $X = V_T$ where V_s is $\mathcal{F}_s$ adapted and $G = \mathcal{F}_t$ for $t < T$ in the above theorem, we get the very useful formula for measure change for conditional expectations on filtered spaces,

$$\mathbb{E}^{\hat{\mathbb{P}}}[V_T|\mathcal{F}_t] = \mathbb{E}^{\mathbb{P}}[\frac{(d\hat{\mathbb{P}}/d\mathbb{P})_T}{(d\hat{\mathbb{P}}/d\mathbb{P})_t}V_T|\mathcal{F}_t] \tag{6.4}$$

Proof.

$$\begin{aligned}
\mathbb{E}^{\hat{\mathbb{P}}}[V_T|\mathcal{F}_t] &= \frac{\mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})V_T|\mathcal{F}_t]}{\mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})|\mathcal{F}_t]} && \text{(abstract Bayes' theorem)}\\
&= \frac{\mathbb{E}^{\mathbb{P}}[\mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})V_T|\mathcal{F}_T]|\mathcal{F}_t]}{\mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})|\mathcal{F}_t]} && \text{(iterated conditioning)}\\
&= \frac{\mathbb{E}^{\mathbb{P}}[\mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})|\mathcal{F}_T]V_T|\mathcal{F}_t]}{\mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})|\mathcal{F}_t]} && (V_T \text{ is } \mathcal{F}_T \text{ measurable})\\
&= \frac{\mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})_T V_T|\mathcal{F}_t]}{(d\hat{\mathbb{P}}/d\mathbb{P})_t} && \text{(martingale property of Radon Nikodym derivative)}\\
&= \mathbb{E}^{\mathbb{P}}[\frac{(d\hat{\mathbb{P}}/d\mathbb{P})_T}{(d\hat{\mathbb{P}}/d\mathbb{P})_t}V_T|\mathcal{F}_t] && ((d\hat{\mathbb{P}}/d\mathbb{P})_t \text{ is } \mathcal{F}_t \text{ measurable}).
\end{aligned}$$

□

We consider processes until terminal time S i.e. $\mathcal{F} = \mathcal{F}_S$ and $(d\hat{\mathbb{P}}/\mathbb{P}) = (d\hat{\mathbb{P}}/\mathbb{P})_S$. We take a strictly positive martingale process to be the Radon-Nikodym derivative:

$$df_t = f_t\sigma(t)dW_t; \;\; f_t = e^{-\int_0^t \frac{1}{2}\sigma^2(s)ds + \int_0^t \sigma(s)dW_s}. \tag{6.5}$$

Theorem 6.3 (Girsanov Theorem)**.** *If $W_{1,t}$ and $W_{2,t}$ are (possibly correlated) Brownian motions in $\mathbb{P}$ and the Radon-Nikodym derivative $(d\hat{\mathbb{P}}/d\mathbb{P})_t = f_t$ is given by $f_t = e^{-\int_0^t \frac{1}{2}\sigma^2(s)ds + \int_0^t \sigma(s)dW^{\mathbb{P}}_{2,s}}$ then $W_{1,t} - \int_0^t \sigma(s)dW_{1,s}dW_{2,s}$ is a Brownian motion in $\hat{\mathbb{P}}$, where $dW_{1,s}dW_{2,s}$ is the covariation rate of W_1 and W_2 (zero if they are independent, $\rho\, ds$ if correlated at ρ, and ds in the case $W_1 = W_2$).*

Proof. We show that $X_t = W_{1,t} - \int_0^t \sigma(s)dW_{1,s}dW_{2,s}$ follows normal distribution with mean 0 and variance t in $\hat{\mathbb{P}}$. Other required properties can be verified easily. We show that the moment generating function of X_t is same as that of normal distribution with mean 0 and variance t.

$$\begin{aligned}
\mathbb{E}^{\hat{\mathbb{P}}}[e^{-yX_t}] &= \mathbb{E}^{\mathbb{P}}[(d\hat{\mathbb{P}}/d\mathbb{P})_t e^{-yX_t}]\\
&= \mathbb{E}^{\mathbb{P}}[e^{-\int_0^t \frac{1}{2}\sigma^2(s)ds + \int_0^t \sigma(s)dW_{2,s}} e^{-yX_t}]\\
&= e^{-\int_0^t \frac{1}{2}\sigma^2(s)ds}\mathbb{E}^{\mathbb{P}}[e^{\int_0^t \sigma(s)dW_{2,s} - yX_t}]\\
&= e^{-\int_0^t \frac{1}{2}\sigma^2(s)ds}\mathbb{E}^{\mathbb{P}}[e^{\int_0^t \sigma(s)dW_{2,s} - y(W_{1,t} - \int_0^t \sigma(s)dW_{1,s}dW_{2,s})}]\\
&= e^{-\int_0^t \frac{1}{2}\sigma^2(s)ds}\mathbb{E}^{\mathbb{P}}[e^{\int_0^t (\sigma(s)dW_{2,s} - ydW_{1,s} + y\sigma(s)dW_{1,s}dW_{2,s})}].
\end{aligned}$$

Define the martingale $Z_{y,t} = \int_0^t (\sigma(s)dW_{2,s} - ydW_{1,s})$. Then $e^{Z_{y,t} - \frac{1}{2}\int_0^t dZ_{y,s}dZ_{y,s}}$ is a martingale as well, with

$$dZ_{y,s}dZ_{y,s} = (\sigma^2(s) + y^2)ds - 2y\sigma dW_{1,s}dW_{2,s}.$$

We then have

$$\begin{aligned}\mathbb{E}^{\hat{\mathbb{P}}}[e^{-yX_t}] &= e^{-\int_0^t \frac{1}{2}\sigma^2(s)ds}\mathbb{E}^{\mathbb{P}}[e^{Z_{y,t}-\frac{1}{2}\int_0^t dZ_{y,s}dZ_{y,s}+\frac{1}{2}\int_0^t(\sigma^2(s)+y^2)ds}] \\ &= e^{\frac{1}{2}y^2t}\mathbb{E}^{\mathbb{P}}[e^{Z_{y,t}-\frac{1}{2}\int_0^t dZ_{y,s}dZ_{y,s}}] \\ &= e^{\frac{1}{2}y^2t}(e^{Z_{y,t}-\frac{1}{2}\int_0^t dZ_{y,s}dZ_{y,s}})|_{t=0} \\ &= e^{\frac{1}{2}y^2t}.\end{aligned}$$

□

Structure (A measure change moves the drift and cannot touch the volatility)**.** The theorem's content is an asymmetry: it changes the drift of a process and leaves its diffusion coefficient exactly where it was. That asymmetry explains several things at once, and stating it precisely matters, because the loose version is misleading in a way that bites later.

Quadratic variation is computed *pathwise*: $\langle M\rangle_t$ is a limit of sums of squared increments of the trajectory that actually occurred, so it is a random variable and not an expectation. Different paths have different quadratic variations — in a stochastic volatility model $\langle S\rangle_t = \int_0^t v_s S_s^2\, ds$, and a path that spent its life in a high-variance regime has a larger one than a path that did not. What is measure-invariant is therefore not a number but a *function of the path*: the map from trajectory to accumulated variance mentions no measure anywhere, so two equivalent measures assign the same quadratic variation to the same path. They disagree about which paths are likely, not about what each path's variance is. Equivalence is doing exactly one job in that sentence: the limit converges in probability, so $\langle M\rangle$ is defined only up to null sets, and equivalent measures are precisely those that agree on which sets those are.

That is why volatility can be estimated from one realisation and drift cannot. An observer sees a single trajectory; the sums of squared increments of *that* trajectory converge without reference to any measure, so no averaging over paths that did not happen is required. A statement about drift is the opposite kind of statement — it is a claim about the average over the paths one did not see — so it needs either many independent histories, of which there is one, or a long span of time. Chapter 21 measures the consequence: over a fixed window, sampling faster drives the volatility error to zero and leaves the drift error exactly where it was. It is also why chapter 22 can measure realised volatility and must argue about expected returns, and why every model in these notes is calibrated on volatilities and never on drifts.

It is also why the risk neutral measure exists at all. Pricing requires the drift to be one particular thing, the measure change is free to set it, and nothing about the observable roughness of the path has to be disturbed to do so.

One consequence deserves drawing out, because collapsing it is the commonest way to misread the paragraph above. The quadratic variation of a given path is measure-free; the *distribution* of quadratic variation across paths is not. So

$$\mathbb{E}^{\mathbb{P}}\big[\langle S\rangle_T\big] \neq \mathbb{E}^{\mathbb{Q}}\big[\langle S\rangle_T\big]$$

is perfectly consistent with everything just said, and it is the variance risk premium that chapter 21 measures — the reason implied volatility exceeds realised on average. Girsanov leaves σ alone as a *coefficient* and does not leave the law of accumulated variance alone at all, because in a stochastic

volatility model it changes the drift of v itself. The two statements to keep apart are that the realised variance of the path one saw is a fact, and that the variance one expects is a choice of measure.

Remark (What a change of measure does, and what it does not do)**.** The algebra above is dense enough to lose the idea in, and the idea is simple. A change of measure does not move anything. Every path the world could take is still available afterwards, ending where it always would have. What changes is how much each one counts.

The die in the example at the start of this chapter is the honest picture of it: the same six faces, reweighted. Girsanov is that operation performed on a continuum of paths rather than six outcomes, and the Radon-Nikodym derivative is the column of weights.

For the geometric Brownian motion of chapter 5 the weights can be written down. If the asset drifts at μ in the real world and at r under the risk neutral measure, the change of measure is governed by

$$\theta = \frac{\mu - r}{\sigma}, \tag{6.6}$$

the excess return per unit of risk, called the market price of risk. A path ending at S_T carries the weight

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = e^{-\theta W_T - \frac{1}{2}\theta^2 T}, \tag{6.7}$$

which decreases in W_T and therefore in S_T: the risk neutral measure counts the good outcomes for less.

Reweighting the paths this way is the same as shifting the noise. Girsanov applied to (6.7) makes $W_t^{\mathbb{Q}} = W_t + \theta t$ a Brownian motion under $\mathbb{Q}$, and substituting $dW = dW^{\mathbb{Q}} - \theta\, dt$ into the real-world dynamics,

$$dS_t = \mu S_t\, dt + \sigma S_t\, dW_t = (\mu - \sigma\theta)\, S_t\, dt + \sigma S_t\, dW_t^{\mathbb{Q}} = r\, S_t\, dt + \sigma S_t\, dW_t^{\mathbb{Q}},$$

using $\sigma\theta = \mu - r$. The drift falls from μ to r; the diffusion coefficient σS_t is the same on every line. That is the whole content of the phrase "removing the risk premium", and (6.6) says exactly how much removing costs.

Structure (One price per risk, and no-arbitrage is what sets it)**.** The scalar $\theta = (\mu - r)/\sigma$ above is a ratio that is easy to read as bookkeeping. With more than one asset it stops being bookkeeping and becomes the fundamental theorem in coordinates.

Take n assets driven by d Brownian motions, with σ the matrix of loadings and $\mu - r\mathbf{1}$ the vector of excess returns. Girsanov shifts each Brownian motion by some θ_j, and asset i's drift falls by its own exposure to those shifts. Demanding that *every* asset end up drifting at r under *one* measure is therefore the linear system

$$\mu - r\mathbf{1} = \sigma\,\theta. \tag{6.8}$$

Read the indices. There is one θ_j per Brownian motion and not one per asset: a price of risk belongs to the *risk*, and each asset's excess return is forced to be its own loadings against those

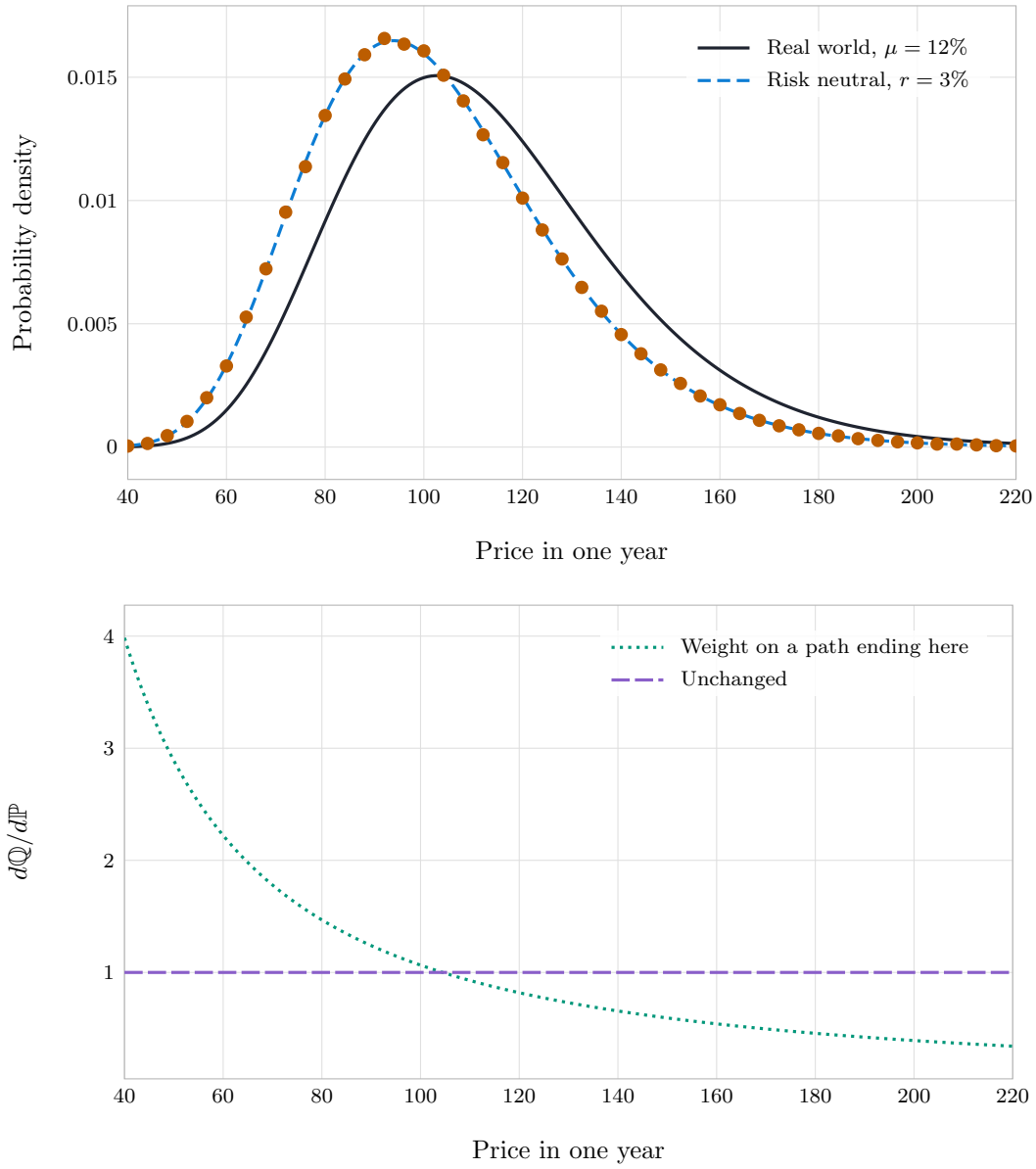

Figure 6.1: A change of measure, performed rather than described. The dotted sample is drawn under the real world measure at $\mu = 12\%$ and never redrawn; each draw is then multiplied by the weight (6.7). It lands on the risk neutral density at $r = 3\%$, which the sampling never saw. The second panel is the weight itself, tilting down across the outcomes. Drag μ onto the interest rate and the weight flattens onto one, the two densities merge, and the sample follows them — with no risk premium there is nothing to reweight, and the two measures coincide.

Drawn by `radon_nikodym` and `reweighted_histogram`.

common prices. Two assets exposed to the same factor must reward that exposure at the same rate, whatever else differs between them.

Which is a statement no-arbitrage can enforce, and does. If two assets on one Brownian motion offered different Sharpe ratios, sell the one paying less per unit of risk, buy the one paying more, in the ratio that cancels the exposure — the position has no randomness left and a positive drift. Solvability of (6.8) is exactly the absence of that trade, which is chapter 4's fundamental theorem written with matrices instead of measures:

Statement about (6.8)	**Statement about measures**
solvable	a risk neutral measure exists, so no arbitrage
uniquely solvable	it is unique, so the market is complete
unsolvable	an arbitrage, and the residual is the portfolio

The last row is constructive rather than merely a diagnosis. Let $\hat{\theta}$ minimise $\|(\mu - r\mathbf{1}) - \sigma\theta\|^2$ and write

$$\rho = (\mu - r\mathbf{1}) - \sigma\hat{\theta}$$

for the residual — the part of the excess-return vector orthogonal to the column space of σ, so $\sigma^\top\rho = 0$. Hold ρ as portfolio weights. Its loading on Brownian motion j is $\sum_i \rho_i\sigma_{ij} = (\sigma^\top\rho)_j = 0$, so the position carries no factor risk; its excess return is $\sum_i \rho_i(\mu_i - r) = \rho^\top(\mu - r\mathbf{1})$, and since $\mu - r\mathbf{1} = \sigma\hat{\theta} + \rho$ with $\sigma^\top\rho = 0$,

$$\rho^\top(\mu - r\mathbf{1}) = \rho^\top\sigma\hat{\theta} + \rho^\top\rho = \|\rho\|^2.$$

So ρ is a riskless portfolio earning $\|\rho\|^2$ over the funding rate — an arbitrage exactly when $\rho \neq 0$, which is exactly when (6.8) has no solution. The failure hands over the trade. `factor_risk_prices` builds it and checks both halves: the factor exposure is zero and the profit is $\|\rho\|^2$.

In a real market ρ is not riskless — it still carries whatever the factors do not span — so a non-zero residual is a *statistical* arbitrage rather than an actual one: a factor-neutral portfolio with a positive expected return. That is what a relative value book trades, and chapter 22 builds one explicitly, then measures whether a year of data can tell its return apart from zero.

This is also where chapter 8's term structure models get their θ. Every bond on the curve is driven by the same short rate, so every bond must reward that one risk at one rate, and forming a riskless combination of two maturities is the argument above with $n = 2$. The market price of interest rate risk is not an extra modelling assumption. It is (6.8) with one column.

Structure (The middle row: an incomplete market and a family of measures)**.** The table's middle row deserves more than a name. Solvable but not unique means (6.8) is consistent and underdetermined: the solutions are $\theta^* + \ker\sigma$ for any one particular solution θ^*, an affine subspace of dimension $d - \operatorname{rank}(\sigma)$, positive exactly when there are more Brownian motions than the assets can pin down.

Every θ in that subspace gives a valid measure $\mathbb{Q}_\theta$ for the n traded assets, and none is preferred over another by anything said so far. Shifting θ by a vector in $\ker\sigma$ changes no asset's drift at all, because only $\sigma\theta$ enters (6.8); it does change the drift of anything driven by a Brownian combination outside the row space of σ, which is to say anything exposed to a risk the n assets do not see. So

the traded market is silent between these measures, and a payoff exposed only to the row space of σ is priced the same under all of them, while a payoff exposed to $\ker\sigma$ is not — the family $\mathcal{Q} = \{\mathbb{Q}_\theta : \theta \in \theta^* + \ker\sigma\}$ gives it a whole range of prices, $\mathbb{E}^{\mathbb{Q}_\theta}[\text{payoff}]$ moving as θ moves.

Definition 6.4 (Super- and sub-replication)**.** For a payoff X paid at T, a *super-replicating* strategy is a self-financing portfolio in the traded assets whose time-T value dominates X almost surely; the *super-replication price* is the cheapest such strategy's cost today. A *sub-replicating* strategy is the mirror image, with time-T value at most X, and the *sub-replication price* is the most expensive such strategy's cost.

Theorem 6.5 (Every price in $\mathcal{Q}$ sits between the hedge prices)**.** *For every $\mathbb{Q}_\theta \in \mathcal{Q}$,*

$$\textit{sub-replication price} \;\le\; \mathbb{E}^{\mathbb{Q}_\theta}\big[D_T X\big] \;\le\; \textit{super-replication price}. \tag{6.9}$$

Proof. Let ϕ super-replicate X, with time-T value $V_T^\phi \ge X$ always. Being built only from the traded assets, ϕ's discounted wealth is a $\mathbb{Q}_\theta$-martingale for *every* θ solving (6.8) — that is what solving it means. So for every such θ,

$$D_0 V_0^\phi \;=\; \mathbb{E}^{\mathbb{Q}_\theta}\big[D_T V_T^\phi\big] \;\ge\; \mathbb{E}^{\mathbb{Q}_\theta}\big[D_T X\big],$$

using the martingale property for the equality and domination for the inequality. This holds for every super-replicating ϕ, so the cheapest one's cost, the super-replication price, is at least $\mathbb{E}^{\mathbb{Q}_\theta}[D_T X]$ for every θ. The mirror argument, run on a sub-replicating strategy, gives the other half. □

Remark (What the theorem does and does not say)**.** It bounds every consistent price by the two hedge costs; it does not say those costs are finite, and it does not say the bound is tight. Finiteness is the subject of the example below. Tightness — that the super-replication price equals $\sup_\theta \mathbb{E}^{\mathbb{Q}_\theta}[D_T X]$ exactly, not merely bounds it, so some strategy actually achieves the supremum — is a genuinely harder fact, proved in general by the optional decomposition theorem.[1]

Example 6.2 (A payoff the market cannot hedge at any price)**.** Take the simplest incomplete market: one asset driven by W_1 alone, and a second, independent Brownian motion W_2 that drives nothing tradable. Then $\sigma = (\sigma_1, 0)$, $\theta_1 = (\mu - r)/\sigma_1$ is pinned, and $\ker\sigma = \{(0, \theta_2) : \theta_2 \in \mathbb{R}\}$ is the free direction of the structure above. Under $\mathbb{Q}_\theta$, W_2 picks up drift $-\theta_2$, so $W_{2,T} \sim N(-\theta_2 T, T)$.

For the bounded payoff $X = \mathbf{1}\{W_{2,T} > 0\}$,

$$\mathbb{E}^{\mathbb{Q}_\theta}[X] \;=\; \Phi(-\theta_2\sqrt{T}) \;\in\; [0,1]$$

for every θ_2 — bounded for free, because an expectation can never exceed the supremum of what it averages, whatever the measure. For the unbounded payoff $X = e^{W_{2,T}}$,

$$\mathbb{E}^{\mathbb{Q}_\theta}[X] \;=\; \exp\big(-\theta_2 T + \tfrac{1}{2}T\big),$$

the moment generating function of a normal, and this diverges as $\theta_2 \to -\infty$.[2] By theorem 6.5 that means the super-replication price of $e^{W_{2,T}}$ is infinite: no finite position in the one traded asset can

[1] Kramkov (1996) in the semimartingale case, after El Karoui and Quenez (1995) for a Brownian filtration.

[2] both closed forms checked against a direct simulation under $\mathbb{Q}_\theta$; the digital, swept over four orders of magnitude in θ_2; the exponential, shown increasing without a ceiling.

ever dominate a payoff that grows without limit in a direction nothing traded is exposed to. That is the theorem correctly reporting that this claim cannot be hedged at any finite cost, not a failure of it. It is the same question chapter 14 asks of a moment strip before trusting a transform: an expectation written down formally is not a price until its finiteness has been checked, and the check is not automatic once the payoff is unbounded.

Remark (Two different reasons a market can be incomplete)**.** Frictions are often filed under the same heading as $\ker \sigma \neq \{0\}$, and one of them belongs there and two do not.

Trading only at finitely many dates is the same mechanism, discretised in time rather than in the number of assets: a position held between two dates produces a payoff linear in the state at the next one, which cannot replicate a general nonlinear function of it, for exactly the reason n assets cannot span $d > n$ Brownian motions. Chapter 5 measures the residual this leaves behind directly, as a hedging error rather than as a price interval.

Transaction costs and position limits are not this. Neither removes any randomness or reduces how many independent risk factors the traded assets see; a market that would be complete without them can become impossible to hedge exactly once they are added, because the frictionless replicating strategy still exists and is merely not admissible — proportional costs make continuous rebalancing cost infinitely much, since the Black-Scholes delta has infinite variation, and a position limit forbids whatever hedge would need more than it allows. The reachable-payoff set still shrinks, and a super/sub-replication interval still results, but $\ker \sigma$ has nothing to do with it: the relevant duality is a separate theory, for transaction costs Jouini and Kallal (1995) and Cvitanić and Karatzas (1996), for portfolio constraints its own constrained-hedging literature, none of it reducible to a solvability question about (6.8).

6.2 Numeraire Measures

Definition 6.6 (Numeraire)**.** A Numeraire is a strictly positive price process of a tradable relative to which prices of all other tradables are expressed.

A savings account which earns interest at the instantaneous interest rate can be taken as a numeraire. The value of the savings account at any point is given by

$$A_t = e^{\int_0^t r(t)dt} = 1/D(t) \tag{6.10}$$

where $D(t)$ is the discount factor.

The fact that discounted trade prices are martingales in risk-neutral measure can then also be stated as: tradable prices in savings account (or money market) numeraire are martingales.

Consider another numeraire N_t. Since the numeraire is itself a price process, $D_t N_t$ is a martingale and we can take it as a Radon-Nikodym derivative, with a suitable normalization so that $\mathbb{E}^{\mathbb{P}}[(d\mathbb{N}/d\mathbb{P})] = 1$, giving

$$\mathbb{E}^{\mathbb{N}}[\frac{X_T}{N_T}|\mathcal{F}_t] = \mathbb{E}^{\mathbb{P}}[\frac{(d\mathbb{N}/d\mathbb{P})_T}{(d\mathbb{N}/d\mathbb{P})_t}\frac{X_T}{N_T}|\mathcal{F}_t] = \mathbb{E}^{\mathbb{P}}[\frac{D_T N_T}{D_t N_t}\frac{X_T}{N_T}|\mathcal{F}_t] = \frac{1}{D_t N_t}\mathbb{E}^{\mathbb{P}}[D_T X_T|\mathcal{F}_t] = \frac{X_t}{N_t} \tag{6.11}$$

where $\mathbb{P}$ is the risk neutral measure and $\mathbb{N}$ is the measure corresponding to the numeraire N_t.

Remark (Prices in a numeraire are martingales in its measure)**.** The above equation shows that tradable prices with respect to a numeraire are martingales in the measure corresponding to the numeraire.

Risk neutral measure corresponds to the savings account (or money market) numeraire.

Definition 6.7 (T-forward measure)**.** If we take N_t to be the price of a riskless bond maturing at time T, the corresponding measure is the T-forward measure.

$$X_t/B(t,T) = \mathbb{E}^T[X_T/B(T,T)] = \mathbb{E}^T[X_T] \tag{6.12}$$

We see that the expectation of X_T in the T-forward measure gives the forward price $X_t/B(t,T)$ of the trade with expiration date T, hence the name T-forward measure. From the above equation we also notice that expiration T forward price process on an asset is martingale in T-forward measure.

Everything so far says which quantities are martingales under which measure, which is a statement about drifts being zero. The working question is usually the neighbouring one: given a process's drift under one numeraire, what is it under another. The answer follows from Girsanov and the density identified above, and it is short enough to record once and use thereafter — chapter 13 in particular does almost nothing else.

Theorem 6.8 (Change of numeraire, in drifts)**.** *Let A and B be numeraires, with $\mathbb{Q}^A$ and $\mathbb{Q}^B$ the corresponding measures, and let X be an Itô process. Then*

$$\operatorname{drift}^B(X) \;=\; \operatorname{drift}^A(X) \;+\; \frac{d}{dt}\Big\langle X,\, \ln \tfrac{B}{A}\Big\rangle_t. \tag{6.13}$$

In particular, for a strictly positive S the drift of $\ln S$ picks up $d\langle \ln S, \ln(B/A)\rangle$, and so does the relative drift in dS_t/S_t.

Proof. B is the price of a tradable and A is the numeraire, so B/A is a price expressed in units of A and is a $\mathbb{Q}^A$ martingale by the remark above. Normalising it,

$$Z_t \;=\; \frac{B_t/A_t}{B_0/A_0},$$

gives a strictly positive $\mathbb{Q}^A$ martingale starting at one, and it is exactly the density $(d\mathbb{Q}^B/d\mathbb{Q}^A)_t$ identified earlier in this section.

Being positive it can be written $dZ_t = Z_t\nu_t\, dW_t^A$, and Girsanov then says $W^B = W^A - \int \nu\, ds$ is a Brownian motion under $\mathbb{Q}^B$. So for any $dX_t = \mu_t^A\, dt + \sigma_t\, dW_t^A$, substituting $dW^A = dW^B + \nu\, dt$ gives

$$dX_t \;=\; \big(\mu_t^A + \sigma_t\nu_t\big)dt + \sigma_t\, dW_t^B.$$

It remains to recognise $\sigma\nu$. By Itô, $d\ln Z = \nu\, dW^A - \frac{1}{2}\nu^2 dt$, so $d\langle X, \ln Z\rangle = \sigma\nu\, dt$; and $\ln Z$ differs from $\ln(B/A)$ by a constant, which no covariation sees. That is (6.13).

For the last claim take $X = \ln S$. The relative drift follows because $dS/S = d\ln S + \frac{1}{2}d\langle \ln S\rangle$, and the quadratic variation term is pathwise and so the same under both measures. □

Remark (Reading the formula)**.** Three things stand out, since the formula gets used more often than it gets derived.

It depends on B/A alone. Neither numeraire matters on its own, only their ratio, which is why changing from A to B and back leaves every drift where it started.

A process uncorrelated with that ratio does not move at all. The adjustment is a covariation, so it vanishes exactly when the process and the numeraire ratio share no randomness, however volatile either happens to be.

And what it adjusts is a drift, never a volatility — the structure block at the head of this chapter, arriving as a formula.

Example 6.3 (Option Pricing)**.** A measure change earns its place in pricing by re-deriving Black-Scholes directly. For an option on a stock with expiry T and strike K, the valuation formula is

$$V_t = \frac{1}{D_t}\mathbb{E}^{\mathbb{P}}[D_T \max(S_T - K, 0)|\mathcal{F}_t]$$

where $\mathbb{P}$ is the risk neutral probability measure, D_t is the discount factor and S_t is the stock price which follows the Black Scholes diffusion

$$\frac{dS_t}{S_t} = rdt + \sigma dW_t$$

where r is the riskfree interest rate and W_t is a Brownian motion in $\mathbb{P}$.

Denote by $\mathbb{I}_{S_T>K}$ the indicator variable which takes value 1 is $S_T > K$ and 0 otherwise. We then have

$$\begin{aligned} V_t &= \frac{1}{D_t}\mathbb{E}^{\mathbb{P}}[D_T\mathbb{I}_{S_T>K}(S_T - K)|\mathcal{F}_t] \\ &= \frac{1}{D_t}\mathbb{E}^{\mathbb{P}}[D_T\mathbb{I}_{S_T>K}S_T|\mathcal{F}_t] - \frac{1}{D_t}\mathbb{E}^{\mathbb{P}}[D_T\mathbb{I}_{S_T>K}K|\mathcal{F}_t] \end{aligned}$$

Taking $\mathbb{S}$ to be the measure corresponding to stock price as numeraire, and taking $\mathbb{T}$ to be the measure corresponding to T-expiry bond as numeraire, we have

$$\begin{aligned} V_t &= \frac{1}{D_t}\mathbb{E}^{\mathbb{S}}[D_T\frac{(d\mathbb{P}/d\mathbb{S})_T}{(d\mathbb{P}/d\mathbb{S})_t}\mathbb{I}_{S_T>K}S_T|\mathcal{F}_t] - \frac{1}{D_t}\mathbb{E}^{\mathbb{T}}[D_T\frac{(d\mathbb{P}/d\mathbb{T})_T}{(d\mathbb{P}/d\mathbb{T})_t}\mathbb{I}_{S_T>K}K|\mathcal{F}_t] \\ &= \frac{1}{D_t}\mathbb{E}^{\mathbb{S}}[D_T\frac{(d\mathbb{S}/d\mathbb{P})_t}{(d\mathbb{S}/d\mathbb{P})_T}\mathbb{I}_{S_T>K}S_T|\mathcal{F}_t] - \frac{1}{D_t}\mathbb{E}^{\mathbb{T}}[D_T\frac{(d\mathbb{T}/d\mathbb{P})_t}{(d\mathbb{T}/d\mathbb{P})_T}\mathbb{I}_{S_T>K}K|\mathcal{F}_t] \\ &= \frac{1}{D_t}\mathbb{E}^{\mathbb{S}}[D_T\frac{D_tS_t}{D_TS_T}\mathbb{I}_{S_T>K}S_T|\mathcal{F}_t] - \frac{1}{D_t}\mathbb{E}^{\mathbb{T}}[D_T\frac{D_t(B(t,T))}{D_TB(T,T)}\mathbb{I}_{S_T>K}K|\mathcal{F}_t] \\ &= S_t\mathbb{E}^{\mathbb{S}}[\mathbb{I}_{S_T>K}|\mathcal{F}_t] - KB(t,T)\mathbb{E}^{\mathbb{T}}[\mathbb{I}_{S_T>K}|\mathcal{F}_t] \end{aligned}$$

where $B(t,T)$ is the bond price with the diffusion

$$dB(t,T)/B(t,T) = rdt + \sigma_2 dW_{2,t}$$

where $W_{2,t}$ is another Brownian motion in $\mathbb{P}$ such that $dW_t dW_{2,t} = \rho dt$. Giving the bond its own volatility goes beyond chapter 5's market, where the rate is constant and $B(t,T)$ has none; the

derivation below specialises to the ordinary Black-Scholes d_1, d_2 when $\sigma_2 = 0$, and otherwise prices the option under a genuinely stochastic discount factor.

Using Girsanov's theorem, we have

$$\begin{aligned} dW_t^{\mathbb{S}} &= dW_t^{\mathbb{P}} - \sigma dW_t^{\mathbb{P}} dW_t^{\mathbb{P}} &&= dW_t^{\mathbb{P}} - \sigma dt \\ dW_t^{\mathbb{T}} &= dW_t^{\mathbb{P}} - \sigma_2 dW_t^{\mathbb{P}} dW_{2,t}^{\mathbb{P}} &&= dW_t^{\mathbb{P}} - \sigma_2 \rho dt \end{aligned}$$

as Brownian motions in the measures corresponding to S, and B(t,T) numeraires respectively.

$$\frac{dS_t}{S_t} = rdt + \sigma(dW_t^{\mathbb{S}} + \sigma dt) = rdt + \sigma(dW_t^{\mathbb{T}} + \sigma_2 \rho dt) \tag{6.14}$$

where $dW_t^{\mathbb{S}} = dW_t - \sigma dt$ is a Brownian motion in $\mathbb{S}$ and $dW_t^{\mathbb{T}} = dW_t - \sigma_2 \rho dt$ is a Brownian motion in $\mathbb{T}$. Rearranging,

$$\frac{dS_t}{S_t} = (r + \sigma^2)dt + \sigma dW_t^{\mathbb{S}} = (r + \sigma\sigma_2\rho)dt + \sigma dW_t^{\mathbb{T}} \tag{6.15}$$

$$S_T = S_t e^{(r+\frac{1}{2}\sigma^2)\tau + \sigma W_\tau^{\mathbb{S}}} = S_t e^{\left(r+\sigma(\sigma_2\rho - \frac{1}{2}\sigma)\right)\tau + \sigma W_\tau^{\mathbb{T}}} \tag{6.16}$$

where $\tau = T - t$.

$$\begin{aligned} \mathbb{E}^{\mathbb{S}}[\mathbb{I}_{S_T > K} | \mathcal{F}_t] &= \mathbb{E}^{\mathbb{S}}[\mathbb{I}_{S_t e^{(r+\frac{1}{2}\sigma^2)\tau + \sigma W_\tau^{\mathbb{S}}} > K} | \mathcal{F}_t] \\ &= \mathbb{E}^{\mathbb{S}}[\mathbb{I}_{(r+\frac{1}{2}\sigma^2)\tau + \sigma W_\tau^{\mathbb{S}} > \log(K/S_t)} | \mathcal{F}_t] \\ &= \mathbb{E}^{\mathbb{S}}[\mathbb{I}_{\sigma W_\tau^{\mathbb{S}} > \log(K/S_t) - (r+\frac{1}{2}\sigma^2)\tau} | \mathcal{F}_t] \\ &= \mathbb{E}^{\mathbb{S}}[\mathbb{I}_{\frac{1}{\sqrt{\tau}} W_\tau^{\mathbb{S}} > \frac{1}{\sigma\sqrt{\tau}}(\log(K/S_t) - (r+\frac{1}{2}\sigma^2)\tau)} | \mathcal{F}_t] \\ &= N(-\frac{1}{\sigma\sqrt{\tau}}(\log(K/S_t) - (r + \frac{1}{2}\sigma^2)\tau)) \\ &= N(d_1) \end{aligned}$$

where N is the cumulative normal distribution, and $d_1 = \frac{1}{\sigma\sqrt{\tau}}(\log(S_t/K) + (r + \frac{1}{2}\sigma^2)\tau)$.

Similarly we have,

$$\begin{aligned} \mathbb{E}^{\mathbb{T}}[\mathbb{I}_{S_T > K} | \mathcal{F}_t] &= \mathbb{E}^{\mathbb{T}}[\mathbb{I}_{S_t e^{\left(r+\sigma(\sigma_2\rho - \frac{1}{2}\sigma)\right)\tau + \sigma W_\tau^{\mathbb{T}}} > K} | \mathcal{F}_t] \\ &= \mathbb{E}^{\mathbb{T}}[\mathbb{I}_{\left(r+\sigma(\sigma_2\rho - \frac{1}{2}\sigma)\right)\tau + \sigma W_\tau^{\mathbb{T}} > \log(K/S_t)} | \mathcal{F}_t] \\ &= \mathbb{E}^{\mathbb{T}}[\mathbb{I}_{\sigma W_\tau^{\mathbb{T}} > \log(K/S_t) - \left(r+\sigma(\sigma_2\rho - \frac{1}{2}\sigma)\right)\tau} | \mathcal{F}_t] \\ &= \mathbb{E}^{\mathbb{T}}[\mathbb{I}_{\frac{1}{\sqrt{\tau}} W_\tau^{\mathbb{T}} > \frac{1}{\sigma\sqrt{\tau}}\left(\log(K/S_t) - \left(r+\sigma(\sigma_2\rho - \frac{1}{2}\sigma)\right)\tau\right)} | \mathcal{F}_t] \\ &= N\left(-\frac{1}{\sigma\sqrt{\tau}}\left(\log(K/S_t) - \left(r + \sigma(\sigma_2\rho - \frac{1}{2}\sigma)\right)\tau\right)\right) \\ &= N(d_2) \end{aligned}$$

where N is the cumulative normal distribution, and $d_2 = \frac{1}{\sigma\sqrt{\tau}}(\log(S_t/K) + (r + \sigma\sigma_2\rho - \frac{1}{2}\sigma^2)\tau)$.

And finally we have,

$$\begin{aligned} V_t &= S_t N(d_1) - KB(t,T)N(d_2) \\ &= S_t N(d + \sigma\sqrt{\tau}) - KB(t,T)N(d + \sigma_2\rho\sqrt{\tau}) \end{aligned}$$

where $d = \frac{1}{\sigma\sqrt{\tau}}(\log(S_t/K) + (r - \frac{1}{2}\sigma^2)\tau)$.

Note that $\sigma\, dt$ and $\sigma_2\rho\, dt$ are the drift correction terms we get to stock price Brownian motion using Girsanov's theorem to change measure to stock and bond numeraires respectively.

Remark (The two terms are one event counted twice)**.** The derivation above deserves one more sentence: it explains something about the Black formula that is usually left as a curiosity.

$N(d_1)$ and $N(d_2)$ are both the probability of $\{S_T > K\}$ — the same event, the option finishing in the money. They are different numbers because they are computed under different measures: $N(d_1)$ under the measure in which the stock is the numeraire, $N(d_2)$ under the T-forward measure. Neither is "the" probability of exercise, and asking which one is misses that the question has no answer until a numeraire is named.

The gap between them is the measure change, and it is $\sigma\sqrt{\tau}$ worth of shift in the argument of N. It is therefore largest for long dated and volatile options and vanishes as $\sigma\sqrt{\tau} \to 0$, where there is no randomness for the reweighting to act on and the two measures agree.

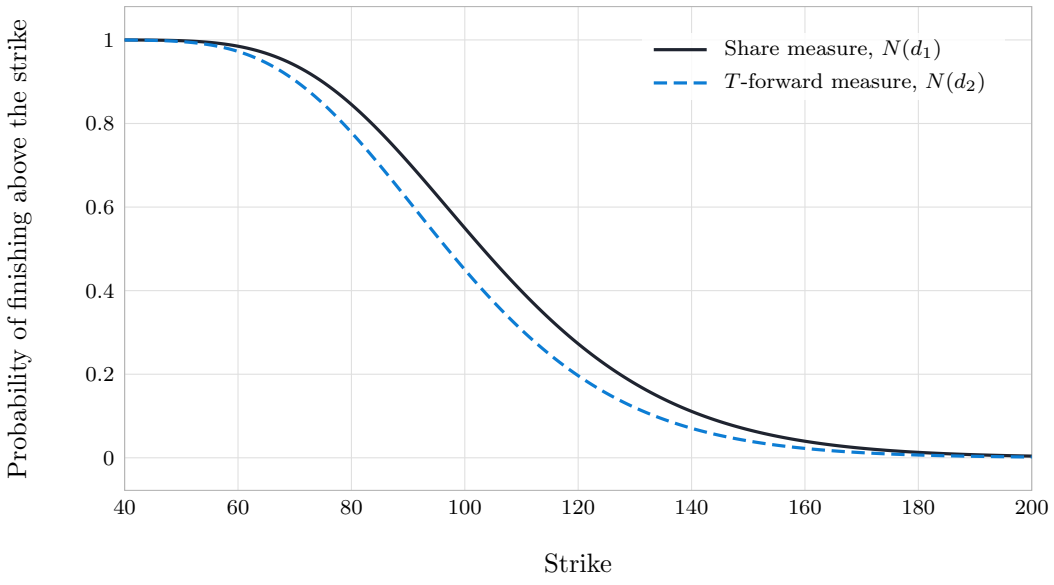

Figure 6.2: The probability that a one year option finishes in the money, at a 25% volatility, under the two measures the Black formula is assembled from. The share measure weights each path by the terminal price itself, so it counts the paths finishing high for more — and those are exactly the paths on which the option is exercised. It therefore assigns the higher probability at every strike. Both curves describe the same event; the market has not been consulted twice.
Drawn by `exercise_probabilities`.

Structure (The pricing kernel)**.** A change to the risk neutral measure does two things at once, reweights the paths and discounts along the way, and the split between them depends on which numeraire was chosen. Their product does not. Write D_t for the discount factor of the savings account and $L_t = (d\mathbb{Q}/d\mathbb{P})_t$ for the density of the risk neutral measure against the real world one, and set

$$M_t \;=\; D_t\, L_t. \tag{6.17}$$

This is the *pricing kernel*, or stochastic discount factor. It is the single random variable that turns a real-world expectation into a price.

Every price is a real-world expectation weighted by M. For a payoff X_T at T, the abstract Bayes formula of this chapter gives

$$V_t \;=\; \mathbb{E}^{\mathbb{Q}}\Big[\tfrac{D_T}{D_t} X_T \,\Big|\, \mathcal{F}_t\Big] \;=\; \mathbb{E}^{\mathbb{P}}\Big[\tfrac{L_T D_T}{L_t D_t} X_T \,\Big|\, \mathcal{F}_t\Big] \;=\; \frac{1}{M_t}\,\mathbb{E}^{\mathbb{P}}\big[M_T X_T \,\big|\, \mathcal{F}_t\big].$$

Equivalently $M_t V_t$ is a $\mathbb{P}$-martingale for every tradable V. There is no discounting and no reweighting left over, and nothing in the formula mentions a numeraire.

Every numeraire measure is the kernel weighted by the numeraire. If N is a tradable, MN is a strictly positive $\mathbb{P}$-martingale, so $(d\mathbb{Q}^N/d\mathbb{P})_t = M_t N_t/(M_0 N_0)$ is a density, and

$$\mathbb{E}^N\Big[\tfrac{X_T}{N_T} \,\Big|\, \mathcal{F}_t\Big] \;=\; \mathbb{E}^{\mathbb{P}}\Big[\tfrac{M_T N_T}{M_t N_t}\tfrac{X_T}{N_T} \,\Big|\, \mathcal{F}_t\Big] \;=\; \frac{V_t}{N_t},$$

which is the statement that prices in a numeraire are martingales in its measure. The savings account gives the risk neutral measure, the bond maturing at T gives the T-forward measure, and the stock gives the share measure of the Black formula. Choosing a numeraire only chooses how to divide the same kernel into a discount and a density.

The price of risk is the kernel's loading. In one factor, $dL/L = -\theta\, dW$ and D decays at the short rate, so $dM/M = -r\, dt - \theta\, dW$. For an asset $dV/V = \mu\, dt + \sigma\, dW$, requiring MV to have no drift under $\mathbb{P}$ gives $\mu - r - \sigma\theta = 0$, which is (6.6) read off the kernel: the excess return is minus the covariance of the asset with the kernel,

$$\mu - r \;=\; -\frac{1}{dt}\,\mathrm{Cov}\Big(\frac{dV}{V}, \frac{dM}{M}\Big).$$

An asset that is high where M is high pays in the states the kernel weights most, so it is a hedge: it is priced above its expected value and earns less than r. An asset that is high where M is low, like the stock, earns a premium.

A forward price is an expectation plus a covariance. At time zero, with $M_0 = 1$ and $P(0,T) = \mathbb{E}^{\mathbb{P}}[M_T]$,

$$F \;=\; \frac{\mathbb{E}^{\mathbb{P}}[M_T X_T]}{\mathbb{E}^{\mathbb{P}}[M_T]} \;=\; \mathbb{E}^{\mathbb{P}}[X_T] + \frac{\mathrm{Cov}^{\mathbb{P}}(M_T, X_T)}{\mathbb{E}^{\mathbb{P}}[M_T]}. \tag{6.18}$$

The forward is the real-world expectation shifted by the covariance of the payoff with the kernel. For the stock in chapter 5, which is high where M is low, the shift is negative, and the forward sits below the real-world mean by exactly $S_0(e^{\mu T} - e^{rT})$. Nothing about the size or sign of the shift is a fact about the payoff alone; it is a fact about how the payoff lines up with the kernel.

The kernel is the same object economics calls the marginal utility of wealth: a representative investor's M is high in bad states, which is why an asset that pays there is worth more than its mean. That reading is an interpretation and the notes use nothing from it. Everything above follows from the existence of $\mathbb{Q}$.[3]

References

- Girsanov, I. V. (1960). On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory of Probability and its Applications*, 5(3), 285–301.

- Geman, H., El Karoui, N., & Rochet, J.-C. (1995). Changes of numeraire, changes of probability measure and option pricing. *Journal of Applied Probability*, 32(2), 443–458.

- Jamshidian, F. (1997). LIBOR and swap market models and measures. *Finance and Stochastics*, 1(4), 293–330.

- El Karoui, N., & Quenez, M.-C. (1995). Dynamic programming and pricing of contingent claims in an incomplete market. *SIAM Journal on Control and Optimization*, 33(1), 29–66.

- Kramkov, D. O. (1996). Optional decomposition of supermartingales and hedging contingent claims in incomplete security markets. *Probability Theory and Related Fields*, 105(4), 459–479.

- Jouini, E., & Kallal, H. (1995). Martingales and arbitrage in securities markets with transaction costs. *Journal of Economic Theory*, 66(1), 178–197.

- Cvitanić, J., & Karatzas, I. (1996). Hedging and portfolio optimization under transaction costs: a martingale approach. *Mathematical Finance*, 6(2), 133–165.

[3] the kernel prices the stock, the bond and the call from the real world, the call against the Black-Scholes formula; the numeraire measures' densities; the covariance identity for the stock.

Part II

Classical Models

Chapter 7

Curve Construction and Linear Products

We build the interest rate curve out of the instruments quoted in the market, and price the products whose value depends on the curve alone and not on its volatility. This is the layer underneath every model in the chapters that follow: a term structure model is calibrated to a curve, and a curve is not handed down, it is constructed.

7.1 Discount Factors, Zero Rates and Forward Rates

The single object underneath all of fixed income is the discount factor.

Definition 7.1 (Discount factor)**.** $P(t,T)$ is the value at time t of a contract paying \$1 with certainty at time T. It is the price at t of a zero coupon bond maturing at T, and $P(T,T)=1$.

Everything else in this chapter is a re-expression of the function $T \mapsto P(0,T)$. The reason interest rates are usually harder to model than equities is because the price of a stock today is a number, whereas the price of borrowing today is a whole curve, one number for every maturity. A model of interest rates has to move a curve, and the constraints linking the points of that curve to each other are what make the subject what it is.

The first such constraint is the one that lets us talk about rates at all.

Definition 7.2 (Simply compounded forward rate)**.** For $T_1 < T_2$ with year fraction $\tau = \tau(T_1,T_2)$, the forward rate $F(t;T_1,T_2)$ is the interest rate, agreed at t, at which one may borrow over the future period $[T_1,T_2]$.

The forward rate is not free to be anything. It is pinned by today's discount factors, and it should be derived rather than quoted, because the derivation is the pattern every no-arbitrage statement in this chapter follows.

Calculation 7.3 (The forward rate from the curve)**.** Consider the following portfolio, set up at time 0 at no cost. Buy one zero coupon bond maturing at T_1, and sell $P(0,T_1)/P(0,T_2)$ zero coupon bonds maturing at T_2. The proceeds of the sale exactly fund the purchase, so

$$\pi_0 = P(0,T_1) - \frac{P(0,T_1)}{P(0,T_2)} \cdot P(0,T_2) = 0.$$

At T_1 the long bond pays \$1. At T_2 the short position costs $P(0,T_1)/P(0,T_2)$. So the portfolio is exactly a contract to receive \$1 at T_1 and repay $P(0,T_1)/P(0,T_2)$ at T_2: it is a loan over $[T_1,T_2]$, entered into today at no cost. Its simple interest rate F satisfies $1+\tau F = P(0,T_1)/P(0,T_2)$, giving

$$F(0;T_1,T_2) = \frac{1}{\tau}\left(\frac{P(0,T_1)}{P(0,T_2)} - 1\right). \tag{7.1}$$

Any other rate for the same period admits an arbitrage: if the market quoted $F' > F$, lend at F' and fund it with this portfolio, and the difference is locked in at T_2 with no position of any kind held in between.

Note what the argument used and what it did not. It used only that the two bonds trade, and it produced a rate for a period entirely in the future without any model of how rates move. This is the sense in which linear products are "just discounting": their prices are consequences of today's curve, not of any assumption about its dynamics.

The same quantity gets quoted in several conventions, and the conversions between them account for a surprising share of the discrepancies between two people who think they are pricing the same trade.

Definition 7.4 (Zero rate)**.** The continuously compounded zero rate, or spot rate, for maturity T is the single rate $z(T)$ that reproduces the discount factor:

$$P(0,T) = e^{-z(T)T} \qquad \Longleftrightarrow \qquad z(T) = -\frac{\ln P(0,T)}{T}.$$

Definition 7.5 (Instantaneous forward rate)**.** The instantaneous forward rate $f(t,T)$ is the rate for borrowing over $[T, T+dT]$, quoted at t: entering at t into a contract to borrow from T to $T+dT$ charges interest $f(t,T)dT$.

Remark. Taking $T_1 = T$, $T_2 = T+\delta$ in (7.1) and letting $\delta \to 0$,

$$f(t,T) = \lim_{\delta\to 0} \frac{1}{\delta}\left(\frac{P(t,T)}{P(t,T+\delta)} - 1\right) = -\frac{\partial}{\partial T}\ln P(t,T),$$

and integrating from t to T using $P(t,t) = 1$,

$$P(t,T) = e^{-\int_t^T f(t,s)\,ds}. \tag{7.2}$$

So the discount curve, the zero curve and the instantaneous forward curve carry exactly the same information. Which of them one works in is a matter of what one wants to be smooth — a point we return to when we interpolate.

Definition 7.6 (Short rate)**.** The short rate is the instantaneous forward rate for immediate borrowing, $r(t) = f(t,t)$.

7.2 The Year Fraction Is Not a Length of Time

Everything above was written with a year fraction $\tau(T_1, T_2)$ in it, and the natural reading of that symbol — elapsed time — is wrong. It is whatever the contract says it is. This section is about the arithmetic that follows from that.

Definition 7.7 (Day count convention)**.** A day count convention is a rule assigning a year fraction to a pair of dates. The three in common use are

- *Act/360*: actual days elapsed, divided by 360. The money market convention, used for USD and EUR floating legs.
- *Act/365*: actual days over 365, used in sterling and several Asian markets.
- *30/360*: every month is treated as thirty days and every year as 360, so a semiannual period is exactly one half. Used for US corporate and agency bonds and for USD swap fixed legs.

A single swap routinely uses two of them, one per leg.

Calculation 7.8 (What the conventions are worth)**.** Take the semiannual period from 15 January to 15 July 2026, which is 181 actual days.[1]

Convention	**Year fraction**	**Interest at** 5%
Act/360	0.502778	2.5139%
Act/365	0.495890	2.4795%
30/360	0.500000	2.5000%

Act/360 against 30/360 is a gap of 1.39 basis points on one period. A swap's fixed leg quoted on the wrong basis is wrong by that on every payment, compounding to something no bid-offer spread absorbs. Act/360 is systematically the largest.

The same arithmetic makes a leap day worth 1.39 basis points on an annual Act/360 period, for the identical reason — both numbers are one day out of 360. Which is the useful way to remember the size of all of this: a day is roughly a basis point and a half at these levels, and every convention question is a question about how many days.

Two further conventions are needed before an actual cashflow can be computed, and both are about dates rather than fractions.

Definition 7.9 (Settlement lag)**.** A trade agreed today does not settle today. A spot-starting swap typically begins two business days forward, a bond purchase settles one or two business days after the trade, and a deposit rate quoted today applies from spot. So the curve built from today's quotes has its origin at spot rather than at today, and a discount factor to a date is a discount factor from settlement.

The lag is small and it does not cancel. A two day gap at 5% is a basis point and a half of value on the whole notional, so a curve whose origin is misplaced by two days misprices every instrument on it by about that, in the same direction, in a way that no repricing check will reveal because the same error is in both the build and the valuation.

[1] `curve::DayCount`, with the year fractions and both gaps below pinned.

Definition 7.10 (Accrued interest, clean and dirty price)**.** A bond that has run part way through a coupon period has earned interest that has not yet been paid. With coupon rate c and τ the year fraction since the last payment, on the bond's own convention,

$$\text{accrued} = 100\, c\, \tau,$$

and the two prices are

$$\text{dirty} = \text{clean} + \text{accrued}.$$

The *dirty* price is what changes hands. The *clean* price is what is quoted.

Remark (Why the quote is the one that is not paid)**.** The dirty price of a bond sawtooths: it climbs by the accrual through each period and drops by the coupon on payment day, so a chart of it looks like the bond is repeatedly falling for reasons unconnected to rates. Removing the accrual removes the sawtooth, and what is left moves only when the market moves. The convention exists to make the quoted series interpretable.

It also produces a specific, common error. A position marked at the clean price is undervalued by up to a full coupon, and a hedge ratio computed from clean prices on both legs is right, while one computed from a clean price on one leg and a dirty on the other is wrong by the accrual — which is a number that grows steadily and then jumps, so the error looks like a slow drift in the P&L punctuated by a step. Chapter 24's explain report is where that pattern shows up, and recognising it as an accrual convention rather than a model failure saves a great deal of time.

7.3 The Instruments

The curve is built from the instruments the market quotes, so we need those first. All of them are linear in the sense that matters here: their payoffs are fixed multiples of rates or of \$1, with no optionality, so their prices are determined by today's curve.

Definition 7.11 (Deposit)**.** A deposit lends \$1 from T_1 to T_2 at a simply compounded rate R, returning $1 + \tau R$ at T_2. Its value today is $P(0,T_2)(1+\tau R) - P(0,T_1)$, and it is at par when $R = F(0;T_1,T_2)$.

Definition 7.12 (Reference rate)**.** A reference rate $L(t,T,T+\delta)$ is the market's published rate for borrowing over $[T,T+\delta]$, as observed at t. Historically these were IBOR rates — panel estimates of unsecured interbank borrowing costs, of which LIBOR was the largest — and they are now overnight risk free rates such as SOFR, compounded over the period.

In a single curve world the reference rate is the curve's own forward rate,

$$L(t,T,T+\delta) = \frac{1}{\delta}\left(\frac{P(t,T)}{P(t,T+\delta)} - 1\right),$$

which is the assumption we make until the last section of this chapter, where we take it apart.

Definition 7.13 (Forward rate agreement)**.** A δ FRA maturing at T pays at T

$$\frac{\delta\big(L(T,T,T+\delta) - R\big)}{1 + \delta L(T,T,T+\delta)},$$

where R is the rate agreed in the contract. The forward rate is the R making this contract worth zero.

The peculiar looking discounting inside the payoff is the market convention of settling at the start of the period rather than the end: the natural payment $\delta(L - R)$ would fall at $T + \delta$, and dividing by $1 + \delta L$ moves it back to T at the rate that has just fixed. Chapter 8's Future-Forward Basis section shows that the fair R is the forward rate under the appropriate measure, and that the corresponding futures contract is not, which is the first place a model becomes necessary.

Interest rate swaps

Definition 7.14 (Swap)**.** A fixed-for-floating interest rate swap exchanges a stream of fixed coupons for a stream of floating coupons on the same notional. The payer of the swap pays fixed and receives floating; the receiver does the opposite. The two legs need not share a payment frequency.

Take the notional to be 1 throughout; prices scale with it. Let the fixed leg pay at dates $\tau_1^r < \cdots < \tau_{N_r}^r$ with accruals δ_i^r, and the floating leg at $\tau_1^p < \cdots < \tau_{N_p}^p$ with accruals δ_j^p, both legs starting at τ_0. The fixed rate is R.

The fixed leg is a known set of cashflows, so it prices by discounting:

$$\text{Fixed}(0) = \Sigma_{i=1}^{N_r} \delta_i^r R\, P(0, \tau_i^r) = R\, \text{Ann}(0),$$

where we have named the sum.

Definition 7.15 (Annuity)**.** The annuity, or level, or PV01, is the value of \$1 paid on each fixed leg date, weighted by the accruals:

$$\text{Ann}(t) = \Sigma_{i=1}^{N_r} \delta_i^r P(t, \tau_i^r).$$

The annuity will do a great deal of work later. It is a price — it is the price of a portfolio of zero coupon bonds — and it is strictly positive, so by chapter 6 it is a numeraire, and it is the numeraire in which swap rates are martingales.

The floating leg takes one more step. Each coupon pays the reference rate fixed at the start of its period, so

Calculation 7.16 (The floating leg telescopes)**.**

$$\begin{aligned}
\text{Float}(0) &= \Sigma_{j=1}^{N_p} \delta_j^p\, L(0, \tau_{j-1}^p, \tau_j^p)\, P(0, \tau_j^p) \\
&= \Sigma_{j=1}^{N_p} \delta_j^p \cdot \frac{1}{\delta_j^p} \left(\frac{P(0, \tau_{j-1}^p)}{P(0, \tau_j^p)} - 1 \right) P(0, \tau_j^p) \\
&= \Sigma_{j=1}^{N_p} \big(P(0, \tau_{j-1}^p) - P(0, \tau_j^p) \big) \\
&= P(0, \tau_0) - P(0, \tau_{N_p}^p).
\end{aligned}$$

Every intermediate term cancels, and a floating leg is worth notional in at the start minus notional out at the end. There is a reason beyond the algebra: a floating rate note that pays the market rate on its own notional is worth par at each reset, because at the reset date the coupon is exactly the rate the market would charge for the next period. Note carefully that the cancellation used L being the curve's forward rate — the same curve that produced the discount factors. In the multi-curve world it is false.

Definition 7.17 (Par swap rate)**.** The par swap rate S is the fixed rate making the swap worth zero.

Setting Fixed = Float,

$$S(t) = \frac{\text{Float}(t)}{\text{Ann}(t)} = \frac{P(t,\tau_0) - P(t,\tau^p_{N_p})}{\Sigma_i \delta^r_i P(t,\tau^r_i)}, \tag{7.3}$$

The numerator is a floating leg of many payments collapsed to a difference of two discount factors, and the collapse is exact rather than approximate: each payment's forward is computed from the same curve that discounts it, so every term's numerator cancels the next term's denominator. It also does not require the two legs to share a frequency, which is why one formula covers a swap paying quarterly floating against annual fixed. and the value of a receiver swap struck at R is $\text{Ann}(0)(R - S(0))$, of a payer swap $\text{Ann}(0)(S(0) - R)$. Two facts fall straight out of this and both matter later. First, the swap rate is a ratio of a price to the annuity, so it is a martingale in the annuity measure — this is the observation that makes chapter 8's swaptions tractable. Second, the sensitivity of a swap to a parallel shift in the fixed rate is the annuity, so the annuity is the swap's risk as well as its numeraire.

Bonds

Definition 7.18 (Bond)**.** A bond pays a specified notional at maturity and periodic coupons until then. A zero coupon bond pays only the notional.

A bond is a set of known cashflows c_i at dates t_i, so its value from the curve is $\Sigma_i c_i P(0,t_i)$. This is the *dirty* or invoice price, and is what changes hands. What is quoted is the *clean* price, the dirty price less the accrued interest earned since the last coupon.

The market also quotes a bond by its yield.

Definition 7.19 (Yield to maturity)**.** The yield y of a bond is the single rate that reprices it:

$$P_{\text{dirty}} = \Sigma_i \frac{c_i}{(1 + y/m)^{m t_i}}, \tag{7.4}$$

with m the coupon frequency. Note which price appears: the yield is defined against the dirty price, since it is the amount actually paid that the discounted cashflows have to reproduce.

The yield is a summary statistic and not a valuation model. It collapses the entire curve onto one number, so two bonds of different maturities have yields that are not comparable in any useful way, and this is exactly why spreads to a curve exist. It is however the natural coordinate for risk.

Definition 7.20 (Duration, DV01 and convexity)**.**

$$D_{\text{mod}} = -\frac{1}{P}\frac{\partial P}{\partial y}, \qquad \text{DV01} = -\frac{\partial P}{\partial y} \times 10^{-4}, \qquad C = \frac{1}{P}\frac{\partial^2 P}{\partial y^2},$$

so that $\Delta P \approx -D_{\text{mod}} P\,\Delta y + \frac{1}{2} C P (\Delta y)^2$.

Exercise (Duration of a zero coupon bond)**.** Show that a zero coupon bond maturing at T, priced with continuous compounding, has $D_{\text{mod}} = T$, and that a coupon bond's duration is the coupon-weighted average of its cashflow times. Deduce that duration is bounded above by maturity, with equality only for a zero coupon bond.

Remark (Two different DV01s, and they are not equal)**.** The definition above differentiates with respect to y, the bond's own yield to maturity, and that is one of two things the same name is used for.

A *yield* DV01 shifts the single number y and reprices with the bond's own discounting. A *curve* DV01 shifts the zero curve — every point of it, by a basis point — rebuilds the discount factors and reprices. They differ: y is defined by (7.4) as the flat rate reproducing the bond's price, so a bond whose cashflows sit on a sloped curve has a y that is a cashflow-weighted average of the curve, and shifting the curve by a basis point does not shift that average by exactly a basis point unless the weights are uniform. The gap grows with the coupon and with the slope, and it is zero for a zero coupon bond, where the two definitions coincide.

Which to use is decided by what the hedge is. A position hedged with another bond wants yield DV01 on both, because the ratio is what is traded. A position hedged with swaps, or held in a book whose risk is bucketed against curve instruments, wants curve DV01, because that is the sensitivity the hedging instruments actually deliver. Reporting one and hedging with the other leaves a residual that looks like a small unexplained P&L which grows when the curve steepens — and the curve risk section below is where that becomes systematic.

Convexity is subject to the same distinction and matters for a further reason. It is why the DV01 itself moves as rates move, so a hedge set today is wrong tomorrow by an amount proportional to C and to the size of the move. That is a rehedging cost rather than a mark, and it is the linear-product version of the gamma cost in chapter 5's calculation The daily profit and loss of a hedged option.

Exercise (A swap's DV01 is its annuity)**.** Show from (7.3) that the change in value of a payer swap for a one basis point rise in the par swap rate is $\text{Ann}(0) \times 10^{-4}$. This is why traders quote swap risk in annuities and why the annuity is called the PV01.

7.4 Bootstrapping the Curve

We now have the two halves of the problem. The instruments above price off the curve; the market quotes their prices. Curve construction is the inverse problem: find the curve that reprices the quotes.

The instruments are chosen by liquidity, and different parts of the curve are pinned by different ones. The front end comes from overnight and short deposits and from overnight indexed swaps.

The belly comes from short term interest rate futures, or from FRAs, which fix forward rates out to two or three years. Beyond that the quotes are par swap rates, out to thirty years and more.

The mental model is to solve outward. The shortest instrument involves only the nearest discount factor, so it determines it. The next instrument, given everything shorter, involves one new discount factor, so it determines that. And so on.

Example 7.1 (Bootstrapping by hand)**.** Suppose the market quotes a one year deposit at 4.00%, a two year annual par swap at 3.80%, and a three year annual par swap at 3.70%, all with $\tau = 1$ and annual fixed payments.

The deposit gives the first discount factor directly:

$$P(0,1) = \frac{1}{1 + 0.0400} = 0.961538.$$

The two year swap is at par, so by (7.3) with $S = 0.0380$,

$$\begin{aligned} 1 - P(0,2) &= 0.0380\,\big(P(0,1) + P(0,2)\big) \\ P(0,2)\big(1 + 0.0380\big) &= 1 - 0.0380 \cdot 0.961538 \\ P(0,2) &= \frac{1 - 0.036538}{1.0380} = 0.928190. \end{aligned}$$

The three year swap then gives, with $S = 0.0370$,

$$\begin{aligned} 1 - P(0,3) &= 0.0370\,\big(P(0,1) + P(0,2) + P(0,3)\big) \\ P(0,3) &= \frac{1 - 0.0370 \cdot (0.961538 + 0.928190)}{1.0370} = 0.896895. \end{aligned}$$

Reading off the continuously compounded zero rates and the simply compounded one year forward rates between the nodes,

	1y	2y	3y
$z(T)$	3.922%	3.726%	3.627%
forward over the year to T	4.000%	3.593%	3.489%

Notice how much further the forwards travel. Over three years the zero rate falls by 30 basis points and the forward rate by 51, nearly twice as far. This is not an accident of these numbers. The zero rate is an average of the forwards up to T, and an average always moves less than the thing being averaged, so a curve that slopes gently in zero space slopes sharply in forward space. Since the forwards are what a product actually fixes on, and the zeros are what one is tempted to draw and interpolate, this is a warning about what comes next.

Exercise (Forwards as the derivative of the zero curve)**.** Show that $f(0,T) = z(T) + Tz'(T)$, and deduce that a zero curve with a maximum at T^* has $f(0,T^*) = z(T^*)$, with the forward curve crossing the zero curve there. Confirm on the example above that a falling zero curve forces the forwards below it.

In practice one does not literally peel the instruments off one at a time. The instruments overlap, their dates do not line up, and the front end has effects a strict ordering cannot express. What is done instead is to set up a global solve: choose a set of curve nodes and an interpolation scheme, and find the node values that reprice every quoted instrument simultaneously. The bootstrap above is the special case in which the system happens to be triangular.

Interpolation is a modelling choice

Whatever the instruments are, there are finitely many of them and infinitely many dates, so the curve between the nodes has to be interpolated, and the choice is not innocent.

Suppose we interpolate linearly in the continuously compounded zero rate. Then on a segment between nodes at T_a and T_b,

$$z(T) = z(T_a) + \frac{T - T_a}{T_b - T_a}\big(z(T_b) - z(T_a)\big),$$

and since $\ln P(0,T) = -z(T)T$, the instantaneous forward rate is

$$\begin{aligned} f(0,T) &= -\frac{\partial}{\partial T}\ln P(0,T) = \frac{\partial}{\partial T}\big(z(T)T\big) = z(T) + Tz'(T) \\ &= z(T) + T\cdot\frac{z(T_b) - z(T_a)}{T_b - T_a}. \end{aligned}$$

This is piecewise linear in T but *discontinuous at every node*, because z' jumps there while T does not. So a perfectly reasonable looking zero curve, smooth to the eye, implies a forward curve with a jump at every instrument in the bootstrap. Every product that fixes between nodes prices off those forwards.

The general lesson is the one already visible in the worked example: the forward rate is a derivative of the curve, so it is one degree less smooth than whatever we chose to interpolate. Interpolating linearly in $\ln P$ makes f piecewise constant, hence a step function. Interpolating in the instantaneous forwards themselves makes them as smooth as we like but gives up exact locality — moving one quote moves the curve everywhere. Monotone convex schemes, which the next remark describes, are built to keep forwards continuous while preventing the overshoot a naive interpolation through the forwards produces.

Remark (What a monotone convex scheme is)**.** It is Hagan and West's scheme, and it is the standard answer to the failure the figure below shows.

Start from what the bootstrap actually determines. It does not pin the instantaneous forward at any single date; it pins the *average* forward across each bucket between adjacent nodes, since that average is what discounts the cashflows in that bucket. Call those the discrete forwards f_i^d on $[T_{i-1}, T_i]$. Any interpolation that reprices the quotes must satisfy

$$\int_{T_{i-1}}^{T_i} f(t)\,dt = f_i^d\,(T_i - T_{i-1})$$

for every bucket — the area under the forward curve over each bucket is fixed, however the curve is shaped inside it. A scheme with this property is called *conserving*, and it is the reason the construction reprices by design rather than by iteration.

The scheme then proceeds in two passes:

Input: the discrete forwards f_i^d on each bucket $[T_{i-1}, T_i]$, as the bootstrap left them.

1. *Node values.* For each interior node T_i, interpolate between the two buckets that meet there,

$$f_i = \frac{T_i - T_{i-1}}{T_{i+1} - T_{i-1}} f^d_{i+1} + \frac{T_{i+1} - T_i}{T_{i+1} - T_{i-1}} f^d_i,$$

   weighting each bucket by the length of the *other* one, so the nearer bucket counts for more. At the two ends there is only one bucket, and f_0 and f_N are collared to it.

2. *A quadratic on each bucket.* On $[T_{i-1}, T_i]$ fit the quadratic through f_{i-1} and f_i whose integral is $f^d_i(T_i - T_{i-1})$. Three conditions, three coefficients, one answer — and the area condition is what makes the scheme conserving, so the quotes are repriced by construction rather than by a solver.

3. *Check the range.* The quadratic overshoots when a node value sits too far from the bucket average it belongs to. Written in terms of

$$g_{i-1} = f_{i-1} - f^d_i, \qquad g_i = f_i - f^d_i,$$

   the curve stays inside the range its own inputs set exactly when (g_{i-1}, g_i) falls in a particular region of the plane, and leaves it otherwise.

4. *Repair, by case.* Where it leaves, replace the quadratic on that bucket with a piecewise quadratic that flattens: it holds one node value across part of the bucket and moves over the rest, with the split point chosen to preserve the area. Which of a handful of cases applies is decided by the signs and relative sizes of g_{i-1} and g_i, and every case preserves the area and stays monotone between the node values.

   Output: a forward curve that is continuous, reprices every quote exactly, stays inside the range its neighbours set, and is positive wherever the inputs are.

The area condition in step two is the whole reason no iteration is needed to reprice — a conserving scheme reprices by algebra. And the case selection in step four is where the scheme stops being linear, since which case applies depends on the data.

Remark (What the sequence shows that the result does not)**.** Two measurements from the same construction, taken as the three year point is passed and then left behind.[2]

Adding the next three quotes after the three year one moves the three year *zero* by nine ten-thousandths of a basis point. Expecting zero is reasonable and the argument for where it fails is the interesting part.

Two different promises are in play, and the naive argument conflates them. Conserving is a promise about the interpolation step alone: given whichever discrete forwards are currently declared, reshaping the curve inside a bucket never changes the zero rate they imply. Repricing the par bonds is a promise about the solve: find the discrete forwards for which every quote hits par. The zero the naive argument expects to survive is built from the first promise; the three year discrete forward is set to satisfy the second, and nothing stops the second from revising a value

[2] `curve::bootstrap_par_stages`.

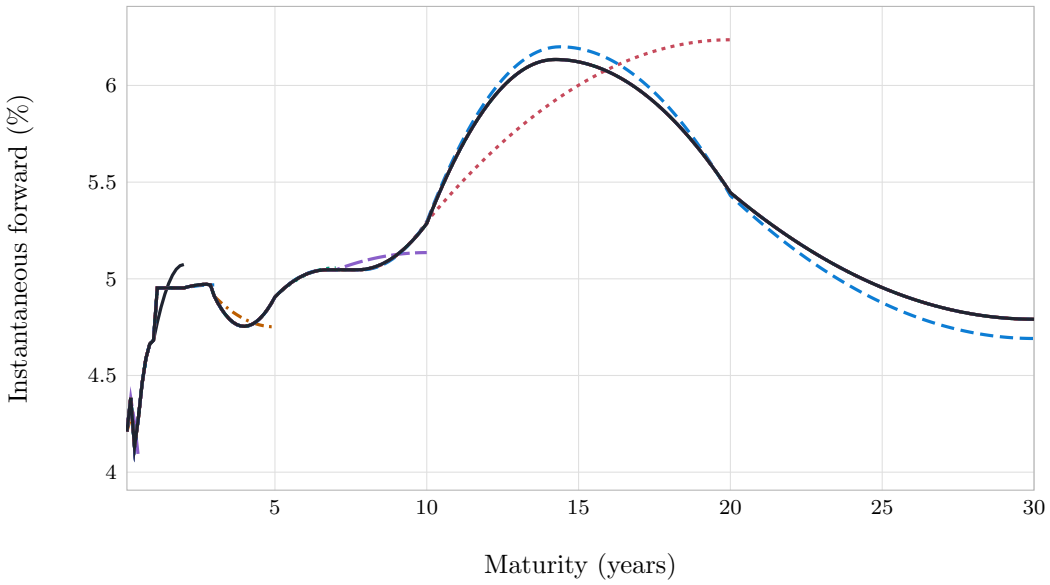

Figure 7.1: The construction as a sequence rather than as a result: the forward curve after each quote is bootstrapped, and then after each sweep of the re-solve. The stage before is drawn faintly behind the current one, because the movement is the point. Two things are visible here that the finished curve hides. Adding a quote does not only extend the curve rightwards — it moves the part already built, since a par bond pays coupons between the nodes and a forward below a node is built partly from the bucket above it. And the stages past the last quote are sweeps, which this scheme needs and the other three do not.
Drawn by `bootstrap_par_stages`. US Department of the Treasury, daily par yield curve rates, as of 2026-09-22 (par yield, semiannual coupon, actual/actual). Retrieved from https://home.treasury.gov/interest-rates-data-csv-archive.

the first was only ever holding steady for a moment. Where every calibrating instrument pays exclusively at nodes the two promises never meet, and the zero really is fixed forever. Here they meet: the coupon at two and a half years sits inside a bucket whose shape moves — conservingly, area intact — the instant the five year quote arrives, and that alone is enough to knock the old three year bond off par and hand the solve a fresh target for that one discrete forward.

The chain is then short. The five year quote creates a bucket beyond three, so the node forward at three years is rebuilt from the buckets on both sides of it and moves by 6.3 basis points. The quadratic on $[2, 3]$ is refitted through that new node value while preserving its area, so the shape inside the bucket changes and the total does not. The coupon at two and a half years is discounted through $\int_2^{2.5} f$, which is half a bucket, and that moves by 1.26×10^{-5} — fifty times the residual left in the whole bucket's area. The three year bond therefore no longer prices to par at the old node value, and the solver moves it until it does.

So the effect is real rather than numerical, and its size is set by how much of the instrument pays between the nodes. It is negligible here and matters in principle: it says the node values are not quite settled even in the region the quotes cover, and that a conserving scheme conserves less than the word suggests once the instruments pay off-node. A three year zero-coupon bond, calibrated instead of the par bond, would have been pinned exactly: its price depends only on the whole-bucket sum $f_1^d + f_2^d + f_3^d$, which the conserving property truly does leave untouched. It is the par bond's coupons landing inside the buckets that breach the guarantee.

Over the same three quotes the *forward* at three years moves by six and a third basis points — seven thousand times as much. Which is the interpolation discussion above's comparison of the four schemes arriving inside a single one: the zeros agree to a rounding error and the forwards do not agree at all, and anything fixing between the nodes is paid on the second.

Set against the other three schemes the effect is not subtle: on the same quotes the monotone convex curve stays inside four per cent to six and a half, where linear interpolation of the forwards runs from under two to nearly eight. The cost is that the case selection depends on the data, so the curve is not a linear function of its inputs. Risk cannot be obtained by applying a fixed matrix to the quotes and has to be bumped numerically instead — and two curves built from the same quotes on different days may not even be in the same case.

None of these is right. They are different priors about what the market would have quoted at maturities it did not quote, and two desks with identical market data and different interpolation will print different prices for anything that fixes off-node.

One might expect the schemes to agree at the nodes and differ only in between, on the grounds that the nodes are pinned by the quotes. That is true only if every calibrating instrument pays exclusively on nodes. It does not: a ten year par bond pays twenty coupons, and only a handful of them land on a node, so the value of the ten year node itself is solved for using interpolated discount factors. The schemes therefore do not produce one curve read three ways. They produce three curves.

7.5 The Curve as a Posterior

The bootstrap finds *the* curve that reprices the quotes exactly. Two things are wrong with that as a statement of the problem, and both were visible earlier in this chapter without being named.

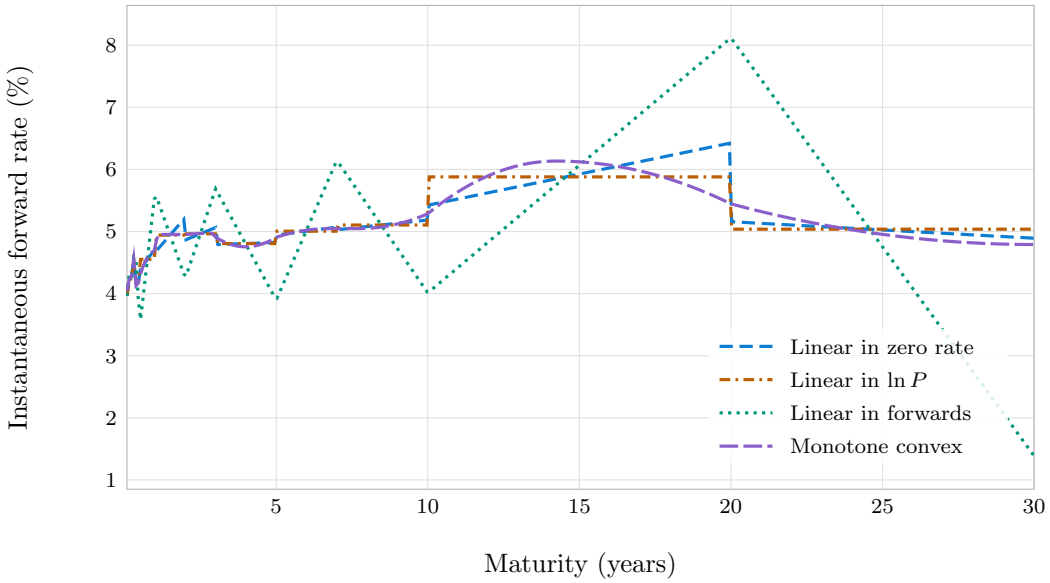

Figure 7.2: The US Treasury par yield curve bootstrapped under four interpolation schemes. All four reprice every quoted instrument exactly, so no examination of the fit distinguishes them, and their zero curves lie within a basis point or two of one another over most of the curve. The instantaneous forward curves, which are what anything fixing between the nodes actually pays, do not agree anywhere. The failures are clearest at the long end, where the quotes are twenty years apart: linear interpolation of the zero rate jumps at every node, linear interpolation of $\ln P$ is a step function, and linear interpolation of the forwards, though continuous, overshoots wildly — forced through the node values, it swings from 3.2% to 6.6% on a curve whose quotes never leave 3.8% to 5.2%. The monotone convex scheme is the repair: continuous like the third, and confined to 3.8% to 6.4% because it is not allowed past the range its neighbours set. It reprices exactly like the others, and it is the only one of the four that cannot be written as a fixed matrix applied to the quotes.

Drawn by `bootstrap_par` and `Curve::inst_forward`. US Department of the Treasury, daily par yield curve rates, as of 2026-09-22 (par yield, semiannual coupon, actual/actual). Retrieved from https://home.treasury.gov/interest-rates-data-csv-archive.

The quotes are mids. A mid is the midpoint of a range in which no trade need have occurred, so repricing it exactly is fitting to a precision the data does not have — and where instruments overlap, as futures and swaps do, exact repricing forces the curve to contort to satisfy quotes that are not mutually consistent to begin with.

And between the quoted maturities the curve is not determined. The discussion above said as much and then chose a scheme; what it did not do is say how undetermined, which is a question with an answer.

Remark (Curve construction is a regression)**.** Put a prior on the forward curve — a Gaussian process, whose kernel says how smooth it is believed to be and over what scale — and treat each quote as a noisy observation of a functional of it. This is Gaussian process regression — specifically what geostatistics calls block kriging, since what is observed is an average over a region rather than a value at a point, exactly as an ore-grade sample is. Since a zero rate is an average of forwards,

$$z(T) = \frac{1}{T}\int_0^T f(u)\,du, \tag{7.5}$$

the observation is *linear* in the curve, so the posterior is Gaussian and available in closed form. What comes out is not a curve but a distribution over curves: a mean, which is the fitted curve, and a standard deviation at every maturity.

Calculation 7.21 (The posterior, in full)**.** Represent the forward curve by its values on a grid $u_1, \dots, u_n$ and give it a Gaussian prior,

$$f \sim N(\mu\mathbf{1}, K), \qquad K_{jk} = \sigma_p^2 \exp\!\left(-\frac{(u_j - u_k)^2}{2\ell^2}\right), \tag{7.6}$$

the squared exponential kernel. Three numbers say everything the prior knows: μ, the level the curve reverts to where nothing is observed; σ_p, how far it may stray from it; and ℓ, the distance over which two forwards stop being related. The figure below uses μ equal to the average quoted rate, $\sigma_p = 1.5\%$ and $\ell = 8$ years.

Now the quotes. A zero rate is an average by (7.5), so observing m of them is

$$z = Hf + \epsilon, \qquad H_{ij} = \frac{\Delta u}{T_i}\mathbf{1}\{u_j \le T_i\}, \qquad \epsilon \sim N(0, \sigma_e^2 I), \tag{7.7}$$

with H the $m \times n$ averaging matrix — row i spreads weight $1/T_i$ evenly over the grid points up to T_i and puts zero beyond — and σ_e the quote error, two basis points here, standing for the bid-offer.

Everything is Gaussian and the observation is linear, so (f, z) is jointly Gaussian with

$$\operatorname{Cov}(f, z) = KH^\top, \qquad \operatorname{Var}(z) = HKH^\top + \sigma_e^2 I,$$

and conditioning a joint Gaussian is a formula:

$$\mathbb{E}[f \mid z] = \mu\mathbf{1} + KH^\top\big(HKH^\top + \sigma_e^2 I\big)^{-1}\big(z - \mu H\mathbf{1}\big), \tag{7.8}$$

$$\operatorname{Var}[f \mid z] = K - KH^\top\big(HKH^\top + \sigma_e^2 I\big)^{-1}HK. \tag{7.9}$$

The matrix inverted is $m \times m$ — eight by eight for the curve below — so the whole construction costs less than the bootstrap it replaces.[3]

[3] `curve::fit_curve` is these two lines and the kernel above.

Remark (Reading the two formulae)**.** In (7.8) the fitted curve is the prior mean plus a correction built from $KH^\top$, the covariance between a forward and each quote. A forward that no quote's averaging window touches gets its correction from the quotes it is still correlated with through K, and that correction fades smoothly rather than vanishing at the last quote — which is the extrapolation beyond it, arrived at by algebra rather than by convention.

In (7.9) the posterior variance does not mention z. The band is fixed by *which* instruments are quoted and how precisely, and not at all by what they say. So it can be computed before the market opens, and a desk can know where its curve is determined without knowing what the curve is.

And $\sigma_e \to 0$ recovers the bootstrap: the inverse becomes that of $HKH^\top$ alone, the mean reproduces every quote exactly, and the posterior variance collapses to zero *in the directions the quotes span* — meaning $H\,\mathrm{Var}[f \mid z]\,H^\top = HKH^\top - HKH^\top(HKH^\top)^{-1}HKH^\top = 0$ exactly, not that (7.9) itself is the zero matrix. A forward's posterior variance is bounded above by K_{jj}, the prior's own — $K - \mathrm{Var}[f \mid z]$ is positive semidefinite, being of the form $MS^{-1}M^\top$ — but it reaches that bound only where the forward is prior-uncorrelated with every quote, which for a kernel like (7.6) happens only in the limit: the band grows back toward the prior gradually, over roughly the length scale, not the instant a quote's window stops covering it.[4] Which is the precise sense in which an exact fit is a limiting case of this rather than a different thing.

Remark (Three constructions, one prior)**.** Writing it this way subsumes the three constructions this chapter has treated as different things. The bootstrap is the limit as the quote error goes to zero with as many degrees of freedom as instruments. A smoothing spline is the posterior mean under a particular prior — that equivalence is a theorem, not an analogy, and it is why the choice of spline and the choice of prior are the same choice. And a parametric family like Nelson-Siegel — $y(T) = \beta_0 + \beta_1\,\ell(T) + \beta_2\,c(T)$, with $\ell(T) = \frac{1-e^{-T/\tau}}{T/\tau}$ and $c(T) = \ell(T) - e^{-T/\tau}$ the loadings behind $\beta_0, \beta_1, \beta_2$'s names of level, slope and curvature — is the same idea pushed to an extreme: give K exactly zero variance in every direction outside a fixed subspace, so no quote, however contradictory, can move the fit out of it. Fix the decay rate τ and that subspace is exactly the one ℓ and c span alongside the constant, a literal infinitely strong prior — zero variance, not merely small. Fit τ too, as is usually done in practice, and the fourth parameter enters through the exponentials rather than linearly, so the same idea becomes an infinitely strong constraint on a four-parameter family rather than a prior on a subspace.

So the earlier point — that interpolation is a modelling choice — can be sharpened: it is a *prior*, and writing it as one makes it something that can be stated, argued about, and checked against how curves actually move.

Remark (The kernel's shape, chosen rather than assumed)**.** The squared exponential in (7.6) is a choice, and the most data-driven alternative is not to assume a functional form at all: take K from the empirical covariance of historical curve changes directly. Write Δf for a day's move in the forward curve on the same grid and estimate $\widehat{\mathrm{Cov}}(\Delta f)$ from a panel of days.

The raw sample covariance is not usable as it stands, for the reason chapter 21's remark The spread is predictable, which is what makes it usable already measured: with a few hundred trading days against dozens of grid points, most of its eigenvalues are noise. The fix is the same test —

[4] curve::the_band_is_tightest_between_the_quotes_and_not_at_them.

keep only the eigenvalues clearing the Marchenko-Pastur edge. On the US Treasury curve, the only one this book's committed data can put through the test, and on the tenors this chapter actually fits a curve to, one year and longer, only the level clears it; put the money-market front end back and a second factor, the slope, clears too. That count is not an artefact of one committed year: it is the same across forty-four years of daily observations spanning several rate cycles. A curve with more cross-sectional structure than the government curve — a swap curve segmented by credit and liquidity tier, say — might put a third eigenvalue above the edge where Treasuries do not; this book's data cannot say, and what follows is scoped to what it can. The curvature chapter 12's principal component analysis always returns as a third shape is a genuine direction, orthogonal to level and slope by construction, but on the Treasury curve, on every panel tested, its eigenvalue sits inside the noise bulk rather than above it, so nothing here treats it as a separately priced factor.[5]

Keeping r eigenvalues this way builds the kernel the same way chapter 12's calculation How the components were found, and what a loading is builds the covariance itself, truncated:

$$\widehat{K} \;=\; \sum_{j=1}^{r} \lambda_j \, v_j v_j^{\top}, \tag{7.10}$$

with λ_j and v_j the jth eigenvalue and loading of $\widehat{\mathrm{Cov}}(\Delta f)$. Taking r equal to the full tenor count reproduces $\widehat{\mathrm{Cov}}(\Delta f)$ exactly; taking it smaller is the truncation, and it is what makes $\widehat{K}$ singular in the discarded directions rather than merely small there — the zero-variance-outside-the-subspace property the squared exponential only approximates with a short length scale.[6]

That rank-truncated covariance is a kernel in the previous remark's exact sense — an infinitely strong prior on the subspace those eigenvectors span, zero variance outside it — except the subspace is read off the data rather than posited as an exponential decay. Nelson-Siegel guesses the shape of the few directions that matter; this measures it.

Remark (Fitting a parametric kernel's own numbers)**.** A smooth kernel like the squared exponential still earns its place alongside the empirical one, because it does something PCA does not: it assigns a definite, if decaying, variance to every direction, including ones no history has yet exercised, where a rank-truncated empirical kernel gives exactly zero.

Its own hyperparameters μ, σ_p, ℓ are fit the same way: maximum likelihood on the joint Gaussian already in use, since $z \sim N(\mu H\mathbf{1}, HKH^{\top} + \sigma_e^2 I)$,

$$\log p(z \mid \mu, \sigma_p, \ell) = -\tfrac{1}{2}(z - \mu H\mathbf{1})^{\top}\big(HKH^{\top} + \sigma_e^2 I\big)^{-1}(z - \mu H\mathbf{1}) - \tfrac{1}{2}\log\big|HKH^{\top} + \sigma_e^2 I\big| + \text{const},$$

maximised over the same $m \times m$ system (7.8) already inverts; the log-determinant term penalises a length scale short enough to track every quote for the flexibility it costs, not only rewarding the fit — an automatic Occam's razor a least-squares fit of the kernel would not have.

One day supplies too few quotes to identify σ_p and ℓ reliably, for the same reason the raw covariance above needed a panel rather than a single day: they describe how curves are shaped,

[5] `risk::curve_spectrum_against_noise`; on the committed year, `risk::only_the_level_clears_the_noise_edge_on_the_bond_curve` and `risk::only_level_and_slope_clear_the_noise_edge_on_the_full_panel`; across decades, `risk::only_level_and_slope_clear_the_noise_edge_across_four_decades`.

[6] `risk::empirical_kernel`, checked against the untruncated covariance by `risk::the_full_rank_kernel_reproduces_the_sample_covariance` and against its own rank by `risk::a_rank_truncated_kernel_has_exactly_that_rank`.

not what one curve says. μ, an intercept, is the exception, refit fresh daily; σ_p and ℓ are estimated rarely, from the same historical panel, and held fixed for the daily fit.

Calculation 7.22 (Where the market has spoken, and where it has not)**.** Fit the committed Treasury snapshot's quoted zero rates from one year out to thirty, with a two basis point error on each, under a prior that the forward curve sits near its own average and can stray one and a half per cent, with an eight year smoothness scale. Two standard deviations of the posterior forward rate:[7]

Maturity	Forward	Band	
1y	4.13%	±2bp	quoted
2y	4.31%	±4bp	quoted
5y	4.68%	±5bp	quoted
10y	5.17%	±15bp	quoted
20y	5.92%	±43bp	quoted
30y	4.49%	±110bp	quoted, and the last one
40y	4.39%	±287bp	extrapolated

Two features, and neither is what one would guess.

The band widens steadily through the quoted region, and is widest at the longest quote. That is not the picture a reader expects from a fit, which is a curve pinned at its knots and loose between them — and (7.5) is why the expectation is wrong. A zero rate constrains an *average* of forwards, not a forward. The front of the curve is covered by eight overlapping averages and is pinned to a couple of basis points; the stretch beyond twenty years enters only through two of them, so one number is asked to determine a decade of curve and cannot. Any intuition that a curve is best known where it is quoted is an intuition about pointwise observations, and these are not pointwise observations.

And past the last quote the band opens towards the prior. At forty years it is nearly three per cent wide, on its way to the four the prior allows. That is the honest statement about extrapolation: beyond the market's reach the curve is whatever was assumed. A construction that extends a flat forward past thirty years is asserting precisely this prior while presenting it as a curve — and pension and insurance liabilities are discounted off exactly that region.

The smoothness scale is doing visible work in both rows. At a three year scale the long end is a sequence of nearly independent points, one average cannot pin any of them, and the thirty year band comes out wider than any curve has ever been. Eight years is a statement that forward rates decades apart move together, and it is that statement, not the data, which makes the long end look determined at all.

Structure (What the band is for)**.** A distribution over curves is not merely a more honest object than a curve. Three things become computable that were not.

Cheapness becomes a z-score. A bond's spread to the curve is currently a number with no scale attached: is four basis points cheap, or is it inside what the construction cannot determine? The posterior answers it. Chapter 22's structure What makes a spread a trade insists a wide spread

[7] `curve::fit_curve`.

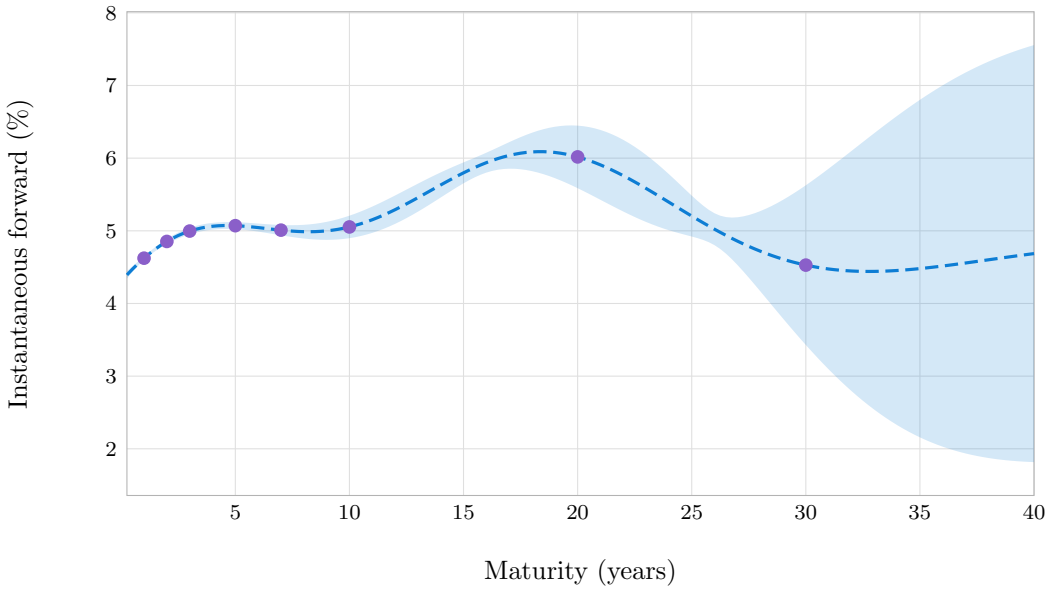

Figure 7.3: The forward curve fitted to the committed Treasury snapshot as a posterior. The shaded region is two standard deviations either side of the mean, and the marks are the eight quoted tenors, from one year to thirty, placed on the fitted curve to show where the information enters — not to claim the curve passes through them, since (7.5) makes each quote a constraint on an *average* of forwards rather than on the forward at its own maturity. The shape of the band is the content: a few basis points across the densely quoted front end, tens of basis points through the twenties where two quotes carry a decade of curve, and opening towards the prior past the last one. Nothing beyond thirty years is observed, and the figure says so rather than drawing a line through it.

Drawn by `fit_curve`. US Department of the Treasury, daily par yield curve rates, as of 2026-09-22 (par yield, semiannual coupon, actual/actual). Retrieved from https://home.treasury.gov/interest-rates-data-csv-archive.

is not a signal; this supplies the width against which to judge, and it is largest exactly where the curve is least constrained, which is where apparent mispricings congregate.

Bucketed risk acquires a covariance. §7.9 reports sensitivities to each calibrating instrument and then needs to know how those buckets move together. Equation (7.9) supplies one, and what to do with it deserves care, because it is the same distinction chapter 24's structure The two requirements want different models draws between pricing and aggregation.

Shock the instruments one at a time to hedge. A bucketed sensitivity is the answer to "how much of instrument i do I trade". It has to be computed by moving instrument i and nothing else, because that is what a hedge does — you buy the ten year swap, not a correlated bundle. Independent bumps are therefore correct for hedging and would be wrong to replace with joint ones.

Shock them together to measure risk. The variance of the book's value is $\Delta^\top \Sigma \Delta$ with Δ the vector of bucketed sensitivities, and no single-instrument bump ever produces that number: a book flat to every bucket individually is flat outright, but a book with offsetting buckets can still be badly exposed if the two do not move together the way the offset assumed. So the covariance enters at aggregation, not at differentiation, and the two are different calculations from the same Δ.

And the eigenvectors are the scenarios worth running. Shocking along the leading eigenvectors of Σ gives level, slope and curvature moves — the same factors as chapter 24's calculation What a one-factor report cannot see, arriving from the construction rather than from a principal component analysis of history. That is a change of basis on the same information, not extra information.

One caution about *which* covariance, since two are in play and they answer different questions. Equation (7.9) is the uncertainty about today's curve given today's quotes: interpolation risk, which does not go away by waiting and is invisible to a historical estimate. A covariance estimated from a history of quotes is how the market moves from day to day. A full risk figure wants both, and adding them is only correct if they are independent, which is a stronger assumption than it looks.

And the extrapolation risk has a number. A thirty year liability discounted off a curve whose last quote is at twenty carries an uncertainty the table above quantifies. Insurers and pension funds hold exactly this exposure, and reporting it as a point estimate is the same error chapter 21's Bayesian calibration argues against, committed at the one place where nobody thinks of the curve as a model.

The cost is a prior, which somebody must choose and defend. That is the same cost as choosing an interpolation scheme, and it is paid in the same currency — the difference is that this version says out loud what is being assumed, and reports what it is worth.

7.6 The Front End Is a Step Function

Everything above treated the curve as a smooth object to be interpolated. At the short end that is false in a way that costs money, because the overnight rate does not drift — it is held by a central bank and moves only when the committee meets.

The meeting dates are published a year ahead. So the forward overnight curve is a step function with known step locations and unknown step sizes, and the construction that says so is the one to use: piecewise constant between meetings, with a knot at each meeting date, solving for the jumps.

Calculation 7.23 (Solving for the steps)**.** It is a forward substitution, with every part determined by the calendar.

Let the overnight rate be r_0 until the first meeting and step to r_k after the kth, at known dates $m_1 < \cdots < m_K$. An overnight indexed swap pays the compounded overnight rate over its life, so to the accuracy that matters here its quote is the average of the path,

$$q_i = \frac{1}{T_i}\int_0^{T_i} r(u)\,du = \frac{1}{T_i}\left[\sum_{k<i} r_k\,(m_{k+1}-m_k) + r_i\,(T_i - m_i)\right], \tag{7.11}$$

writing $m_0 = 0$. Choose the quotes so that the ith matures just after the ith meeting and every term in that bracket but the last is already known, so

$$r_i = \frac{q_i T_i - \sum_{k<i} r_k (m_{k+1}-m_k)}{T_i - m_i}.$$

One division per meeting, taken in order. There is no optimiser and no residual: the system is triangular because each instrument spans exactly one more meeting than the last, and the calendar is what makes it so.[8]

On four meetings at a quarter's spacing with swaps maturing a fortnight after each, quotes generated from a path that hikes, holds, hikes and cuts return exactly that: $+25$, 0, $+25$, -25 basis points.

Twenty-eight basis points is many times the bid-offer, and it is not noise: it is one-signed and determined by the calendar, so any instrument whose accrual straddles a meeting is mispriced systematically. This sharpens the point above at its limit — the interpolation scheme is a statement about how the underlying rate behaves, and here the correct statement is available for free from a published calendar.

Remark (The output is a schedule of decisions)**.** Building the curve this way changes what the calibration produces. Instead of a smooth function whose values have no individual meaning, the output is a set of numbers indexed by meeting, each of which is the market's expected policy move there — and since the outcome space is a short list of twenty-five basis point multiples, chapter 4's binomial model showed how to read a probability off it. The four steps above divide by the size of a move to give $+1$, 0, $+1$, -1: a hike fully priced, a hold, another hike, a cut.

Nothing in 7.23 consults an opinion. The quotes are market instruments and the steps are whatever reprices them, so the curve reports what has already been priced rather than what anyone believes. That is what makes the number useful: a mispricing acquires a location, at a named meeting, rather than being a general sense that the front end is rich.

Calculation 7.24 (Risk to a decision rather than to a tenor)**.** A front end book is not naturally described by tenor buckets, because what it is exposed to is decisions. The risk report that says so is indexed by meeting: move one step by a basis point, leave the others alone, reprice.

[8] `events::build_meeting_curve`, with the inversion checked against a known path.

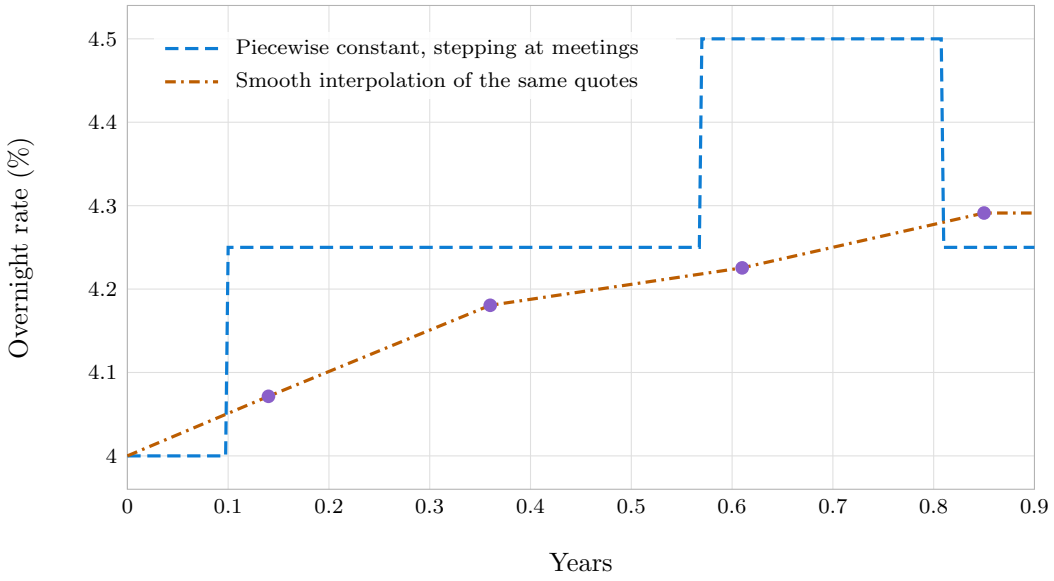

Figure 7.4: The same four quotes built two ways. Both reprice every swap exactly; they disagree about the overnight rate on almost every day in between. The step curve holds the rate flat between decisions because that is what the rate does, and the smooth one spreads each move across the weeks either side of it. The error is worst just *after* a decision — nineteen basis points a week after the first meeting against four and a half a week before — because by then the smooth curve is still catching up with a move that has already happened. Over the year its largest error is twenty-eight basis points, and its sign is set by where in the accrual period the meeting falls, which is public information.

Drawn by `build_meeting_curve`.

That is not a bucket bump. Moving a step shifts the entire curve *after* that meeting and nothing before it, so what it measures is the exposure to everything the instrument accrues from that date onwards. Differentiating (7.11) gives the answer in closed form:

$$\frac{\partial q}{\partial(\text{step at } m)} = \frac{T-m}{T}, \tag{7.12}$$

the fraction of the instrument's life falling after the meeting.[9]

For a one year swap against the four meetings above, that is 90%, 68%, 43% and 19%. The risk is heavily front-loaded — the nearest meeting carries nearly five times the exposure of the furthest — which no tenor bucketing reveals, because in tenor coordinates a one year swap is simply a one year swap.

And the localisation is exact in a way a bucket cannot be. A three month swap has substantial exposure to the meeting inside its life and *identically zero* to every meeting after it, however close the next one falls. A key rate bucket, which shifts a region of the curve smoothly, always leaks some sensitivity across a date on the far side of maturity; this does not, because the curve genuinely does not move there.

Calculation 7.25 (Trading a view on one meeting)**.** The sensitivities suggest a use. A trader who thinks a meeting will deliver more than the curve has priced wants a position exposed to *that* step and to no other — otherwise the trade is also a bet on the neighbouring meetings, and being right about one while wrong about another is indistinguishable from being wrong.

Such a position exists, and the reason is the structure that made 7.23 a forward substitution. In profit and loss terms a step of size δ at m adds $\delta(T-m)$ to what a unit notional swap maturing at T accrues, so the exposure matrix

$$E_{ik} = (T_i - m_k)^+$$

is lower triangular, hence invertible, and the weights isolating any one meeting are a back substitution — the same calendar structure used the other way round.[10]

The trade is a butterfly. Isolating the third of four meetings gives weights of about $+4.6$, -8.7, $+4.3$ on the three instruments spanning it and nothing on the fourth: wings up, body down, roughly $1:-2:1$.[11] That is not a coincidence of these dates. Exposure $(T-m)$ is *linear* in the meeting date, killing a linear function is what a second difference does, and it is the same reason chapter 22's curve fly kills level and slope. A meeting view is a butterfly in meeting space.

And the choice of instruments decides whether it can be put on at all. The back substitution divides at each step by $T_i - m_i$, the exposure of an instrument to the meeting it barely spans. Swaps maturing a fortnight after each meeting make that divisor a fortnight, and the notionals compound: isolating the third meeting then needs a gross position of a thousand times the exposure it buys. Swaps maturing just *before the next* meeting make the divisor a whole segment, and the same trade

[9] `events::meeting_sensitivities`, checked against a numerical bump.

[10] `events::isolate_meeting`, with the resulting exposures checked to be one at the target meeting and zero at every other.

[11] `events::isolating_a_meeting_is_a_butterfly`.

needs eighteen. The instruments that are convenient for the bootstrap are the wrong ones for the trade, and the difference is a factor of sixty.[12]

Remark (What the position is really short)**.** With unit exposure the profit is exactly the gap between what the committee does and what the curve had priced, which makes the trade a clean expression of the view and is the reason to normalise it that way.

A butterfly in meeting space is short the assumption that the meetings either side of the target are correctly priced — and that is the same fact as the isolation itself, not an extra risk left over from it: giving the position zero exposure to the neighbours' steps is exactly what leaves it fully exposed to the surprise landing at one of them instead. If the committee moves at the neighbouring meeting rather than the target one, the trade loses on both wings. So the view being expressed is sharper than "rates will rise": it is a bet on which meeting delivers the surprise, and a position built to isolate that bet cannot also be hedged against having picked the wrong one — the two are the same construction.

Structure (Three constructions, three priors)**.** *Against monotone convex.* It is tempting to reach for the monotone convex machinery above and put its knots at the meetings, and it is the wrong tool. That scheme exists to stop a *continuous* forward curve overshooting between sparse quotes. Here the curve is meant to be discontinuous and the discontinuities are at known dates, so piecewise constant with knots at meetings is not a crude monotone convex — it is the right shape for a rate held constant by decision and moved by decision. The two therefore govern different regions and are stitched rather than blended: stepping out to the horizon over which meetings are scheduled and quoted, smooth beyond it. Where they join is a modelling choice.

As a posterior. The step construction is a Gaussian process prior, written in the kernel rather than in prose. Saying the forward is constant between meetings and free to jump at them is saying that $K(u, v)$ of (7.6) is replaced by a block kernel: perfect correlation between two dates in the same inter-meeting segment, and none across a meeting. The smooth schemes are the opposite extreme — correlation decaying with distance and no knowledge of the calendar at all.

Two things follow. The block prior can be *softened* — correlation across a meeting near one rather than at zero — which is the honest description of a curve whose meeting dates are known but whose committee sometimes moves between them. And running the posterior rather than the exact solve puts error bars on the steps, so the implied probability of a hike arrives with a width: not that the market prices a hike, but that it prices one to within so many basis points given the bid-offer on the swaps that determine it.

7.7 Two Curves: Projection and Discounting

Everything so far assumed a single curve doing two jobs: producing the forward rates the floating leg pays, and discounting the cashflows. That assumption is what made the floating leg telescope, and it is false.

It was always slightly false and became visibly so in 2008, when the spread between overnight rates and term interbank rates — which had been a basis point or two — opened to hundreds of

[12] `events::the_instrument_set_decides_whether_the_trade_is_possible`: 1066× against 18× gross notional to isolate the same meeting.

basis points and stayed open. Two rates that had been treated as the same rate were plainly not. The reason is credit and liquidity: lending to a bank for three months is not the same as lending overnight and rolling, because the three month loan cannot be withdrawn if the bank deteriorates.

The modern picture separates the two jobs.

- The *projection* curve produces the forward rates a floating leg pays. It is specific to the tenor and the index: a three month curve and a six month curve on the same currency are different curves.
- The *discount* curve present-values the cashflows. It is determined by how the trade is collateralised, and not at all by which index the coupons reference.

So the floating leg becomes

$$\text{Float}(0) = \Sigma_j \delta_j^p \, F^{\text{proj}}(0; \tau_{j-1}^p, \tau_j^p) \, P^{\text{disc}}(0, \tau_j^p),$$

in which nothing cancels, and the par swap rate is the general ratio in (7.3) rather than its telescoped form. Both the projection curves and the discount curve are bootstrapped, from their own instruments, usually simultaneously.

Remark (What the projection curve's $P(0,T)$ is not)**.** It is tempting to read $P^{\text{proj}}(0,T)$ the way $P(0,T)$ was defined at the start of this chapter, as the price of a zero coupon bond. It cannot be one: lending for six months is worth more than rolling two three month loans, so there is no tradable claim whose price $P^{\text{proj}}(0, 6\text{m})$ could honestly be.

What is tradable is a forward rate agreement, cash settled against the three month index *fixing* observed at the forward date rather than funded for the period it covers — it commits no one to six months of credit exposure, only to a view on what the fixing will be. The projection curve is chosen exactly so that $F^{\text{proj}}(0; 3\text{m}, 6\text{m})$, computed from (7.1)'s ratio, reproduces that quoted FRA rate. So $P^{\text{proj}}(0,T)$ is bookkeeping engineered to make the algebra output the right forward, not the price of anything a desk could buy; only the discount curve, built from collateral that is real cash, prices anything at all.

The same point, sharpened: at the six month date there are three different numbers with three different meanings. $P^{\text{disc}}(0, 6\text{m})$, from the daily rolled overnight curve, discounts real collateral cash, so it prices something that actually changes hands. $P^{\text{proj},3m}(0, 6\text{m})$ makes (7.1)'s ratio against $P^{\text{proj},3m}(0, 3\text{m})$ reproduce the 3×6 FRA's forward three month fixing — pure bookkeeping. $P^{\text{proj},6m}(0, 6\text{m})$ is different again: six months is that curve's own first tenor, so it happens to coincide with the one real transaction on it, the spot six month deposit; every later point on the same curve — matched against a 6×12, 12×18, ... FRA in turn — is bookkeeping once more. None of the three answers the same question, and past the first tenor none prices anything that changes hands. The gap between the last two is exactly the tenor basis below: rolling two three month resets is worth more than one six month lock-in, so $P^{\text{proj},3m}(0, 6\text{m}) > P^{\text{proj},6m}(0, 6\text{m})$, and both sit below $P^{\text{disc}}(0, 6\text{m})$ for the credit and liquidity reasons that opened this section.

The claim that collateral determines the discount curve deserves a derivation rather than an assertion.

Calculation 7.26 (Collateral sets the discount rate)**.** Consider a derivative, fully collateralised in cash under a standard agreement: at every instant the party that is out of the money posts collateral equal to the trade's value V_t, and the party holding it pays interest on it at the collateral rate $c(t)$, which for a standard cash agreement is the overnight rate.

Take the position of the party holding the trade as an asset. It owns the derivative, worth V_t, and it owes back the collateral it holds, also V_t, on which it pays $c(t)V_t\,dt$. Consider the self-financing portfolio consisting of the derivative, the collateral obligation, and a money market account funded at the risk free rate $r(t)$. Being fully collateralised, the position costs nothing to hold: the collateral received funds the trade exactly. Its total value is therefore zero at all times, and its change is

$$dV_t - c(t)V_t\,dt$$

— the change in the derivative's value, less the interest owed on the collateral against it.

This combination is a traded, self-financing position with zero cost. By the no-arbitrage argument of chapter 4's fundamental theorem it must have zero drift in the risk neutral measure, so $\mathbb{E}^{\mathbb{P}}[dV_t] = c(t)V_t\,dt$, and hence the process

$$e^{-\int_0^t c(s)ds}V_t$$

is a martingale under $\mathbb{P}$, giving

$$V_t = \mathbb{E}^{\mathbb{P}}\Big[e^{-\int_t^T c(s)ds}V_T \,\Big|\, \mathcal{F}_t\Big]. \tag{7.13}$$

The discount rate is c, the rate paid on the collateral, and the risk free rate r has dropped out entirely. It has dropped out because a collateralised derivative is never funded — the collateral funds it — so the cost of funding never enters. This is why the market discounts collateralised trades on the overnight curve, and it is also why the answer changes when the collateral terms change: post collateral in a different currency and c becomes that currency's overnight rate adjusted by the cross currency basis; post no collateral and there is no single c at all, and the value acquires funding and credit adjustments instead.

Remark (The step that assumes rehypothecation)**.** One sentence in that derivation is carrying more than it looks: *the collateral received funds the trade exactly.* That is only true if the party holding the collateral may actually spend it.

Under a standard title transfer agreement it may. The cash is transferred outright rather than pledged, the receiver owns it, and it can be used — rehypothecated — to fund the hedge. Then no separate funding arises, r cancels, and (7.13) follows.

Segregate the collateral instead — hold it at a third party custodian, with no right to use it — and the cancellation fails. The receiver must now fund the position from its own balance sheet at its own rate while the collateral sits earning c somewhere it cannot reach. Two rates appear where one did, the difference between them does not cancel, and the value picks up a funding adjustment.

This is not a technicality about paperwork. It is exactly why variation margin and initial margin are treated differently: variation margin is rehypothecable and therefore discounts at c, while initial margin must be segregated by regulation, cannot be used, and so costs the posting party its funding spread for the life of the trade. That cost has its own name, a margin valuation adjustment, and it exists precisely because the sentence above is false for it.

Remark. Notice that (7.13) is the ordinary risk neutral pricing formula of chapter 4's fundamental theorem with the accumulation factor $A_t = e^{\int_0^t r}$ replaced by $e^{\int_0^t c}$. In the language of chapter 6, the collateral account is the numeraire. Nothing about the theory changes; what changes is which account the theory is applied to, and the answer is the one the contract actually pays interest on.

The differences between all these curves are traded in their own right. Tenor basis swaps exchange one tenor's floating leg for another's, overnight indexed swaps exchange fixed for compounded overnight, and cross currency basis swaps exchange floating legs in two currencies. In a single curve world all three would be worthless by construction. That they are not is the measure of how wrong the single curve was.

7.8 The Rest of the Multicurve Stack

Separating projection from discounting is the first split. There are three more major splits, and they compose: a trade is priced off a projection curve chosen by its index, a discount curve chosen by its collateral, a currency adjustment if the two are not the same currency, and — if it is a bond with optionality — a spread reconciling the model to the quote. Each is a different object and they are frequently confused with one another.

Tenor basis

Two swaps on the same currency and maturity, one paying three month rates and one paying six month, do not have the same par rate. The difference is the *tenor basis*, and it is quoted directly as a basis swap: exchange three month payments plus a spread for six month payments.

Remark (What the tenor basis is the price of)**.** Lending for six months rather than rolling three month loans twice differs in two ways, and both are worth money.

The first is credit. The six month loan cannot be withdrawn if the borrower deteriorates at the three month point, so the lender is exposed for longer and charges for it. This is the same argument that separated the discount curve from the projection curve, applied between two term rates rather than between term and overnight.

The second is optionality, and it survives even when the credit component does not. Rolling gives the lender a decision at the three month point — to roll, or not — and a decision has non-negative value. So the rolled position is worth at least the term position, and the basis is the price of that flexibility.

Since 2008 the basis has been persistently non-zero and it widens in stress, which is what a credit and optionality premium should do. In an index transition to overnight-based rates much of the credit component goes away; the optionality does not, and neither does the fact that a curve must be built per index.

The consequence for construction is that there is no single curve. Each index and tenor has its own projection curve, they are built simultaneously with the discount curve because the basis instruments involve two of them at once, and a book of swaps on four indices carries four projection curves and one discount curve, all of which must be arbitrage-free jointly.

Cross-currency basis

The same problem across currencies. Borrowing dollars directly and borrowing euros and swapping them into dollars should cost the same, since chapter 17 derives the forward exchange rate as a ratio of numeraires with no freedom in it. They do not cost the same, and the wedge is the cross-currency basis.

Definition 7.27 (Cross-currency basis swap)**.** Two parties exchange floating payments in different currencies over the same schedule, and — unlike a single-currency swap — they exchange *notionals* as well, at the start and again at maturity, both at the spot rate fixed on day one. A euro-dollar basis swap therefore has one party pay dollar floating on a dollar notional and receive euro floating on a euro notional, having handed over the euros at the start and taken them back at the end.

The basis is a spread b added to the floating rate on the *non-dollar* leg, by convention for every pair quoted against the dollar:

$$\text{pay } \big(\text{EUR floating} + b\big) \quad \text{against} \quad \text{receive USD floating.}$$

A negative b, which is the usual sign, means the party receiving dollars pays for them.

Remark (Why the notional exchange is the whole point)**.** The instrument is often described as though it were an interest rate swap in two currencies and it is not.

It is a funding trade. The notional exchange makes it a loan of one currency against another for the life of the swap: euros in at the start, euros back at the end, dollars the other way. The floating legs are the interest on those loans. So the basis is the price of borrowing dollars against euro collateral rather than a view on either rate, and it moves with the scarcity of dollar funding rather than with the level of either curve.

And it is the reason a discount curve depends on the collateral currency. A dollar cashflow collateralised in euros earns the euro collateral rate on the margin, which has to be converted; the conversion is priced by this instrument, so the discount curve for that trade is the dollar curve adjusted by the basis. That is not a small correction applied afterwards — it is a different curve, and a book with agreements in several currencies carries one per agreement.

For curve construction the effect is mechanical and unavoidable: a dollar cashflow collateralised in euros is not discounted on the dollar curve. It is discounted on a dollar curve adjusted by the basis, because the collateral is earning a euro rate and the conversion is not free. So the discount curve depends on the collateral *currency* as well as the collateral rate, and a book with agreements in several currencies has a discount curve per agreement. Chapter 22's catalogue of constraint trades treats the basis as one of them; here it is an input, and treating it as a small correction is how a cross-currency book comes to be systematically mismarked.

Option-adjusted spread

The last object in the stack is different in kind from the others.

Definition 7.28 (Option-adjusted spread)**.** Let $V(s)$ be the model value of an instrument with all cashflows discounted on the curve shifted by a constant spread s, where a model has been used to value any optionality the instrument contains. The option-adjusted spread is the s solving

$$V(s) = \text{market price.}$$

For a bond with no optionality this is just a yield spread expressed against a curve rather than against a single yield. The interesting case is a callable bond or a mortgage pass-through, where the quoted price is low partly because the issuer or the homeowner holds an option. A plain yield spread attributes all of that cheapness to credit and liquidity. The OAS values the option separately and reports what is left.

Remark (An OAS is a property of a model, not of a bond)**.** This is the most misused number in fixed income, and the reason is visible in the definition: $V(s)$ requires a model for the optionality, so the OAS is whatever spread reconciles *that model* to the price. Two desks with different volatility assumptions compute different OAS for the same bond at the same price, and neither is wrong. The number is a residual, and a residual carries every error in everything that produced it.

Two consequences follow.

Comparing OAS across desks or across model versions is meaningless, whereas comparing it across bonds priced by one model on one day is exactly what it is for — the model errors are common and largely cancel in the comparison.

And the spread should be *uncorrelated with volatility*, which makes a diagnostic. Suppose the volatility market rises. The bond cheapens, because the option the issuer holds is worth more. A model given the new volatility reprices that option by the same amount, so at an unchanged s the model value falls with the market's and the spread does not move — which is exactly what "option-adjusted" is claiming to have achieved. If instead the model's volatility is stale, or its option is misvalued, the model price does not fall as far as the market's and s has to move to close the gap. So an OAS that tracks the volatility surface is the warning sign: the optionality has not been removed, only relabelled as spread. Regressing changes in OAS on changes in implied volatility costs nothing, and a significant loading means the number is not doing the job its name claims.

7.9 Curve Risk

Finally, how the risk of all this is expressed. A book of linear products has one exposure, the curve, but the curve is a function and not a number, so its risk has to be bucketed.

The crudest measure moves the whole curve. The parallel DV01 is the change in value for a one basis point shift in every rate, and for a single swap it is the annuity, by the exercise above.

The parallel number is not enough, because a book can be flat to a parallel shift and heavily exposed to the curve steepening. What is computed instead are key rate durations: shift the quote of one calibrating instrument by a basis point, rebuild the entire curve, and reprice. The result is the sensitivity to that instrument, in the units the trader can actually hedge in, because the hedge is a position in that instrument. Summing the buckets recovers the parallel number.

This is a general principle rather than a fixed income one, and it deserves stating plainly. The risk is reported against the instruments used to build the curve, not against the curve's own internal parametrisation, and the two differ. Change the interpolation scheme and the bucketed risk moves even though the curve reprices the same instruments, because a shift in one quote propagates differently. Risk is a property of the construction, not only of the prices.

References

- Hagan, P. S., & West, G. (2006). Interpolation methods for curve construction. *Applied Mathematical Finance*, 13(2), 89–129.

- Piterbarg, V. (2010). Funding beyond discounting: collateral agreements and derivatives pricing. *Risk*, 23(2), 97–102.

- Bianchetti, M. (2010). Two curves, one price. *Risk*, 23(8), 66–72.

Chapter 8

Term Structure Models

We put the interest rate curve in motion. Chapter 7 built today's curve out of today's quotes, which is enough to price anything linear; here we ask how the whole curve may diffuse without admitting arbitrage, and price the products whose value depends on the answer. We work in a single curve framework, so that there is one interest rate in the market and the projection and discount curves of chapter 7 coincide.

8.1 What We Are Modelling

We take from chapter 7 the discount factors $P(t,T)$, the instantaneous forward rate

$$f(t,T) = -\frac{\partial}{\partial T}\ln P(t,T), \qquad P(t,T) = e^{-\int_t^T f(t,s)ds},$$

the short rate $r(t) = f(t,t)$, and the initial curve $f(0,\cdot)$ inferred from market instruments. All of that was a description of the curve at one instant.

The difference between interest rates and equities now becomes the whole problem. In chapter 5 the state of the market was a number S_t and we gave it a diffusion. Here the state is the curve $T \mapsto f(t,T)$, an entire function, and we have to give *that* a diffusion. Two questions arise immediately and neither has an analogue in the equity case. What drift is consistent with no arbitrage, when every point of the curve is the price of a traded bond and they must all be consistent with one another? And can the resulting infinite dimensional object be made to depend on finitely many state variables, so that anything can actually be computed?

The first question is answered completely by the Heath-Jarrow-Morton framework, which is the subject of the next section. The second is answered only under a restriction on the volatility, which gives the Hull-White model here and the quasi-Gaussian models of chapter 12.

8.2 Heath-Jarrow-Morton Framework

The first question of the previous section — which drifts are consistent with no arbitrage — has a complete answer, that there is no freedom at all. Once the volatility of the curve is chosen, the drift is determined. This is the Heath-Jarrow-Morton drift condition, and everything in chapter 12 and chapter 14 that concerns interest rates rests on it.

The starting point, and why it is not an assumption

Take as given a filtered probability space carrying a Brownian motion W, and suppose each discount bond follows

$$\frac{dP(t,T)}{P(t,T)} = r(t)\,dt + \nu(t,T)\,dW_t, \tag{8.1}$$

with $\nu(t,T)$ any adapted process. The volatility is free; the drift is not.

Chapter 4's fundamental theorem establishes that a measure under which every traded asset discounted by the money market account is a martingale exists if and only if there is no arbitrage. So (8.1) is the general form of an arbitrage-free bond price, not a special one, and ν is where all the modelling freedom lives.

Two constraints on ν come free and both matter below. A bond pays one unit at its own maturity, so $P(T,T) = 1$ with certainty, which forces

$$\nu(t,t) = 0. \tag{8.2}$$

A bond about to mature cannot be volatile. And $\nu(t,T)$ is a volatility of a *price*, indexed by the maturity of the bond, so the family $\{\nu(t,\cdot)\}$ is called the volatility term structure.

The drift condition

Theorem 8.1 (Heath, Jarrow and Morton)**.** *Let bond prices follow* (8.1) *under the risk-neutral measure and set*

$$\sigma(t,T) = -\frac{\partial \nu}{\partial T}(t,T). \tag{8.3}$$

Then the instantaneous forward rate satisfies

$$df(t,T) = \sigma(t,T)\left(\int_t^T \sigma(t,s)\,ds\right)dt \;+\; \sigma(t,T)\,dW_t. \tag{8.4}$$

The drift is a function of the volatility alone: no choice remains.

Proof. Apply Itô's lemma to $\ln P(t,T)$ using (8.1). The logarithm of a process with proportional volatility ν picks up the usual variance correction,

$$d\ln P(t,T) = \Big(r(t) - \tfrac{1}{2}\nu^2(t,T)\Big)dt + \nu(t,T)\,dW_t.$$

Now differentiate in the maturity. Since $f(t,T) = -\partial_T \ln P(t,T)$ and the two derivatives commute,

$$df(t,T) = -\frac{\partial}{\partial T}\, d\ln P(t,T) = \frac{\partial}{\partial T}\Big[\tfrac{1}{2}\nu^2(t,T)\Big]dt - \frac{\partial \nu}{\partial T}(t,T)\, dW_t,$$

the short rate $r(t)$ having no T in it and so contributing nothing. Carrying out the derivative of the square and substituting (8.3),

$$df(t,T) = \nu(t,T)\,\frac{\partial \nu}{\partial T}(t,T)\, dt + \sigma(t,T)\, dW_t = -\nu(t,T)\,\sigma(t,T)\, dt + \sigma(t,T)\, dW_t.$$

It remains to identify $-\nu$. Integrating (8.3) in the maturity from t to T and using (8.2) to fix the constant,

$$-\nu(t,T) = \int_t^T \sigma(t,s)\, ds, \tag{8.5}$$

which is (12.1). □

Remark (Where each hypothesis went)**.** The martingale measure supplied the drift rP, and with it the fact that the drift of the *bond* carries no maturity dependence at all. That is what makes $r(t)$ disappear on differentiating, and it is the whole reason a condition emerges: the forward rate's drift is left holding only the variance correction.

The variance correction $-\frac{1}{2}\nu^2$ is what makes the condition non-trivial. Without Itô's second-order term the forward rate would be driftless.

And $P(T,T) = 1$ fixed the constant of integration in (8.5). Without it ν would be determined only up to a function of t, and the drift with it. A bond maturing at par is what closes the system.

Remark (The same volatility, a different drift, under the historical measure)**.** Everything above is written under the risk-neutral measure, where (8.1)'s drift is not a choice — it is r, and (12.1) is what that forces on the forward rate. Call the Brownian motion driving (8.1) $W^{\mathbb{Q}}$ to be explicit about it.

Under the historical measure the bonds keep the same volatility ν, by chapter 6's invariance argument — a change of measure moves drifts, not diffusion coefficients — but the drift is free again. Write $\theta(t)$ for chapter 6's market price of risk, one number per Brownian motion and not per bond, since chapter 6 showed every bond driven by the same risk must reward it at the same rate. Girsanov gives $dW_t^{\mathbb{Q}} = dW_t^{\mathbb{P}} + \theta(t)\, dt$, and substituting into (8.1),

$$\frac{dP(t,T)}{P(t,T)} = \big(r(t) + \nu(t,T)\theta(t)\big)\, dt + \nu(t,T)\, dW_t^{\mathbb{P}}.$$

Repeating the argument that produced (12.1) with this drift in place of r adds exactly one term to it,

$$df(t,T) = \left[\theta(t)\,\sigma(t,T) + \sigma(t,T)\int_t^T \sigma(t,s)\, ds\right] dt + \sigma(t,T)\, dW_t^{\mathbb{P}},$$

the risk-neutral drift of (12.1) plus $\theta(t)\sigma(t,T)$. The volatility term structure σ is calibrated once, from today's prices, and does not change between the two equations. The market price of risk θ is a second, independent object, estimated from history rather than quoted, and chapter 21 is where that estimation is actually attempted.

Structure (The bonds are not independent assets)**.** The result is surprising on first meeting — an entire curve of assets and no drift may be chosen — and stops being surprising once one notices what the objects are.

Bonds of different maturities are not separate securities that happen to be correlated. They are all functions of one object, the curve, and a bond is the integral of the curve over an interval. So the family $\{P(t,T)\}_T$ is a single degree of freedom per Brownian motion dressed as a continuum, and demanding that every member of it be a martingale after discounting is demanding one condition of one object, over and over. That the conditions are compatible at all is the content; that they leave no freedom follows.

The equity case has no analogue because there is nothing to be consistent with. A stock's drift is pinned by the same argument — it is r under the martingale measure — but a stock is one asset and the condition is one equation. Here it is an equation for every maturity, and (12.1) is what the whole system collapses to.

Remark (Testing it, and what failure looks like)**.** The drift condition has a consequence sharp enough to check: a model built with it reprices the curve it started from. Simulating the whole forward surface under (12.1) and averaging $\exp(-\int_0^T r)$ returns $P(0,T)$, and nothing in the simulation is told what that number is.[1]

Setting the drift to zero instead — which is what one writes down if the convexity term is forgotten — breaks it. The drift is second order in the volatility, so at one per cent volatility the error is under a basis point of discount factor at one year and grows to half a per cent at ten.[2] It also has a sign: without the convexity drift the forwards are too low and the discount factors too high. So the failure is the uncomfortable kind — nearly invisible on the short instruments a desk would check first, and compounding out to the maturities where the exotic actually sits.

The same equation in time to maturity

Equation (12.1) indexes the curve by maturity date, which is how the market quotes it, and is the wrong coordinate for almost every other purpose. A bond's maturity is a fixed date; a modeller cares about the ten year point, which is a different bond every day.

Definition 8.2 (Musiela parametrisation)**.** Write the curve by time to maturity rather than by maturity date,

$$r_t(x) = f(t, t+x), \qquad x \geq 0. \tag{8.6}$$

Calculation 8.3 (The transport term)**.** Holding x fixed means letting the maturity date advance with the clock, so the chain rule contributes an extra term. Writing $\bar\sigma(t,x) = \sigma(t,t+x)$,

$$dr_t(x) = \left(\frac{\partial r_t}{\partial x}(x) + \bar\sigma(t,x) \int_0^x \bar\sigma(t,y)\, dy \right) dt \;+\; \bar\sigma(t,x)\, dW_t. \tag{8.7}$$

The new piece is $\partial_x r_t$, and it is not a modelling term: it is the statement that if nothing happens, tomorrow's nine year rate is today's ten year rate. The curve slides along itself.

[1] checked at one, three and five years.

[2] measured, and one-signed.

Structure (Rolling down is a transport equation)**.** Equation (8.7) is a stochastic partial differential equation on a space of curves, and the operator $\partial/\partial x$ appearing in its drift is the generator of translation. In the absence of volatility it reduces to $\partial_t r = \partial_x r$, whose solution is $r_t(x) = r_0(t+x)$ — the initial curve, read further along. That is precisely the carry and roll-down that chapter 22 computes as a trading quantity, arriving here as the deterministic part of the dynamics.

The parametrisation earns its keep twice more in these notes. Chapter 12 needs it because the criterion for a finite dimensional realisation is a statement about the Lie algebra generated by the drift and volatility vector fields, and the transport term is one of those fields — bracketing against it is what differentiates the maturity profile and produces the quasi-exponential condition. And it is the right coordinate for any question about stationarity, since a model with time-homogeneous $\bar{\sigma}$ is one whose dynamics look the same on every date, which is not visible in the maturity-date coordinate at all.

What is left to choose

If $\sigma(t,T)$ is deterministic then (12.1) makes $f(t,T)$ Gaussian and bond prices lognormal, which is the Gaussian HJM class and the setting for everything in this chapter.

One Brownian motion has been used for the whole curve, and that is a restriction rather than a simplification. Every point of the curve is driven by the same dW_t, in proportion to its own $\sigma(t,T)$, so the increments of any two forward rates are perfectly correlated: the curve can shift and steepen as it shifts, but it cannot twist. Chapter 12's calculation How much a factor costs measures what that costs on a spread option and shows the cost is large. Relaxing it means several Brownian motions with their own volatility term structures, and (12.1) generalises with a sum over factors and nothing else changed.

8.3 Hull-White Model

Consider one-factor HJM diffusion. Instrument pricing is much more tractable if we restrict the volatility term-structure to be deterministic, separable and of the form:

$$\sigma(t,T) = \sigma(t)e^{-\int_t^T \alpha(s)ds}$$

Hull-White in general permits both $\alpha(s)$ and $\sigma(t)$ to be time-dependent. We take both to be constant, $\alpha(s) = \alpha$ and $\sigma(t) = \sigma$, which is the version usually meant by the Hull-White (or extended Vasicek) model and the version used in practice, since a fitted $\sigma(t)$ makes the future volatility term structure implied by the model non-stationary. Nothing structural in what follows depends on σ being constant, only on $\sigma(t,T)$ being deterministic and separable, so the derivation goes through for a deterministic $\sigma(t)$ with the closed forms replaced by integrals against σ^2.

In this case,

$$\begin{aligned}
df(t,T) =& \sigma e^{-\alpha(T-t)}\left(\int_t^T \sigma e^{-\alpha(u-t)}du\right)dt + \sigma e^{-\alpha(T-t)}dW_t \\
=& \sigma^2 e^{-\alpha(T-t)}\left(\int_t^T e^{-\alpha(u-t)}du\right)dt + \sigma e^{-\alpha(T-t)}dW_t \\
=& \frac{\sigma^2}{\alpha}e^{-\alpha(T-t)}\left(1-e^{-\alpha(T-t)}\right)dt + \sigma e^{-\alpha(T-t)}dW_t
\end{aligned}$$

$$\begin{aligned}
f(t,T) =& f(0,T) + \int_0^t \frac{\sigma^2}{\alpha}e^{-\alpha(T-u)}\left(1-e^{-\alpha(T-u)}\right)du + \int_0^t \sigma e^{-\alpha(T-u)}dW_u \\
=& f(0,T) + \frac{\sigma^2}{\alpha}\int_0^t \left(e^{-\alpha(T-u)} - e^{-2\alpha(T-u)}\right)du + \sigma\int_0^t e^{-\alpha(T-u)}dW_u \\
=& f(0,T) + \frac{\sigma^2}{\alpha^2}\left(e^{-\alpha(T-t)} - e^{-\alpha T} - \frac{1}{2}(e^{-2\alpha(T-t)} - e^{-2\alpha T})\right) + \sigma\int_0^t e^{-\alpha(T-u)}dW_u
\end{aligned}$$

Using the Itô isometry, $f(t,T)$ therefore is normally distributed as

$$f(t,T) \sim N\left(f(0,T) + \frac{\sigma^2}{\alpha^2}\left(\frac{e^{\alpha t}-1}{e^{\alpha T}} - \frac{e^{2\alpha t}-1}{2e^{2\alpha T}}\right), \frac{\sigma^2}{2\alpha}\frac{e^{2\alpha t}-1}{e^{2\alpha T}}\right)$$

where $N(\mu,\sigma^2)$ is a normal distribution with mean μ and variance σ^2.
We then have

$$\begin{aligned}
\int_t^T f(t,s)ds &= \int_t^T \left(f(0,s) + \frac{\sigma^2}{\alpha^2}\left(\frac{e^{\alpha t}-1}{e^{\alpha s}} - \frac{e^{2\alpha t}-1}{2e^{2\alpha s}}\right)\right)ds + \sigma\int_t^T\int_0^t e^{-\alpha(s-u)}dW_u ds \\
&= \int_t^T f(0,s)ds + \frac{\sigma^2}{\alpha^3}\left((e^{\alpha t}-1)(e^{-\alpha t}-e^{-\alpha T}) - \frac{1}{4}(e^{2\alpha t}-1)(e^{-2\alpha t}-e^{-2\alpha T})\right) \\
&\quad + \sigma\int_t^T\int_0^t e^{-\alpha(s-u)}dW_u ds \\
&= \int_t^T f(0,s)ds + \frac{\sigma^2}{\alpha^3}\left((1-e^{-\alpha t})(1-e^{-\alpha(T-t)}) - \frac{1}{4}(1-e^{-2\alpha t})(1-e^{-2\alpha(T-t)})\right) \\
&\quad + \sigma\int_t^T\int_0^t e^{-\alpha(s-u)}dW_u ds \\
&= \int_t^T f(0,s)ds + \frac{\sigma^2}{\alpha^3}\left((1-e^{-\alpha t})(1-e^{-\alpha(T-t)}) - \frac{1}{4}(1-e^{-2\alpha t})(1-e^{-2\alpha(T-t)})\right) \\
&\quad + \frac{\sigma}{\alpha}(e^{-\alpha t}-e^{-\alpha T})\int_0^t e^{\alpha u}dW_u
\end{aligned}$$

Therefore,

$$\int_t^T f(t,s)ds \sim N\Big(\int_t^T f(0,s)ds + \frac{\sigma^2}{\alpha^3}\big((1-e^{-\alpha t})(1-e^{-\alpha(T-t)}) - \tfrac{1}{4}(1-e^{-2\alpha t})(1-e^{-2\alpha(T-t)})\big), \\ \frac{\sigma^2}{2\alpha^3}(1-e^{-\alpha(T-t)})^2(1-e^{-2\alpha t})\Big) \tag{8.8}$$

and the bond price

$$P(t,T) = e^{-\int_t^T f(t,s)ds}$$

is log-normally distributed.

The Hull-White model can be specified completely in terms of the short rate diffusion as well, and is therefore called a short-rate model. The re-parametrisation goes as follows.

$$\begin{aligned} r(t) =& f(t,t) \\ =& f(0,t) + \frac{\sigma^2}{\alpha^2}\left(1 - e^{-\alpha t} - \frac{1}{2}(1-e^{-2\alpha t})\right) + \sigma\int_0^t e^{-\alpha(t-u)}dW_u \\ =& f(0,t) + \frac{\sigma^2}{\alpha^2}\left(1 - e^{-\alpha t} - \frac{1}{2}(1-e^{-2\alpha t})\right) + \sigma e^{-\alpha t}\int_0^t e^{\alpha u}dW_u \\ =& f(0,t) + \frac{\sigma^2}{2\alpha^2}\left(1-e^{-\alpha t}\right)^2 + \sigma e^{-\alpha t}\int_0^t e^{\alpha u}dW_u \end{aligned}$$

Notice here that the stochastic evolution of both $\int_t^T f(t,s)ds$ and $r(t)$ is determined by the term $\int_0^t e^{\alpha u}dW_u$ and therefore the bond yields $\ln P(t,T)$ can be specified as a linear function of $r(t)$ with coefficients of this linear relationship being deterministic functions of t and T. Models with such linear relationship between short rate and yield term structure are known as affine term structure models.

$$\begin{aligned} P(t,T) =& \exp\left(A_1(t,T) + B_1(t,T)\int_0^t e^{\alpha u}dW_u\right) \\ =& \exp\left(A_2(t,T) + B_2(t,T)f(t,T)\right) \\ =& \exp\left(A_3(t,T) + B_3(t,T)r(t)\right) \end{aligned}$$

where $A_i(t,T)$ and $B_i(t,T)$ are deterministic functions of t and T.

Differentiating the short rate diffusion,

$$\begin{aligned} dr(t) =& \left(\frac{\partial}{\partial t} f(0,t) + \frac{\sigma^2}{\alpha}(e^{-\alpha t} - e^{-2\alpha t}) - \sigma\alpha e^{-\alpha t} \int_0^t e^{\alpha u} dW_u \right) dt + \sigma e^{-\alpha t} e^{\alpha t} dW_t \\ =& \left(\frac{\partial}{\partial t} f(0,t) + \frac{\sigma^2}{\alpha}(e^{-\alpha t} - e^{-2\alpha t}) - \alpha \left(r(t) - (f(0,t) + \frac{\sigma^2}{2\alpha^2}\big(1 - e^{-\alpha t}\big)^2) \right) \right) dt + \sigma dW_t \\ =& \left(\frac{\partial}{\partial t} f(0,t) + \frac{\sigma^2}{\alpha}(e^{-\alpha t} - e^{-2\alpha t}) + \alpha(f(0,t) + \frac{\sigma^2}{2\alpha^2}\big(1 - e^{-\alpha t}\big)^2) - \alpha r(t) \right) dt + \sigma dW_t \\ =& \big(\theta(t) - \alpha r(t)\big) dt + \sigma dW_t \end{aligned}$$

where $\theta(t)$ is specified in terms of α, σ and the initial forward curve $f(0,\cdot)$. From this specification, α has another reading, as the mean-reversion strength of the short rate.

Structure (What the reduction buys)**.** Every quantity above — the forward rate, the short rate, the bond price, the drift $\theta(t)$ — came out of the single restriction on $\sigma(t,T)$ at the top of this section, and none of it needed the drift condition (12.1) solved again from scratch. That is the whole appeal of the model: the short rate is Gaussian and mean-reverting by construction, $\theta(t)$ is chosen so the bond prices it implies match today's curve exactly, and every bond price is log-normal, so a swaption or cap has a closed form. What is bought is one Brownian motion's worth of dynamics, exactly as flexible as (12.1) allows and no more — the same restriction chapter 12's How much a factor costs measured above, priced there on a spread option.

8.4 Future-Forward Basis

Here is the first place a model is unavoidable. Chapter 7 priced a forward rate agreement off today's curve alone, with no assumption about how rates move. A futures contract on the same rate cannot be priced that way, and the difference between the two prices is a pure model quantity. We work it through for LIBOR futures and forwards. LIBOR itself has been decommissioned, but the calculation is the instructive one and the logic transfers unchanged to futures on other benchmark rates.

Recall from chapter 7 the reference rate

$$L(t,T,T+\delta) = \frac{1}{\delta}\left(\frac{P(t,T)}{P(t,T+\delta)} - 1 \right),$$

and the forward rate agreement, which pays $\delta(L(T,T,T+\delta) - R)/(1+\delta L(T,T,T+\delta))$ at T. We write FRA for the value of R making it worth zero. First we confirm what chapter 7 asserted, that this is the forward rate, and identify the measure in which that is the natural statement.

$$\begin{aligned}
0 =& P(0,T)\mathbb{E}^T[\frac{\delta(L(T,T,T+\delta)-\text{FRA}(0))}{1+\delta L(T,T,T+\delta)}] \\
=& P(0,T)\mathbb{E}^T[\delta(L(T,T,T+\delta)-\text{FRA}(0))P(T,T+\delta)] \\
=& P(0,T)\mathbb{E}^{T+\delta}[\delta(L(T,T,T+\delta)-\text{FRA}(0))P(T,T+\delta)\frac{P(T,T)P(0,T+\delta)}{P(0,T)P(T,T+\delta)}] \\
=& P(0,T+\delta)\mathbb{E}^{T+\delta}[\delta(L(T,T,T+\delta)-\text{FRA}(0))] \\
\text{FRA}(0) =& \mathbb{E}^{T+\delta}[L(T,T,T+\delta)]
\end{aligned}$$

Notice that $L(t,T,T+\delta)$ is a ratio of two prices plus a constant, and scaled by a constant. Therefore it is a martingale in the measure with $P(t,T)$ as numeraire i.e. T-forward measure. Therefore

$$\text{FRA}(t) = \mathbb{E}_t^{T+\delta}[L(T,T,T+\delta)] = L(t,T,T+\delta)$$

is known as forward LIBOR rate.

Definition 8.4 (Eurodollar Futures)**.** Eurodollar futures are daily margined contracts that at maturity T settle at value $1 - L(T,T+\delta)$ which is published at T.

Eurodollar futures are settled in market at value $1 - L(T,T+\delta)$ but for ease of comparison we will take futures value at T $\text{Fut}(T)$ to be $L(T,T+\delta)$. We can get the true futures value from $1-\text{Fut}(0)$. Nothing is lost in the substitution: the margin condition below sets a price to zero and sees the quote only through $\frac{\partial}{\partial s}\text{Fut}(s)$, so it is blind to both the sign and the additive constant, and $\mathbb{E}^{\mathbb{P}}[1-L] = 1-\mathbb{E}^{\mathbb{P}}[L]$ carries the martingale property between the two conventions. That is a fact about $L \mapsto 1-L$ being affine, and it fails for a contract settling on anything else — the Treasury futures of chapter 22 settle against a bond price, convex in its yield, so there the implied yield is not a martingale even though the price is.

Futures contracts on LIBOR are different from forwards in two major ways.

1. Futures contracts are daily margined. That is, the buyer does not pay for the contract initially, except for initial margin which is usually much smaller than the futures price. Then the movement in futures prices are settled daily; if the futures price goes up, the buyer has cash put in his margin account and if the futures price goes down, the buyer has cash withdrawn from this account.

2. Futures contracts settle at T, the start of the period, for the rate that has just fixed, and *without discounting*. A forward rate agreement also settles at T by market convention — but with the $1/(1+\delta L)$ factor of chapter 7, which is exactly $P(T,T+\delta)$, so its payoff is worth the same as $\delta(L-R)$ paid at $T+\delta$. The calculation below shows that equivalence rather than assuming it.

The futures price $\text{Fut}(t)$ is such that price of future cashflows from daily margining is 0 i.e. for all t

$$
\begin{aligned}
0 =& \frac{1}{D_t}\mathbb{E}_t^{\mathbb{P}}\left[\int_t^T D_s \frac{\partial}{\partial s}\mathrm{Fut}(s)ds\right] \\
0 =& \mathbb{E}_t^{\mathbb{P}}\left[\int_t^T D_s \frac{\partial}{\partial s}\mathrm{Fut}(s)ds\right]
\end{aligned}
$$

where $\frac{\partial}{\partial s}\mathrm{Fut}(s)ds$ is the cashflow from margin in the time interval $(s, s+ds)$. Differentiating with respect to t,

$$
\begin{aligned}
0 =& \mathbb{E}_t^{\mathbb{P}}\left[-D_t\frac{\partial}{\partial t}\mathrm{Fut}(t)\right] \\
=& -D_t\mathbb{E}_t^{\mathbb{P}}\left[\frac{\partial}{\partial t}\mathrm{Fut}(t)\right] \\
0 =& \mathbb{E}_t^{\mathbb{P}}\left[\frac{\partial}{\partial t}\mathrm{Fut}(t)\right]
\end{aligned}
$$

Integrating and using iterated conditioning, we get

$$
\begin{aligned}
\mathrm{Fut}(0) =& \mathbb{E}_0^{\mathbb{P}}[\mathrm{Fut}(T)] \\
=& \mathbb{E}_0^{\mathbb{P}}[L(T,T,T+\delta)]
\end{aligned}
$$

which says that the current futures price is the expected settlement price of the future in risk neutral measure.

Remark (Quoting in price or in rate)**.** The basis derived below is positive in rate terms, so the futures rate sits above the forward rate; in the market's price quote the ordering reverses and the future trades below the forward.

The future-forward basis is largely sensitive to the covariance between rate fixing and money market discount factor.

$$
\begin{aligned}
\mathrm{Fut}(0) - \mathrm{FRA}(0) =& \mathbb{E}^{\mathbb{P}}[L(T,T,T+\delta)] - \mathbb{E}^{T+\delta}[L(T,T,T+\delta)] \\
=& \mathbb{E}^{\mathbb{P}}[L(T,T,T+\delta)] - \mathbb{E}^{\mathbb{P}}[L(T,T,T+\delta)\frac{P(T,T+\delta)D_T}{P(0,T+\delta)D_0}] \\
=& \frac{P(0,T+\delta)\mathbb{E}^{\mathbb{P}}[L(T,T,T+\delta)] - \mathbb{E}^{\mathbb{P}}[L(T,T,T+\delta)P(T,T+\delta)D_T]}{P(0,T+\delta)} \\
=& \frac{\mathbb{E}^{\mathbb{P}}[P(T,T+\delta)D_T]\mathbb{E}^{\mathbb{P}}[L(T,T,T+\delta)] - \mathbb{E}^{\mathbb{P}}[L(T,T,T+\delta)P(T,T+\delta)D_T]}{P(0,T+\delta)} \\
=& -\frac{\mathrm{Cov}^{\mathbb{P}}(P(T,T+\delta)D_T, L(T,T,T+\delta))}{P(0,T+\delta)} \\
=& -\frac{\mathrm{Cov}^{\mathbb{P}}(P(T,T+\delta)D_T, \frac{1}{P(T,T+\delta)})}{\delta P(0,T+\delta)}
\end{aligned}
$$

where $\mathrm{Cov}^{\mathbb{P}}(X,Y)$ is the covariance of X and Y in risk neutral measure.

Note the sign. The covariance itself is negative: when rates rise the fixing $L(T,T,T+\delta)$ rises, while both the bond price $P(T,T+\delta)$ and the money market discount factor D_T fall, so the Radon-Nikodym density $P(T,T+\delta)D_T/P(0,T+\delta)$ moves against the fixing. With the leading minus sign the basis is therefore positive, and the futures rate sits above the forward rate. This is the precise sense in which daily margining is worth something to the buyer of a future: the contract pays the fixing undiscounted, so the cash it throws off is largest exactly when it is worth least to reinvest.

Worked case 8.5 (The basis in the Hull-White model)**.** The covariance above is exact but is not a number until a model is chosen. Hull-White is the model this chapter has already built, so the computation is available without anything new, and it shows how a basis is extracted once a model is fixed. The theory does not rest on it: a reader may take expression 8.9 off the last line and carry on.

Both legs are simply compounded yields, $\delta L(T,T,T+\delta) = 1/P(T,T+\delta)-1$ and $\delta L(0,T,T+\delta) = P(0,T)/P(0,T+\delta)-1$, and $\mathbb{E}^{T+\delta}[L(T,T,T+\delta)] = L(0,T,T+\delta)$ because the forward rate is a martingale in its own forward measure. The constants cancel and

$$\mathrm{Fut}(0) - \mathrm{FRA}(0) = \frac{1}{\delta}\left(\mathbb{E}^{\mathbb{P}}\Big[\frac{1}{P(T,T+\delta)}\Big] - \frac{P(0,T)}{P(0,T+\delta)}\right),$$

so everything turns on $\mathbb{E}^{\mathbb{P}}[1/P(T,T+\delta)] = \mathbb{E}^{\mathbb{P}}[e^{\int_T^{T+\delta} f(T,s)ds}]$. Expression 8.8 makes the exponent normal with mean μ and variance V, and a lognormal has mean $e^{\mu+V/2}$. Setting $t = T$ there, replacing the upper limit T by $T+\delta$, and abbreviating $u = 1-e^{-\alpha T}$ and $v = 1-e^{-\alpha\delta}$ so that $1-e^{-2\alpha T} = u(2-u)$ and $1-e^{-2\alpha\delta} = v(2-v)$,

$$\mu = \int_T^{T+\delta} f(0,s)ds + M, \qquad M = \frac{\sigma^2}{4\alpha^3}uv\big(4-(2-u)(2-v)\big) = \frac{\sigma^2}{4\alpha^3}uv\big(2u+2v-uv\big)$$
$$V = \frac{\sigma^2}{2\alpha^3}u(2-u)v^2$$

where M names the part of the mean that today's curve does not already contain. The curve supplies the rest, since $P(0,T)/P(0,T+\delta) = e^{\int_T^{T+\delta} f(0,s)ds} = 1+\delta L(0,T,T+\delta)$, and therefore

$$\mathrm{Fut}(0) - \mathrm{FRA}(0) = \frac{1}{\delta}\frac{P(0,T)}{P(0,T+\delta)}\big(e^{M+\frac{V}{2}}-1\big) = \left(\frac{1}{\delta}+L(0,T,T+\delta)\right)\big(e^{M+\frac{V}{2}}-1\big).$$

Since $u,v \in (0,1)$ both $2u+2v-uv$ and $u(2-u)$ are positive, so M and V are, and the basis is strictly positive: the futures rate always sits above the forward rate, as the covariance argument above anticipated.

The Ho-Lee model is the $\alpha \to 0$ limit, in which the forward rate volatility $\sigma(t,T) = \sigma$ no longer decays with maturity. Then $u \to \alpha T$ and $v \to \alpha\delta$, so $M \to \frac{\sigma^2}{2} T\delta(T+\delta)$ and $V \to \sigma^2 T\delta^2$, giving $M + \frac{V}{2} \to \frac{\sigma^2}{2} T\delta(T+2\delta)$. Both are of order $\sigma^2 T^2$ and hence small, so $e^{M+V/2} - 1 \approx M + V/2$; and $\delta L(0,T,T+\delta) \ll 1$, so the leading factor is approximately $1/\delta$. The Ho-Lee basis is therefore

$$\text{Fut}(0) - \text{FRA}(0) \approx \frac{\sigma^2}{2} T(T+2\delta) \tag{8.9}$$

Remark. The basis is often quoted in textbooks as $\frac{1}{2}\sigma^2 T_1 T_2$ with $T_1 = T$ and $T_2 = T+\delta$, which is not what expression 8.9 says. The discrepancy is instructive, and it is entirely a question of which compounding convention the rate is quoted in.

The term M on its own measures the drift of the yield $\int_T^{T+\delta} f(T,s)ds$ away from today's forward yield, i.e. the basis on the *continuously compounded* rate $R_c = -\frac{1}{\delta}\ln P(T,T+\delta)$:

$$\mathbb{E}^{\mathbb{P}}[R_c] - \frac{1}{\delta}\int_T^{T+\delta} f(0,s)ds = \frac{M}{\delta} \to \frac{\sigma^2}{2} T(T+\delta) = \frac{\sigma^2}{2} T_1 T_2$$

This is exactly the textbook expression. In Hull-White it can be written in the familiar form, with $B(t,T) = \frac{1}{\alpha}(1 - e^{-\alpha(T-t)})$ so that $B(0,T) = u/\alpha$ and $B(T,T+\delta) = v/\alpha$,

$$\frac{M}{\delta} = \frac{\sigma^2}{4\alpha\delta} B(T,T+\delta)\Big(B(T,T+\delta)\big(1 - e^{-2\alpha T}\big) + 2\alpha B(0,T)^2 \Big)$$

LIBOR, however, is simply compounded, and $\frac{1}{\delta}(\frac{1}{P(T,T+\delta)} - 1)$ is a convex function of the yield. That convexity contributes a further $\frac{1}{2}\sigma^2 T\delta$ to the basis and turns $T+\delta$ into $T+2\delta$. The two agree to relative order δ/T, so for a three month rate the textbook formula understates the LIBOR basis by about 11% at $T = 2$ and about 5% at $T = 5$; mean reversion pulls in the opposite direction and is the larger correction at realistic α.

The picture explains why futures are used in the belly of the curve and swaps beyond it, as chapter 7 described without saying why. The convexity correction is negligible where the futures are liquid and becomes both large and model-dependent exactly where they stop being used.

8.5 Swaptions

Chapter 7 introduced the swap, its annuity

$$\text{Ann}(t) = \Sigma_{i=1}^{N_r} \delta_i^r P(t, \tau_i^r),$$

and the par swap rate

$$S(t) = \frac{P(t,\tau_0) - P(t,\tau_{N_p}^p)}{\text{Ann}(t)},$$

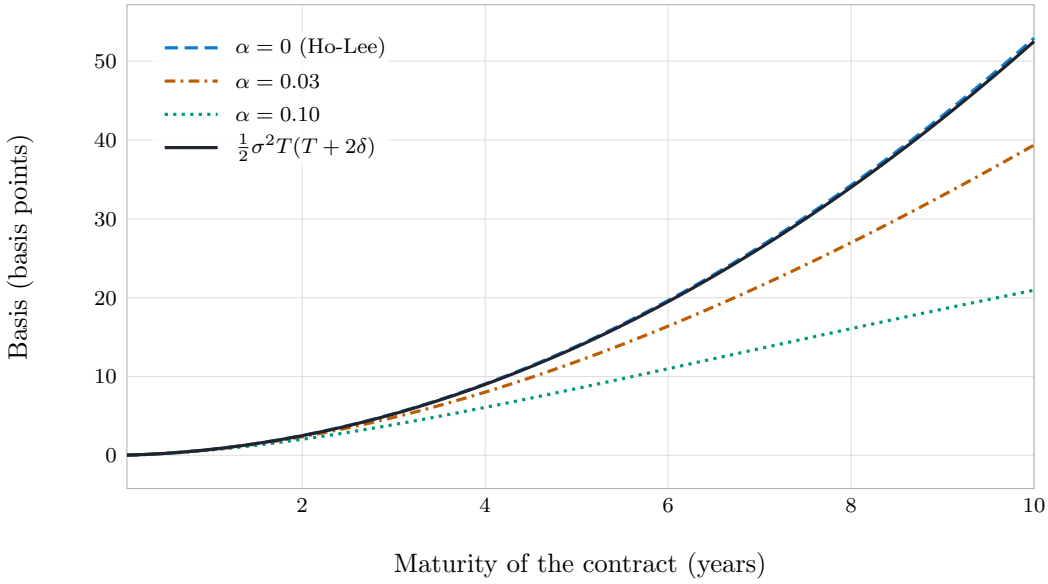

Figure 8.1: The basis against the maturity of the contract, for a three month rate. It grows like $T(T + 2\delta)$ rather than like T^2 — doubling the maturity from one year to two multiplies it by 3.3, not by 4, because the accrual period is still a quarter of the maturity — and only well beyond the accrual does the growth look quadratic. Mean reversion pulls the curve down, and increasingly so with maturity, since it is the accumulated decay that matters. At the front end the basis is under a basis point and can be ignored; at ten years it is not.

Drawn by `basis` and `ho_lee_basis`.

together with the observations that the annuity is a price and is strictly positive, so it is a numeraire, and the swap rate is a ratio of a price to it, so by chapter 6

$$S(t) = \mathbb{E}_t^{\text{Ann}}[S(T)]$$

is a martingale in the annuity measure. We also had the value of a receiver swap struck at R as $\text{Ann}(0)(R - S(0))$, and of a payer swap as $\text{Ann}(0)(S(0) - R)$.

None of that needed a model. What follows does, and the annuity measure is what makes it manageable.

Definition 8.6 (Swaption)**.** A swaption is an option to enter a swap with swaption strike as the swap fixed rate. A call option is an option to enter a payer swap and a put option is an option to enter a receiver swap. The swap start date is the option exercise date.

Price of call swaption with exercise date T and strike K is therefore

$$\begin{aligned}
\text{Swo}(0) =& P(0,T)\mathbb{E}^T[\max(\text{Ann}(T)(S(T) - K), 0)] \\
=& P(0,T)\mathbb{E}^{\text{Ann}}[\max(\text{Ann}(T)(S(T) - K), 0)\frac{\text{Ann}(0)P(T,T)}{\text{Ann}(T)P(0,T)}] \\
=& \text{Ann}(0)\mathbb{E}^{\text{Ann}}[\max(S(T) - K, 0)] \\
=& \text{Ann}(0)\mathbb{E}^{\text{Ann}}[(S(T) - K)^+]
\end{aligned}$$

The same swaption inside a one-factor model

That was the market's calculation. A model has to produce the same price from its own dynamics, and for a one-factor short rate model there is a decomposition that makes it a closed form rather than a numerical integral. It is Jamshidian's.

Start by seeing the swaption as an option on a bond. At expiry T the floating leg of the swap is worth par, so entering a receiver struck at K is worth

$$\sum_i c_i\, P(T,\tau_i) \;-\; 1, \qquad c_i = \delta_i K \;\; (i < N), \qquad c_N = 1 + \delta_N K,$$

the difference between a coupon bond and par. A receiver swaption is therefore a call on that coupon bond struck at 1, and a payer swaption a put.

Theorem 8.7 (Jamshidian's decomposition)**.** *Suppose every discount bond at T is a strictly decreasing function of a single state x. Let x^* be the state at which the coupon bond equals the strike, and write $P_i^* = P(T,\tau_i)$ evaluated at x^*. Then, for every x,*

$$\Big(\sum_i c_i P(T,\tau_i) - 1\Big)^+ \;=\; \sum_i c_i \big(P(T,\tau_i) - P_i^*\big)^+. \tag{8.10}$$

An option on a coupon bond is a portfolio of options on the zeros.

Proof. The coupon bond is a positive combination of decreasing functions and so is decreasing, which gives a unique x^* and, by definition, $\sum_i c_i P_i^* = 1$.

Now fix x. If $x < x^*$ then every bond exceeds its own value at x^*, since each is decreasing, so every bracket on the right is positive and the sum is $\sum_i c_i P(T, \tau_i) - \sum_i c_i P_i^*$, which is the left side. If $x > x^*$ then every bracket is negative, the right side is zero, and the coupon bond is below par so the left side is zero too. The two sides agree at every state, not merely in expectation.[3] □

Each term of (8.10) is an option on a single discount bond, and in a Gaussian model the bond is lognormal under the appropriate forward measure, so each is a Black formula. The swaption is a sum of them.

Remark (The hypothesis is the one factor, and it is also the defect)**.** Everything rested on the bonds being decreasing functions of *one* number, which is what makes x^* a point.

With two factors the exercise boundary is a curve rather than a point. Two states on it give the same coupon bond value and different individual bond prices, so no single set of strikes P_i^* can serve — whichever state supplied them, the other would decompose wrongly.[4] The decomposition is unavailable, and a multi-factor model prices its swaptions by simulation or by a grid.

This says something about the trick. It works because the model can only move the whole curve as a function of one number — the very property chapter 12's calculation How much a factor costs measures the cost of, and the reason a one-factor model misprices anything sensitive to the curve's shape. The closed form is not a reward for a good model. It is a consequence of the restriction that makes the model wrong.

Remark (Where this stops)**.** The swaption is priced and no smile has been needed: one volatility, one strike, Black's formula in the annuity measure.

References

- Kirikos, G., & Novak, D. (1997). Convexity Conundrums: Presenting a treatment of swap convexity in the Hull-White framework. *Risk*, 10, 60–61.

- Schlenkrich, S., & Ursachi, I. (2015). Multi-Curve Pricing of Non-Standard Tenor Vanilla Options. Available at SSRN 2695011.

- Heath, D., Jarrow, R., & Morton, A. (1992). Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation. *Econometrica*, 60(1), 77–105.

- Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, 5(2), 177–188.

- Hull, J., & White, A. (1990). Pricing interest-rate-derivative securities. *Review of Financial Studies*, 3(4), 573–592.

[3] `checked pointwise across the state.`

[4] `measured.`

- Jamshidian, F. (1989). An exact bond option formula. *Journal of Finance*, 44(1), 205–209. The decomposition of Theorem 8.7. Not to be confused with the same author's 1997 paper, cited in chapter 13 for the spot Libor measure, which is an unrelated result.

Chapter 9

Implied and Local Volatility

We take seriously the fact that the market does not price options the way chapter 5 said it should. We show that the option prices at one expiry are the same information as the probability distribution of the underlying at that expiry, that there is always a model with a single deterministic volatility function reproducing every one of those prices, and we derive the formula for it. We finish with Gyongi's theorem, which explains both why that model exists and why it is not the end of the story.

Where the Curve Went

Chapter 8 put the whole curve in motion and priced a swaption from it, and the last thing it did was reduce that swaption to Black's formula: one volatility, one strike, in the annuity measure. That reduction is exact, and it is what the next three chapters stand on. A swap rate is a martingale under its own annuity and a forward rate is one under its own forward measure, so once the numeraire is chosen, a caplet or a swaption is an option on *one* underlying at *one* date — the same object as an equity option, with the curve already integrated into the numeraire.

So these chapters are not ignoring the term structure. They are working in the measure in which it has already been dealt with.

Structure (Marginal and joint, and why the marginals never pin down the joint)**.** There are two distinct things a rates model has to get right.

The first is the *marginal* question. For one rate, at one date, what is the distribution? This is what a smile is, entirely — the Breeden-Litzenberger theorem below says that the prices at one expiry *are* that distribution, with nothing left over. It needs no curve, because a numeraire has removed it, and it is the subject of this chapter, chapter 10 and chapter 11.

The second is the *joint* question. How do several rates, at several dates, move together? A swaption on the ten-year rate says nothing about how the ten-year moves relative to the two-year, and no collection of single-rate smiles ever will: they are marginals, and marginals do not determine

a joint distribution. That is the subject of chapter 12 and chapter 13, and chapter 18 is about the general fact underneath it.

The separation is real rather than expositional. It is why a CMS caplet needs one rate's smile and no dynamics at all, as chapter 15 shows by replicating it statically, while a Bermudan needs the joint law and cannot be replicated from vanillas at any price. It is also why the two halves are calibrated to different instruments and, on most desks, by different people.

What the separation costs is stated in chapter 10: a smile fixes the marginal and leaves the dynamics free, so several models agreeing on every quoted price today can disagree about tomorrow. Fitting the first question perfectly does not begin to answer the second.

One convention, since it saves writing every statement twice. These three chapters say "spot", "forward" and "strike" because the results hold for any underlying, and the equity notation is the one they were discovered in. For a rates reader the translation is fixed: read the underlying as a forward rate under its forward measure or a swap rate under its annuity, and read a discounted expectation as one taken under that numeraire. Nothing else changes; chapter 12 picks the curve back up where this leaves it.

9.1 Implied Volatility

Chapter 5 gave us a formula. Given the forward F, a strike K, an expiry T and a volatility σ, it returns a price. Everything on that list except σ can be read off a contract or a screen. So a trader quoting an option is really quoting one number, and rather than quote it as a price, the market quotes it as the volatility that would produce that price.

Definition 9.1 (Implied volatility)**.** The implied volatility of an option with observed price V is the number $\sigma_{\text{imp}}(K,T)$ satisfying

$$V = \text{Black}\big(F, K, \sigma_{\text{imp}}(K,T), T\big).$$

Before using this we should check it makes sense.

Lemma 9.2. *For a price strictly between its no-arbitrage bounds, the implied volatility exists and is unique.*

Proof. The Black price is continuous in σ and its derivative, the vega, is

$$\frac{\partial V}{\partial \sigma} = F\, n(d_1)\sqrt{T} > 0$$

for $F, T > 0$, where n is the standard normal density. A continuous, strictly increasing function is invertible on its range. As $\sigma \to 0$ the price tends to the intrinsic value $(F-K)^+$ and as $\sigma \to \infty$ it tends to F, so the range is exactly the interval between the bounds. □

In words: turning the volatility dial always moves the price, and always in the same direction, so from a price you can always work backwards to exactly one volatility. That is all a quoting convention needs. Notice what the lemma does *not* say. It says nothing about whether the underlying is

lognormal, or whether any model is right. It says only that this particular formula is a reversible way of writing down a number. The implementation is `implied_vol_black76` in `quant/src/black.rs`, and it does exactly what the proof does: it brackets the answer and squeezes.

Now the observation that this whole part of the notes exists to explain. If the Black-Scholes model were literally true, the same σ would appear in every option on the same underlying, whatever its strike. So plotting implied volatility against strike would give a horizontal line. It does not. What one sees instead is called the smile, or, when it slopes, the skew.

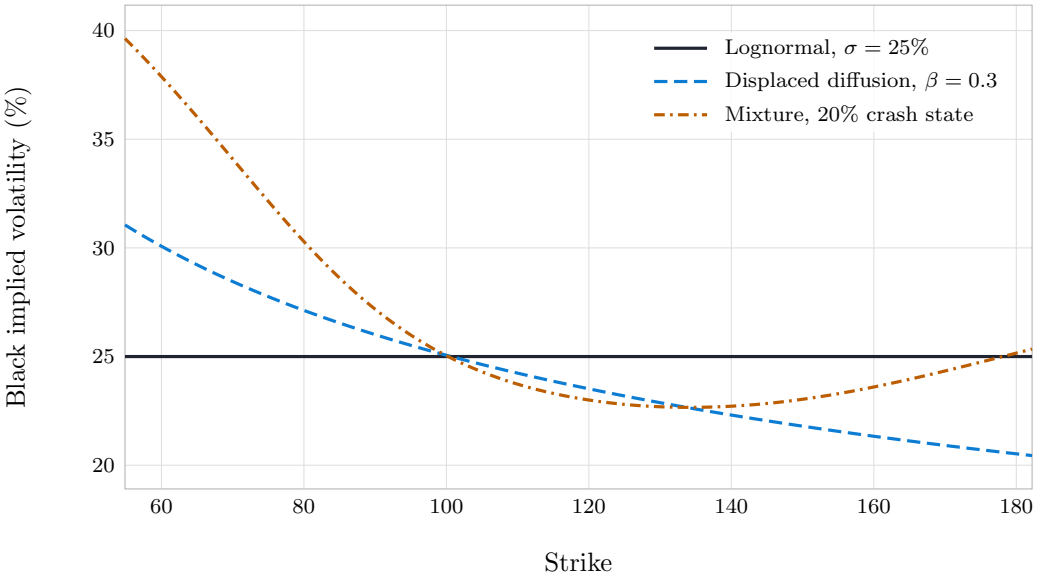

Figure 9.1: The implied volatility of three models, all with the same at-the-money volatility. A lognormal forward gives a flat line at every setting of every parameter, and could not do otherwise: implied volatility is defined by inverting the lognormal formula, so lognormal prices invert to one number. The other two are the two smallest departures from it, and they produce the two features every real market shows — a tilt and a curvature.

Drawn by `Model::price`, `Model::implied_vol` and `implied_vol_black76`.

Play with the figure for a moment, because it makes a distinction that is easy to lose. The displaced diffusion tilts the smile but leaves it nearly straight; the mixture bends it. These are different statements about the world. Tilt says that a fall in the forward is more violent than a rise. Curvature says that large moves in either direction are likelier than the lognormal model allows. Real markets show both, and a model that produces only one of them will be wrong about half of the options it is asked to price.

9.2 The Prices Are the Distribution

If the smile is a map of something, what exactly is it a map of? The answer is completely precise, and it is the first theorem of this chapter.

Fix an expiry T and work in the T-forward measure of chapter 6, in which the forward F_T is a

martingale and a call is worth

$$C(K) = P(0,T)\,\mathbb{E}^T\big[(F_T - K)^+\big]. \tag{9.1}$$

Theorem 9.3 (Breeden-Litzenberger)**.** *Let p be the probability density of F_T in the T-forward measure. Then*

$$\frac{\partial C}{\partial K} = -P(0,T)\,\mathbb{P}^T(F_T > K), \qquad \frac{\partial^2 C}{\partial K^2} = P(0,T)\,p(K).$$

Proof. Write (9.1) as an integral and differentiate under it:

$$C(K) = P(0,T)\int_0^\infty (x-K)^+ p(x)\,dx = P(0,T)\int_K^\infty (x-K)p(x)\,dx$$
$$\frac{\partial C}{\partial K} = P(0,T)\bigg(-\big[(x-K)p(x)\big]_{x=K} - \int_K^\infty p(x)\,dx\bigg) = -P(0,T)\int_K^\infty p(x)\,dx,$$

the boundary term vanishing because $(x-K)$ is zero at $x=K$. That is the first claim. Differentiating once more,

$$\frac{\partial^2 C}{\partial K^2} = -P(0,T)\frac{\partial}{\partial K}\int_K^\infty p(x)\,dx = P(0,T)\,p(K). \qquad \square$$

Analytically: the second derivative of the call price with respect to strike is the density, discounted. Knowing the price of every strike at one expiry is therefore *exactly* knowing the distribution of the underlying at that date.

Financially: consider what you would have to trade to construct that second derivative. A second derivative is a second difference, so buy a call struck at $K-h$, sell two struck at K, and buy one struck at $K+h$. This is called a butterfly.

Example 9.1 (A butterfly is a bet on where the forward lands)**.** Take $h = \$1$ and $K = \$100$, so we buy the \$99 call, sell two \$100 calls, and buy the \$101 call. Its payoff at expiry, as a function of where F_T lands:

F_T	\$98	\$99.50	\$100	\$102
\$99 call	0	0.50	1.00	3.00
$-2\times$ \$100 call	0	0	0	−4.00
\$101 call	0	0	0	1.00
total	0	0.50	1.00	0

The payoff is zero outside $[\$99, \$101]$ and rises linearly to \$1 at the middle: a narrow tent pitched over K. It pays if and only if the forward lands near \$100, and it is worthless otherwise.

Now price it. The tent has height h and base $2h$, so its area is h^2, and for small h the expected payoff is approximately $p(K)h^2$. Its price is therefore about $P(0,T)p(K)h^2$; dividing by h^2 gives the theorem. So $\partial^2 C/\partial K^2$ is not merely a derivative that happens to equal a density. It is the price of the cheapest available bet that the forward finishes at K, and its price is the probability of that happening.

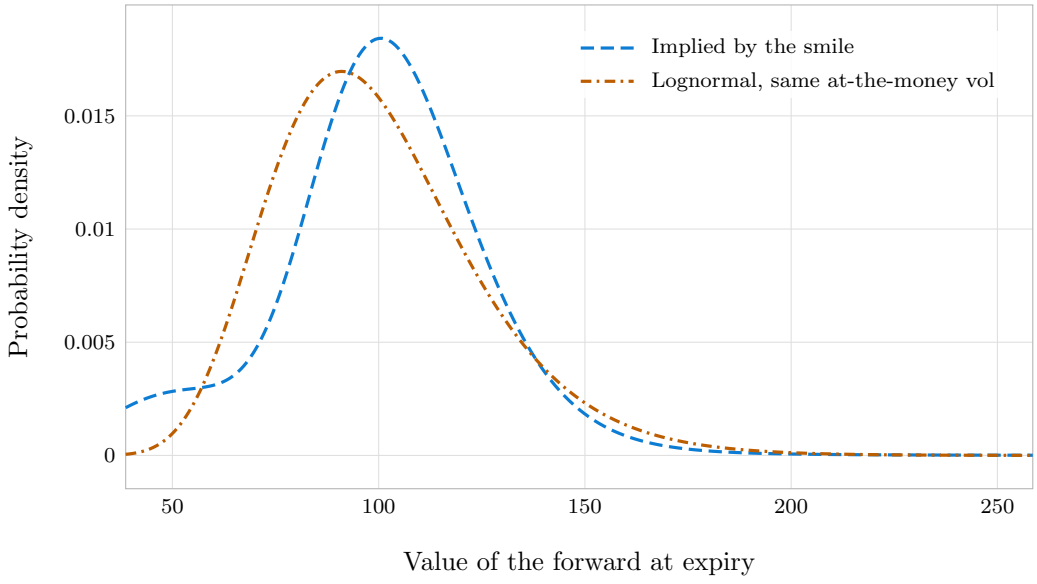

Figure 9.2: The density implied by a smile, next to the lognormal density with the same at-the-money volatility. Both are obtained by differencing option prices twice in the strike, so the picture is the theorem being applied rather than an illustration of it. The market's density has a fatter left tail, a thinner middle, and a slightly fatter right tail: the extra mass in the tails is what the raised wings of the smile are paying for, and it has to come from somewhere.

Drawn by `Model::density` and `lognormal_density`.

Practically: it tells you what a smile is a map of. It is a map of the market's probability distribution — and since a flat smile corresponds to the lognormal density, the shape of the smile is precisely the shape of the market's disagreement with lognormality.

The theorem also hands us the arbitrage constraints on a surface, for free. A density is non-negative, so $\partial^2 C/\partial K^2 \geq 0$: the call price must be *convex* in the strike. A probability is at most one, so $-\partial C/\partial K/P(0,T) \leq 1$. A quoted surface violating either of these is not merely ill-behaved, it is arbitrageable, and the trade that exploits it is the butterfly above — sold for a negative price.

Exercise (Finding the arbitrage)**.** Suppose three calls on the same expiry are quoted at $K = 99, 100, 101$ for \$6.10, \$5.50 and \$4.60. Show that the surface is not convex, construct the portfolio that captures the arbitrage, and state what it pays in every state of the world. Then check your answer against the bounds computed by `implied_vol_black76`, which returns no implied volatility at all for prices outside them.

9.3 What the Standard Combinations Price

The butterfly was not chosen for its convenience. It is a trade people actually put on, and so are its relatives, and each of them exists because it isolates one feature of the smile and is insensitive to the others. Setting them out makes the surface easier to read, because a quoted market is usually a market in these rather than in individual options.

Take the smile near the money and write it as a level, a slope and a curvature,

$$\sigma_{\text{imp}}(y) \approx \sigma_0 + s\,y + \tfrac{1}{2}c\,y^2, \qquad y = \ln\frac{K}{F},$$

and each combination picks off one coefficient.

- A *straddle* — a call and a put at the same strike — is delta neutral at the money and long vega. It has no exposure to s by symmetry, so it prices σ_0, the level.

- A *risk reversal* — long an out-of-the-money call, short the equally out-of-the-money put — is antisymmetric about the money and so cancels the level and the curvature, leaving s. It is the market's instrument for the skew, and it is quoted directly as a volatility difference: "the twenty-five delta risk reversal is at minus two" is a statement about s and nothing else.

- A *butterfly* cancels the level and the slope and leaves c. It prices the curvature, and, as the previous section showed, in the limit of narrow wings it prices the density itself.

- A *call spread* is the difference of two calls, so by Theorem 9.3 its price is $-\partial C/\partial K$ times the width: it is a digital, and it prices the *cumulative* distribution rather than the density.

- A *collar* is a risk reversal sized to cost nothing — long a put, short a call, strikes chosen so the premiums cancel. Its usefulness as a hedge is a consequence of the skew: in a market with $s < 0$ the put one wants is expensive and the call one gives up is cheap, so the strikes come out asymmetric, and how asymmetric is a direct reading of s.

Remark. Each combination is a finite-difference operator applied to the price in the strike: a call spread is a first difference, a butterfly a second, a risk reversal an antisymmetric first difference. By Breeden-Litzenberger the price surface *is* the distribution, so differencing it in strike reads off successive features of that distribution — the cumulative, then the density, then its shape. A market that quotes straddles, risk reversals and butterflies is quoting the first three moments of its own risk neutral distribution, in a coordinate system chosen so that each can be traded without disturbing the others.

9.4 How the Density Moves

Dupire's formula needs to know how the risk neutral density changes with maturity, and chapter 3 has already supplied it. The generator's adjoint was computed there, and the forward equation followed from the definition of the adjoint in two lines: the density satisfies $\partial p/\partial t = \mathcal{L}^* p$, with the coefficients inside the derivatives rather than outside.

Written out for the risk neutral price process of this chapter, whose drift is $(r-q)S$ and whose diffusion coefficient is $\sigma_{\text{loc}}(T,S)\,S$, that reads

$$\frac{\partial p}{\partial T} = -\frac{\partial}{\partial S}\big[(r-q)S\,p\big] + \frac{1}{2}\frac{\partial^2}{\partial S^2}\big[\sigma_{\text{loc}}^2(T,S)S^2 p\big]. \tag{9.2}$$

That is the only new thing the next section needs.

Remark (It is a conservation law)**.** (9.2) is chapter 3's probability flux, written out for this chapter's own drift and diffusion coefficient — probability moved rather than created or destroyed:

$$\frac{\partial p}{\partial T} = -\frac{\partial J}{\partial S}, \qquad J = (r-q)S\,p - \frac{1}{2}\frac{\partial}{\partial S}\big[\sigma_{\text{loc}}^2 S^2 p\big]. \tag{9.3}$$

That matters before this equation is used, because it is what makes it safe to solve numerically. Chapter 12 solves a forward equation on a grid, and the scheme in `QuasiGaussian::density` is written in the flux form (9.3) for exactly this reason: it moves mass between cells rather than evaluating a second derivative, so the total is conserved by construction rather than by hope. Another version evaluated the derivative directly, and returned a "density" integrating to 1.42.

9.5 Local Volatility and Dupire's Formula

We now know that at each expiry the market has told us a distribution. It has told us nothing about how the underlying travels between expiries, and infinitely many processes are consistent with any given set of distributions. The question Dupire asked is whether at least *one* of them can be found in a particularly simple class.

Definition 9.4 (Local volatility model)**.** A local volatility model is one in which the diffusion coefficient is a deterministic function of time and of the current level of the underlying:

$$dS_t = (r-q)S_t\,dt + \sigma_{\text{loc}}(t,S_t)\,S_t\,dW_t.$$

There is no new randomness here. Volatility is not a risk factor; it is a lookup table indexed by where the underlying happens to be and what time it is. That makes the model complete, in the sense of chapter 4 — the underlying and the money market still replicate everything.

Theorem 9.5 (Dupire)**.** *If call prices $C(T,K)$ for all expiries and strikes come from a local volatility model, then*

$$\sigma_{loc}^2(T,K) = \frac{2\Big(\dfrac{\partial C}{\partial T} + (r-q)K\dfrac{\partial C}{\partial K} + q\,C\Big)}{K^2\dfrac{\partial^2 C}{\partial K^2}}. \tag{9.4}$$

Proof. Let $p(T,S)$ be the risk neutral density of S_T, which moves according to (9.2). Write the call price as $C(T,K) = e^{-rT}\int_K^\infty (S-K)p(T,S)\,dS$ and differentiate in T:

$$\frac{\partial C}{\partial T} = -r\,C + e^{-rT}\int_K^\infty (S-K)\frac{\partial p}{\partial T}\,dS.$$

Substitute the forward equation and take the two resulting integrals in turn. For the drift term, one integration by parts gives

$$\begin{aligned}\int_K^\infty (S-K)\Big(-\frac{\partial}{\partial S}\big[(r-q)Sp\big]\Big)dS &= \Big[-(S-K)(r-q)Sp\Big]_K^\infty + \int_K^\infty (r-q)Sp\,dS \\ &= (r-q)\int_K^\infty Sp\,dS,\end{aligned}$$

the boundary term vanishing at $S=K$ because of the factor $(S-K)$, and at infinity because the density decays. Splitting $S = (S-K)+K$ and using Theorem 9.3,

$$\int_K^\infty Sp\,dS = \int_K^\infty (S-K)p\,dS + K\int_K^\infty p\,dS = e^{rT}C - Ke^{rT}\frac{\partial C}{\partial K}.$$

For the diffusion term, integrate by parts twice. Writing $\phi(S) = \sigma_{\text{loc}}^2(T,S)S^2p(T,S)$ for brevity,

$$\begin{aligned}\frac{1}{2}\int_K^\infty (S-K)\frac{\partial^2\phi}{\partial S^2}\,dS &= \frac{1}{2}\left(\Big[(S-K)\frac{\partial\phi}{\partial S}\Big]_K^\infty - \int_K^\infty \frac{\partial\phi}{\partial S}\,dS\right) \\ &= \frac{1}{2}\Big(0 - \big[\phi\big]_K^\infty\Big) = \frac{1}{2}\phi(K) \\ &= \frac{1}{2}\sigma_{\text{loc}}^2(T,K)K^2p(T,K) = \frac{1}{2}\sigma_{\text{loc}}^2(T,K)K^2e^{rT}\frac{\partial^2 C}{\partial K^2}.\end{aligned}$$

Collecting the pieces and cancelling e^{-rT} against e^{rT},

$$\begin{aligned}\frac{\partial C}{\partial T} &= -rC + (r-q)\Big(C - K\frac{\partial C}{\partial K}\Big) + \frac{1}{2}\sigma_{\text{loc}}^2(T,K)K^2\frac{\partial^2 C}{\partial K^2} \\ &= -qC - (r-q)K\frac{\partial C}{\partial K} + \frac{1}{2}\sigma_{\text{loc}}^2(T,K)K^2\frac{\partial^2 C}{\partial K^2},\end{aligned}$$

and rearranging for σ_{loc}^2 gives (9.4). □

What has just happened deserves restating slowly, because the formula is usually met as something to be memorised.

We started with a question about a function of two variables — what is $\sigma_{\text{loc}}(t, S)$? — and we have answered it by rearranging an equation. There was no search, no fitting, and no approximation. Given a surface of call prices smooth enough to differentiate, formula (9.4) *tells* you the local volatility, at every point, in closed form. And since the surface was arbitrary, this says something remarkable:

Remark (Any smile can be fitted exactly)**.** For any arbitrage-free surface of European option prices, there is a local volatility model reproducing every one of them exactly.

The simplest reading of the formula is dimensional. The denominator, $K^2\partial^2 C/\partial K^2$, is the density at K weighted by K^2, and the numerator is essentially the rate at which option value accrues as expiry is pushed out. So local volatility is the ratio of how much value time buys you at that strike to how much probability is sitting there. Where the market is confident the underlying will not go, the density is small, the denominator is small, and the local volatility that has to be assigned there is correspondingly large.

Before trusting it, we should check it on a case where we know the answer.

Example 9.2 (Dupire on a flat smile)**.** Suppose the surface came from Black-Scholes with a constant σ, and take $r = q = 0$ so that (9.4) reduces to $\sigma^2_{\text{loc}} = 2C_T/(K^2 C_{KK})$. From chapter 5,

$$\frac{\partial C}{\partial T} = \frac{S_0\, n(d_1)\sigma}{2\sqrt{T}}, \qquad \frac{\partial^2 C}{\partial K^2} = \frac{n(d_2)}{K\sigma\sqrt{T}},$$

so

$$\sigma^2_{\text{loc}} = \frac{2}{K^2} \cdot \frac{S_0 n(d_1)\sigma}{2\sqrt{T}} \cdot \frac{K\sigma\sqrt{T}}{n(d_2)} = \sigma^2\, \frac{S_0\, n(d_1)}{K\, n(d_2)}.$$

It remains to show the last fraction is 1. Since $d_1 - d_2 = \sigma\sqrt{T}$ and $d_1 + d_2 = 2\ln(S_0/K)/(\sigma\sqrt{T})$,

$$\frac{n(d_1)}{n(d_2)} = e^{-\frac{1}{2}(d_1^2 - d_2^2)} = e^{-\frac{1}{2}(d_1 - d_2)(d_1 + d_2)} = e^{-\ln(S_0/K)} = \frac{K}{S_0},$$

so $S_0 n(d_1) = K n(d_2)$ and $\sigma_{\text{loc}} = \sigma$. A flat smile implies a constant local volatility, as it must.

Remark (Why the formula is harder to use than to derive)**.** Formula (9.4) asks for one derivative in T and two in K of a function known only at the strikes and expiries someone chose to quote. Differentiating twice amplifies noise ferociously, and the second derivative sits in the denominator, where a small error becomes a large one and a negative value becomes a catastrophe. In practice one never differences raw quotes: the surface is first fitted with something smooth and arbitrage-free by construction, and Dupire is applied to that. The formula is exact; the input never is.

Two things are commonly used for that fit, and they are not the same kind of object. SABR is a model, pressed into service as an interpolator; its implied volatility is an asymptotic expansion, and chapter 10 locates precisely where that expansion stops being arbitrage-free — at low strikes it can imply a negative density, which is the very quantity this formula divides by. The stochastic volatility inspired parameterisation, SVI, is not a model at all: it writes the total implied variance

directly as a function of log-moneyness with five parameters to a slice, has no dynamics, and so cannot price a barrier or say how the smile moves. Chapter 20 sets it out, and says how far the shape is dictated by what a smile has to do and how far it is simply the simplest choice that does it. What it can do is be constrained to admit neither butterfly nor calendar arbitrage.

Which of the two is the better tool here follows from the remark above that any smile can be fitted exactly. Dupire does not need a model, because the model is already whatever the surface says it is. It needs a surface convex in the strike and increasing in total variance with maturity, and nothing besides. A parameterisation guaranteeing exactly those is doing the job that was asked; a stochastic volatility model in this role is doing more than was asked everywhere, and in the wings less than is needed.

9.6 Dupire in the Coordinates We Actually Have

Formula (9.4) is written in call prices, and nobody quotes call prices. What is quoted is implied volatility, so it pays to have Dupire in those coordinates too — and the translation turns out to reveal something the price version hides completely.

Use the two natural coordinates. Log-moneyness measures how far a strike is from the forward,

$$y = \ln \frac{K}{F},$$

and total implied variance measures how much uncertainty an option contains,

$$w(y,T) = \sigma_{\text{imp}}^2(y,T)\,T.$$

Total variance rather than volatility, because variance is what accumulates: two independent periods contribute variances that add, and it is w, not σ_{imp}, that is the natural quantity in time.

Substituting $C = \text{Black}(F, K, \sqrt{w/T}, T)$ into (9.4) and grinding through the chain rule — the algebra is mechanical and long, and adds nothing once you have seen where it goes — gives

$$\sigma_{\text{loc}}^2(y,T) = \frac{\dfrac{\partial w}{\partial T}}{1 - \dfrac{y}{w}\dfrac{\partial w}{\partial y} + \dfrac{1}{4}\left(-\dfrac{1}{4} - \dfrac{1}{w} + \dfrac{y^2}{w^2}\right)\left(\dfrac{\partial w}{\partial y}\right)^2 + \dfrac{1}{2}\dfrac{\partial^2 w}{\partial y^2}}. \tag{9.5}$$

Do not memorise it. Read what each piece is for. The numerator says local variance is fed by implied variance accumulating with maturity: if a surface's total variance were flat in T, no local volatility could be extracted at all. The denominator is a set of corrections for the surface having shape in the strike direction — the first for its slope, the second for the slope's interaction with the level, the third for its curvature. And since it is a denominator, it must stay positive, which is the arbitrage condition of the previous section reappearing in volatility coordinates: a surface whose skew is too steep for its level makes the denominator vanish and the implied local variance blow up, and that is the calendar or butterfly arbitrage announcing itself.

Now the case to do by hand, because it is the one chapter 10 is built on. Suppose that near the money the smile is a straight line,

$$\sigma_{\text{imp}}(y) = \sigma_0 + s\,y,$$

with a small slope s. Then to first order in s,

$$w = (\sigma_0 + sy)^2 T = \sigma_0^2 T + 2\sigma_0 syT + O(s^2),$$
$$\frac{\partial w}{\partial y} = 2\sigma_0 sT + O(s^2), \qquad \frac{\partial^2 w}{\partial y^2} = O(s^2), \qquad \frac{\partial w}{\partial T} = \sigma_0^2 + 2\sigma_0 sy + O(s^2).$$

Every term in the denominator of (9.5) carrying $(\partial w/\partial y)^2$ or $\partial^2 w/\partial y^2$ is second order and drops, leaving

$$1 - \frac{y}{w}\frac{\partial w}{\partial y} = 1 - \frac{y}{\sigma_0^2 T} \cdot 2\sigma_0 sT = 1 - \frac{2sy}{\sigma_0}.$$

So

$$\sigma_{\text{loc}}^2 = \frac{\sigma_0^2 + 2\sigma_0 sy}{1 - 2sy/\sigma_0} \approx \left(\sigma_0^2 + 2\sigma_0 sy\right)\left(1 + \frac{2sy}{\sigma_0}\right) \approx \sigma_0^2 + 4\sigma_0 sy,$$

and taking the square root,

$$\sigma_{\text{loc}}(y) \approx \sigma_0 + 2sy. \tag{9.6}$$

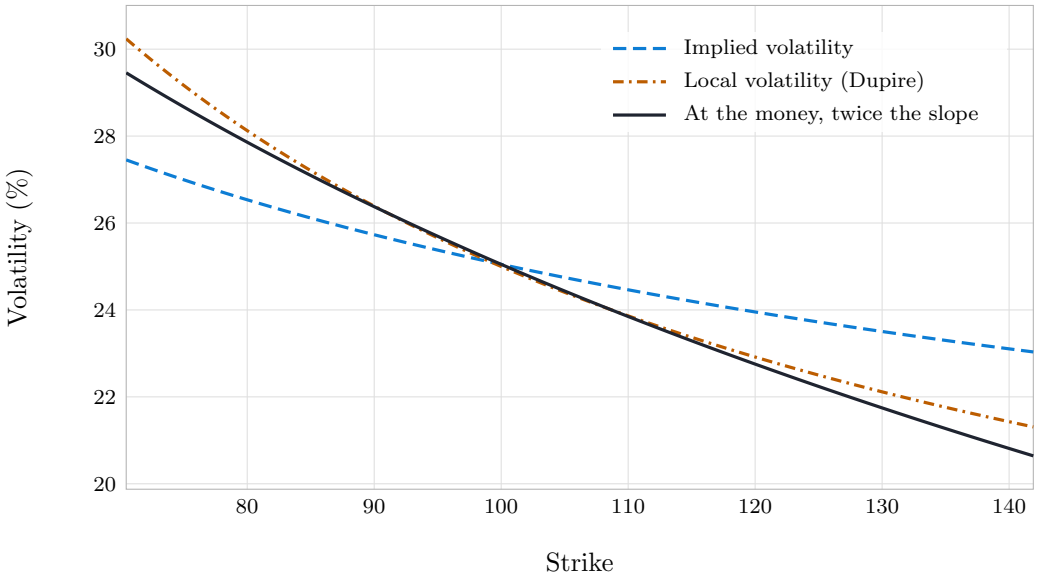

Figure 9.3: A displaced diffusion's implied volatility, the local volatility Dupire's formula extracts from it, and the straight line predicted by (9.6). The local volatility curve is twice as steep, and the prediction is a line drawn through the at-the-money point at twice the observed slope — not a fit. It tracks the extracted curve closely near the money and drifts away in the wings, which is what a first-order result should do.

Drawn by `dupire_local_vol`.

Remark (The rule of two)**.** The local volatility curve has *twice* the slope of the implied volatility curve that produced it. If the market's smile falls by half a volatility point per ten percent of moneyness, the local volatility falls by a full point.

An option struck at K does not care about volatility at K alone; the underlying travels from the forward to wherever it ends, and the option's implied volatility is a kind of average of the local

volatility over that journey. Averaging spot and strike halves a slope. So to produce an observed implied slope s by averaging, the underlying local slope must have been $2s$ — the same factor of two, read backwards.

This apparently innocuous factor is the whole of chapter 10. It says that a local volatility model, having been fitted to today's skew, contains a volatility function twice as steep as that skew. That steepness is not a free parameter: it was forced on us by the fit. And it determines how the model thinks the smile moves when the underlying moves, which turns out to be badly.

9.7 Gyongi's Theorem

The remark above — that *any* smile can be fitted — should be unsettling. We fitted every European option price exactly, using a model with no volatility risk in it at all — a model in which volatility is a deterministic table. Either European options contain no information about the randomness of volatility, or something has been quietly assumed. It is the former, and Gyongi's theorem is the statement of exactly how much information they do contain.

Theorem 9.6 (Gyongi)**.** *Let X be an Itô process*

$$dX_t = \beta_t\,dt + \delta_t\,dW_t$$

with β and δ bounded adapted processes, not necessarily Markov and not necessarily functions of X. Define

$$a(t,x) = \mathbb{E}\big[\beta_t \,\big|\, X_t = x\big], \qquad b^2(t,x) = \mathbb{E}\big[\delta_t^2 \,\big|\, X_t = x\big].$$

Then the Markov diffusion

$$dY_t = a(t,Y_t)\,dt + b(t,Y_t)\,dW_t, \qquad Y_0 = X_0,$$

has the same distribution as X_t at every single time t.

Proof. Two processes have the same distribution at each time if the expectations of every smooth test function agree at each time, so fix such an f and track $\mathbb{E}[f(X_t)]$. By Itô's lemma,

$$df(X_t) = f'(X_t)\beta_t\,dt + \tfrac{1}{2}f''(X_t)\delta_t^2\,dt + f'(X_t)\delta_t\,dW_t,$$

and the stochastic integral has zero expectation under the boundedness assumed, so

$$\frac{d}{dt}\mathbb{E}\big[f(X_t)\big] = \mathbb{E}\Big[f'(X_t)\beta_t + \tfrac{1}{2}f''(X_t)\delta_t^2\Big].$$

Now the one step that matters. Condition on X_t and use the tower property. Because $f'(X_t)$ and $f''(X_t)$ are functions of X_t alone, they pass through the inner conditional expectation as constants:

$$\mathbb{E}\Big[f'(X_t)\beta_t\Big] = \mathbb{E}\Big[\mathbb{E}\big[f'(X_t)\beta_t \,\big|\, X_t\big]\Big] = \mathbb{E}\Big[f'(X_t)\,\mathbb{E}\big[\beta_t \,\big|\, X_t\big]\Big] = \mathbb{E}\Big[f'(X_t)\,a(t,X_t)\Big],$$

and identically $\mathbb{E}[f''(X_t)\delta_t^2] = \mathbb{E}[f''(X_t)b^2(t,X_t)]$. Therefore

$$\frac{d}{dt}\mathbb{E}\big[f(X_t)\big] = \mathbb{E}\Big[f'(X_t)a(t,X_t) + \tfrac{1}{2}f''(X_t)b^2(t,X_t)\Big].$$

Running the same calculation for Y, whose coefficients are a and b by construction, produces the identical equation. So $\mathbb{E}[f(X_t)]$ and $\mathbb{E}[f(Y_t)]$ solve the same evolution equation from the same starting value, and where that equation has a unique solution they agree for all t and all f. □

Remark (What the theorem says, without symbols)**.** Take any process at all — let its volatility be random, let it depend on the whole history, let it be driven by other factors we cannot see. Now ask what it looks like at a single fixed date, ignoring how it got there. Gyongi says that this snapshot is always reproduced by some Markov diffusion, and tells you which one: at each point, use the *average* of the true volatility over all the ways the world could have arrived at that point.

The averaging is where the information goes. Two worlds, one in which volatility at $S = 100$ is always 20% and another in which it is 10% half the time and about 27% the other half, have the same average squared volatility there, and therefore the same distribution of S_T, and therefore the same European option prices. Nothing you can construct from options at a single expiry can tell them apart — because the option price is an expectation over that expiry's distribution, and the distributions are the same.

They are, of course, completely different worlds to be in if you have to hedge.

Applying the theorem to a general stochastic volatility model, in which $dS_t/S_t = \cdots + \sigma_t\, dW_t$ with σ_t random, and comparing with Dupire gives the identity that the next two chapters revolve around.

Theorem 9.7 (Local volatility is a conditional expectation)**.** *If a model with random instantaneous volatility σ_t reproduces the market's European option prices, then the Dupire local volatility of that same surface satisfies*

$$\sigma_{loc}^2(T,K) = \mathbb{E}\big[\sigma_T^2 \,\big|\, S_T = K\big]. \tag{9.7}$$

Proof. Both sides of (9.7) are read off the same object — the price surface — so the proof consists of computing that surface twice and comparing.

Write $C(T,K) = \mathbb{E}[(S_T - K)^+]$ for the undiscounted call surface of the stochastic volatility model, and differentiate it in maturity. The payoff has chapter 3's kink at K, so it is not classically twice differentiable there; its second derivative exists only distributionally, as a Dirac delta at the kink, and Itô's lemma applied to the payoff with that derivative gives

$$d(S_T - K)^+ = \mathbf{1}_{\{S_T > K\}}\, dS_T + \tfrac{1}{2}\delta(S_T - K)\,\sigma_T^2 S_T^2\, dT,$$

and the first term has zero expectation because S is a martingale. So

$$\frac{\partial C}{\partial T} = \tfrac{1}{2}\mathbb{E}\big[\delta(S_T - K)\,\sigma_T^2 S_T^2\big] = \tfrac{1}{2}K^2\,\mathbb{E}\big[\sigma_T^2 \,\big|\, S_T = K\big]\,p(T,K), \tag{9.8}$$

where the last step is the definition of a conditional expectation: integrating against the delta picks out the density $p(T,K)$ of S_T at K and, with it, the average of σ_T^2 over exactly the paths that arrive there. This is the step that produces the conditioning, and it produces it from nothing more than the fact that the second derivative of a call payoff is supported on a single level.

Now the same surface in the local volatility model. There the volatility is a function, so the identical computation gives

$$\frac{\partial C}{\partial T} = \tfrac{1}{2}K^2\,\sigma_{\text{loc}}^2(T,K)\,p(T,K),$$

with no expectation to take. And the density is the same density: $\partial^2 C/\partial K^2 = p(T,K)$ for any model, which is the Breeden-Litzenberger identity, so two models agreeing on the surface agree on the marginal law of S_T.

Equating the two expressions for $\partial C/\partial T$ and cancelling $\frac{1}{2}K^2 p(T,K)$, which is non-zero wherever the density is, gives (9.7). □

Remark (What the proof shows about the scope)**.** Three features of the argument stand out, because they say exactly how far the theorem reaches.

It used the martingale property of S and nothing else about the drift, so it is indifferent to how many factors the volatility has, whether they are correlated with S, and whether σ is Markovian. Any of that would appear only inside the conditional expectation.

It matched one derivative of one surface. That is why the conclusion is about one-dimensional marginals and about European prices, and why nothing follows about the joint law across maturities. Chapter 10 is where that gap does damage: two models with the same σ_{loc} can disagree completely about a forward-starting payoff, and (9.7) does not notice.

It divided by $p(T,K)$. Where the density is small — far wings, short maturities — the local volatility is a ratio of two small numbers, which is the analytic reason for the numerical fragility of (9.4) rather than a defect of any particular scheme for evaluating it.

There is one case where the conditional expectation is available in closed form: a forward whose volatility is drawn once, at time zero, from two values. That is a genuine stochastic volatility — random and unknown — and its mimicking local volatility can be written down and simulated. The two models agree on every European price to within a few hundredths of a volatility point, which is a strong statement given that they are nothing like each other path by path: one has a volatility that never moves and is a function of the level, the other a volatility that is constant along each path and unknown at the start.[1]

Constructively, it is the tool of chapter 11: it says that if you have a stochastic volatility model whose dynamics you like but whose fit you do not, you can multiply its volatility by whatever function makes (9.7) hold, and recover an exact fit without disturbing the dynamics. That is what a local-stochastic volatility model is.

Destructively, it says that local volatility is an average of something, and an average is a summary. The local volatility surface is not the market's volatility; it is the market's volatility with all of its own randomness integrated away. Every model matching the smile shares this one average, and they differ in everything the average discarded — which is to say, in everything that depends on more than one date at a time.

That is the subject of the next chapter. Local volatility fits every European option in the market exactly, and it is nevertheless the wrong model for almost everything else, in a way we can make quantitative rather than rhetorical.

[1] `checked`, with `a guard against the check being vacuous` — if the mixture's smile were flat a constant volatility would reproduce it and nothing would have been shown.

9.8 Scheduled Events on the Surface

The curve chapter built the front end as a step function because the overnight rate moves only at meetings. The surface has the same structure and the same remedy, in the coordinate that matters here: variance rather than rate.

Between events, variance accumulates diffusively — linearly in time, which is what makes total variance the natural interpolation coordinate rather than volatility. At an event it jumps, because a policy decision or a labour report delivers a move in an instant. So the surface is

$$\sigma^2_{\text{imp}}(T)\,T = \int_0^T \sigma^2_{\text{diff}}(u)\,du + \sum_{\text{events } i \le T} v_i, \tag{9.9}$$

with v_i the variance contributed by the ith event.

The construction that follows is the same manoeuvre as the meeting-date curve. Strip the event lumps out of the quoted total variances, interpolate what is left — which is now smooth, because the jumps have been removed — and add the lumps back at the right dates. Interpolating the raw surface instead produces the familiar pathology: two options expiring three days apart across a payroll date have implied volatilities that differ by several points, a smooth interpolation reads that as a volatility term structure, and every expiry in between is marked wrong.

Remark (Calibrating the lumps)**.** The v_i are not free parameters to be fitted alongside everything else. They are pinned by the pairs of expiries that straddle each event: the difference in total variance between an option expiring just before a release and one expiring just after *is* v_i, up to the diffusive variance of the days between. Where such a pair is quoted, the event variance is a read-off rather than a fit.

Where it is not quoted — and for most events it is not — the lump is marked from history: the average squared move of the rate on past occurrences of that release. That is a genuinely different kind of number from the rest of the surface, being a historical estimate sitting inside a risk-neutral object, and it should be understood as a marking convention rather than as a calibration. It is also where a desk's view lives: a trader who thinks this meeting is more consequential than the last four marks the lump up, and that is a position rather than a parameter.

A jump or a lump of volatility?

Chapter 4's example Reading the hike probability off the futures established that a policy decision has a finite and small outcome space — twenty-five basis point multiples, three or four possibilities — so it can be represented literally, as a discrete distribution over moves. The alternative is to represent it as extra Gaussian variance with the same second moment, which is what (9.9) does. The two are not the same model and the question is when the difference is paid for.

Calculation 9.8 (Where the two representations part)**.** A meeting with risk-neutral outcomes $\{0, +25, +50\}$ basis points at probabilities $\{0.35, 0.55, 0.10\}$ — read off market prices the way chapter 4's example did, not assumed — a forward of 4%, and a diffusive normal volatility of sixty basis points a year. What follows compares undiscounted prices, $\mathbb{E}^T[(F_T - K)^+]$ in the convention this chapter has used throughout, so every time below is elapsed time since the meeting resolves rather

than time from today: discounting from today to the meeting is a factor common to both representations, and touches neither the ratio between their prices nor the implied volatility either reads back out. Compare pricing calls on the rate — and, from them, a digital — under the three-point mixture against under a single Gaussian matched to the same total variance.[2]

Write $\mu = \sum_i p_i m_i$ for the mean move — already embedded in the forward — and $\mathrm{Var}[m] = \sum_i p_i (m_i - \mu)^2$ for its variance around that mean. The mixture prices as a probability-weighted sum of Bachelier prices at shifted forwards,

$$C_{\mathrm{mix}}(K) = \sum_i p_i \,\mathrm{Bachelier}\big(F + m_i - \mu,\, K,\, \sigma,\, T\big),$$

and the matched Gaussian carries the event's variance forward instead of its shape,

$$\sigma_{\mathrm{matched}} = \sqrt{\sigma^2 + \mathrm{Var}[m]/T}.$$

Both reproduce the same total variance at T by construction — that is what "matched" means — so only the mixture's shape survives inverting C_{mix} into an implied volatility.

At the money they agree, to within a fraction of a per cent, because a straddle is priced off the variance and the variance was matched by construction. This is why the choice can go unnoticed indefinitely on an at-the-money book.

Across strikes they do not. The mixture puts mass at three places and less between them, so its implied volatility is not flat: at a one month expiry it runs from about 70 basis points at a strike a hundred below the forward to 82 at the money, an eleven basis point smile manufactured entirely by the event. A single Gaussian has no smile at all.

And the size depends on how much diffusion sits between the event and the expiry, which is the finding that decides everything. A digital struck between two outcomes prices at 0.194 under the mixture against 0.230 under the Gaussian three days after the meeting — a sixth of the price. At one month the two agree to four decimal places, and at a year they are indistinguishable. Seventeen basis points of diffusion against a twenty-five basis point spacing is enough to smear three atoms into something very nearly Gaussian.

Structure (The instrument decides, and so does the calendar)**.** Calculation 9.8 gives a rule with two conditions rather than a preference.

What is the payoff reading? An instrument priced off the second moment — a straddle, a variance swap, anything approximately quadratic — cannot tell the two representations apart, because they were constructed to agree there. An instrument priced off the shape of the density — a digital, a tight butterfly, a range accrual, a barrier sitting between two policy outcomes — reads exactly what differs. So the choice is decided by the book, and a desk trading at-the-money volatility genuinely does not need the discrete representation.

And how much diffusion follows the event? The discrete structure survives only while the diffusion is small compared with the spacing between outcomes. For an option expiring days after a meeting that is the whole picture; a month later it is nearly gone; a year later the event is a lump

[2] `events::PolicyEvent`, priced in closed form as a probability weighted sum of Bachelier prices rather than simulated.

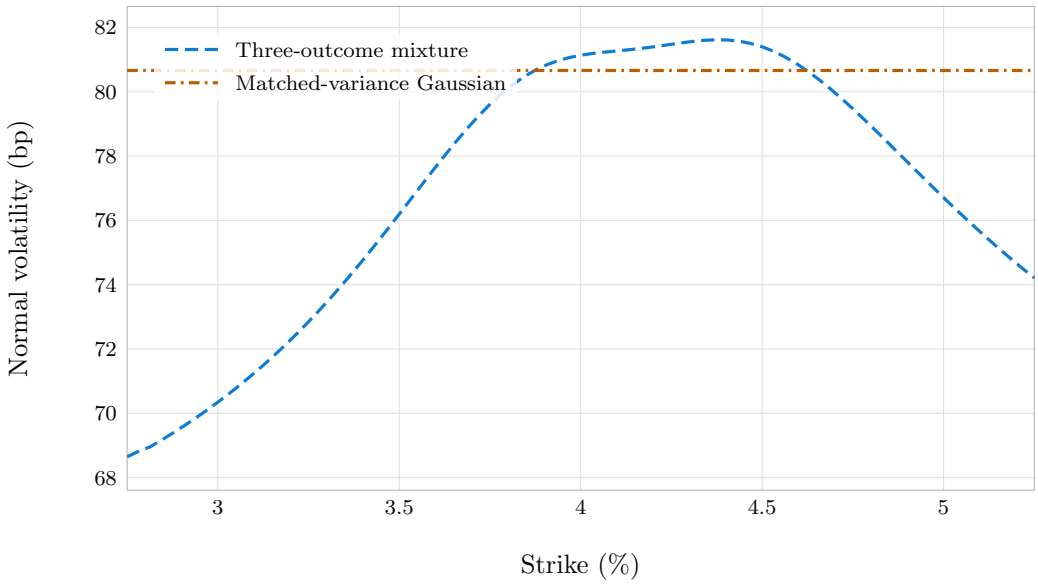

Figure 9.4: The implied normal volatility of the mixture against the matched Gaussian, across strikes, one month before the meeting. The Gaussian is flat by construction — it was built to match the mixture's variance and nothing else — while the mixture's three atoms and the gaps between them show up as a genuine skew: lowest below the forward, rising through it, matching the numbers in the text. The shape is not fitted to anything; it is what three point masses look like once a month of diffusion has blurred their edges.

Drawn by `representation_gap`.

of variance and nothing else. This is why the discrete representation is a front-end tool: it earns its complexity on the short-dated options that straddle a meeting, and is wasted effort further out.

Both conditions have to hold for the extra machinery to be worth it, which is a narrower criterion than the usual advice to model events properly. The general lesson is the one chapter 18 makes about copulas and chapter 10 about the forward smile: matching a moment is not matching a distribution, and whether the difference matters is a question about the payoff rather than about the model.

References

- Breeden, D. T., & Litzenberger, R. H. (1978). Prices of state-contingent claims implicit in option prices. *Journal of Business*, 51(4), 621–651.

- Dupire, B. (1994). Pricing with a smile. *Risk*, 7(1), 18–20.

- Gyongi, I. (1986). Mimicking the one-dimensional marginal distributions of processes having an Itô differential. *Probability Theory and Related Fields*, 71(4), 501–516.

- Derman, E., & Kani, I. (1994). Riding on a smile. *Risk*, 7(2), 32–39.

Chapter 10

Smile Dynamics and Stochastic Volatility

We show that fitting every option in the market is not enough. We derive, rather than assert, the relationship between the slope of a smile and the way that smile moves, and find that a local volatility model gets it wrong by a factor of two in the wrong direction. We then introduce models in which volatility is genuinely random, and show that the market's quotes do not determine which one to use.

10.1 What Is Left to Get Wrong

Chapter 9 ended in an uncomfortable place. Dupire's formula fits every European option in the market exactly, with no fitting and no error, using a model in which volatility carries no randomness at all. If a model reproduces every price we can see, in what sense can it be wrong?

Gyongi's theorem already told us. A local volatility model matches the *marginal* distribution of the underlying at each date. It says nothing about the joint distribution across dates, and infinitely many joint distributions share a set of marginals. So the model is guaranteed to be right about anything determined by one date at a time, and is unconstrained about everything else.

- Anything path dependent. A barrier option cares whether the underlying visited a level, which is a statement about the whole path.
- Anything with an early exercise decision. A Bermudan swaption's value depends on what the smile will look like on each exercise date.
- Anything forward starting. A cliquet struck at the money in a year's time is a bet on volatility a year from now.
- *The hedge.* This one is easy to miss and is the most important, because it applies to the plain vanilla options themselves.

The last point deserves spelling out. Suppose we and the market agree exactly on the price of every option today. We then hold one and hedge it. The hedge ratio is

$$\Delta = \frac{dV}{dS},$$

the total derivative of the option's value with respect to the underlying — and V depends on S both directly and through the volatility we will use to price it after S has moved. Writing $V = \text{Black}(S, K, \sigma_{\text{imp}}(K, S), T)$ and differentiating,

$$\Delta = \underbrace{\frac{\partial \text{Black}}{\partial S}}_{\text{the textbook delta}} + \underbrace{\frac{\partial \text{Black}}{\partial \sigma} \cdot \frac{\partial \sigma_{\text{imp}}}{\partial S}}_{\text{the smile's contribution}} . \tag{10.1}$$

The first term is what chapter 5 gave us. The second is vega multiplied by the rate at which the implied volatility of *this* option changes when the underlying moves. Vega is large — for a one year at-the-money option it is roughly 0.4 of the underlying per unit of volatility — so the second term is not a refinement.

And $\partial \sigma_{\text{imp}}/\partial S$ is not observable today. It is a statement about the future, and therefore a statement the model has to make. Two models agreeing on every price today will disagree on the hedge unless they also agree about this. That is the subject of the chapter.

10.2 Two Rules of Thumb, and What They Would Mean

Before deriving what a model says, it helps to have the two things a trader might say, because they bracket the answer.

Definition 10.1 (Sticky strike)**.** The smile is sticky strike if the implied volatility of a *fixed strike* does not change when the underlying moves: σ_{imp} is a function of K alone.

Definition 10.2 (Sticky delta, or sticky moneyness)**.** The smile is sticky delta if the implied volatility depends only on the strike relative to the forward: $\sigma_{\text{imp}} = \varphi(K/F)$. The smile rides along with the market.

To compare them we need two quantities that are easy to confuse, so let us name them carefully. Both are measured at the money and in log-moneyness $y = \ln(K/F)$, which is the coordinate that makes them comparable.

Definition 10.3 (Skew and backbone)**.** The *skew* is the slope of today's smile in the strike direction, with the forward held still:

$$s = \left.\frac{\partial \sigma_{\text{imp}}}{\partial y}\right|_{y=0}.$$

The *backbone* is the rate at which the at-the-money volatility itself moves when the forward moves:

$$b = \frac{\partial \sigma_{\text{ATM}}}{\partial \ln F}, \qquad \sigma_{\text{ATM}}(F) = \sigma_{\text{imp}}(K = F, F).$$

The skew is a photograph. The backbone is a motion. They are different numbers, they are measured on different days. The market lets us observe the first and not the second.

Before the rules, one identity, because these two numbers are constantly confused and the identity is what keeps them apart. There is a third derivative in play — how the volatility of *one fixed strike* moves — and all three are related by the chain rule.

Lemma 10.4 (The three derivatives)**.** *For any smile,*

$$b = s + \left.\frac{\partial \sigma_{imp}}{\partial \ln F}\right|_{K \text{ fixed}}, \tag{10.2}$$

both derivatives on the right evaluated at the money.

Proof. $\sigma_{\text{ATM}}(F) = \sigma_{\text{imp}}(F, F)$, so moving the forward moves both arguments. Differentiating,

$$b = \frac{d}{d \ln F}\, \sigma_{\text{imp}}(F, F) = \left.\frac{\partial \sigma_{\text{imp}}}{\partial \ln K}\right|_{K=F} + \left.\frac{\partial \sigma_{\text{imp}}}{\partial \ln F}\right|_{K=F},$$

and the first term is the skew by definition. □

Now the two rules.

Under **sticky delta**, $\sigma_{\text{ATM}} = \varphi(1)$ is a constant, so $b = 0$: the at-the-money volatility never moves, however far the market travels. From (10.1), with K fixed and $\sigma_{\text{imp}} = \varphi(K/F)$, the smile's contribution to delta is $\partial \sigma_{\text{imp}} / \partial \ln F = -s$ at the money.

Under **sticky strike**, $\sigma_{\text{imp}}(K)$ does not move, so as the forward slides along the fixed smile the at-the-money volatility takes the value of the smile at the new level: $\sigma_{\text{ATM}}(F) = \sigma_{\text{imp}}(F)$, and $b = s$. The smile's contribution to delta is zero, since the volatility of the option we hold does not change at all.

So the two conventions bracket the answer: $b = 0$ at one end and $b = s$ at the other. In an equity market with a downward skew, $s < 0$, so the two say the at-the-money volatility either stays put or falls a little when the market rises.

We are now in a position to ask what a local volatility model says — and the answer is outside the bracket.

10.3 The Backbone of a Local Volatility Model

The one ingredient we need from outside is a short-maturity result about how a local volatility function turns into an implied volatility. It has a clean statement and a clean intuition.

Theorem 10.5 (Berestycki, Busca and Florent)**.** *In a local volatility model, as the expiry shrinks to zero the implied volatility converges to the harmonic mean of the local volatility along the path in log-space from the forward to the strike:*

$$\sigma_{imp}(K, F) \;\longrightarrow\; \left(\frac{1}{\ln K - \ln F} \int_{\ln F}^{\ln K} \frac{dz}{\sigma_{loc}(e^z)} \right)^{-1}. \tag{10.3}$$

Why it is true, before why it is a theorem. Think of the underlying as diffusing through a medium whose diffusivity varies from place to place. The option struck at K is asking how hard it is to get from F to K. In a medium of constant diffusivity σ, covering a log-distance d takes a characteristic time $(d/\sigma)^2$, so the natural measure of separation is d/σ, and in a varying medium the separation of F and K is the accumulated $\int dz/\sigma_{\text{loc}}(e^z)$. Implied volatility is by definition the constant volatility reproducing the same option price, hence the same separation, so

$$\frac{\ln K - \ln F}{\sigma_{\text{imp}}} = \int_{\ln F}^{\ln K} \frac{dz}{\sigma_{\text{loc}}(e^z)},$$

which is (10.3) rearranged. Implied volatility is a harmonic average because it is an average of a *rate*, and rates average harmonically — for the same reason that driving one mile at 30 and one mile at 60 does not average to 45. □

Remark (What turning that into a proof requires)**.** The argument above is not a proof and the gap is not a technicality, so it pays to see where the work is. The rigorous chain has four links.

One. The short-time behaviour of the transition density is governed by a distance. Varadhan's theorem says

$$-2t \log p(t, x, y) \;\longrightarrow\; d(x, y)^2,$$

where d is the Riemannian distance in the metric $ds = dz/\sigma_{\text{loc}}(e^z)$ — precisely the separation the heuristic wrote down. So the heuristic names the right object; what it does not do is establish the limit. The geometry that this metric carries is taken up later in the chapter.

Two. The same asymptotics transfer from the density to the option price, since an out-of-the-money call is an integral of the density over a region whose nearest point to the forward is the strike, and a Laplace-type estimate says an integral of $e^{-d^2/2t}$ is governed by the smallest d in the region.

Three. Black-Scholes with a constant volatility is the special case $\sigma_{\text{loc}} \equiv \sigma_{\text{imp}}$, where the distance is $|\ln K - \ln F|/\sigma_{\text{imp}}$. Equating the two distances is the theorem.

Four. Step one is classically available for a uniformly elliptic operator with smooth coefficients on a compact space, and none of those three conditions holds here: the operator degenerates as $S \to 0$, the domain is not compact, and σ_{loc} coming out of a calibration is at best continuous. The route that survives this substitutes $u_t = -2t \log C$ directly into the pricing equation, which turns a linear parabolic equation into a nonlinear one, and the limit as $t \downarrow 0$ is a first-order Hamilton-Jacobi equation — an eikonal equation $|\nabla u| = 1$, whose solutions are exactly distance functions. Two facts about that equation then have to be supplied: that the limit satisfies it, and that it has only one solution with the given boundary behaviour. Neither is available in the classical sense, because a distance function is not differentiable where two geodesics meet. Viscosity solutions are the notion of solution designed for exactly that, and the comparison principle for them is what delivers uniqueness.

A degenerate limit of a linear problem became a nonlinear problem, and a nonlinear first-order equation with non-smooth solutions needs a weak notion of solution to be well posed at all. The same substitution and the same difficulty appear wherever a small-noise limit is taken; large deviations theory is this argument in general form, with the exponential rate playing the part of the distance here.

From this we get the fact we actually use.

Lemma 10.6 (The midpoint rule)**.** *To first order in* $\ln K - \ln F$*,*

$$\sigma_{imp}(K,F) = \sigma_{loc}\big(\sqrt{FK}\big) + O\big((\ln K - \ln F)^2\big),$$

that is, the implied volatility of a strike is the local volatility evaluated halfway between the forward and the strike, measured in log-space.

Proof. Write $g(z) = 1/\sigma_{\mathrm{loc}}(e^z)$, and let

$$m = \frac{\ln F + \ln K}{2}, \qquad \delta = \frac{\ln K - \ln F}{2},$$

so the integral in (10.3) runs from $m-\delta$ to $m+\delta$ and the mean is taken over a width 2δ. Expanding g about the midpoint,

$$\frac{1}{2\delta}\int_{m-\delta}^{m+\delta} g(z)\,dz = \frac{1}{2\delta}\int_{-\delta}^{\delta}\Big(g(m) + u\,g'(m) + \tfrac{1}{2}u^2 g''(m) + \dots\Big)du = g(m) + \frac{\delta^2}{6}g''(m) + \dots,$$

the term linear in u integrating to zero by symmetry. So the harmonic mean is $1/g(m)$ up to a correction of order δ^2, and $1/g(m) = \sigma_{\mathrm{loc}}(e^m) = \sigma_{\mathrm{loc}}(\sqrt{FK})$. □

The midpoint rule is the whole story in one line: *implied volatility is local volatility, averaged over the journey.* Averaging halves a slope, and that is where the factor of two lives.

Remark (The same two as chapter 9)**.** This factor is not a new one. Chapter 9's remark The rule of two established, statically, that the local volatility curve has twice the slope of the implied volatility curve — and for the same reason: implied volatility is an average of local volatility over the interval between the forward and the strike, and averaging a function over an interval halves its slope.

The market's verdict is shared too. Chapter 9 measured the ratio and found it near two only close to the money, decaying further out, because both statements are leading order. And the backbone measured in the market is not $2s$ either — which is the complaint this chapter is built around, arriving in a second form.

Theorem 10.7 (The local volatility backbone)**.** *In a local volatility model, the at-the-money volatility responds to a move in the forward at twice the rate of the skew:*

$$b = 2s. \tag{10.4}$$

Proof. The essential point is that σ_{loc} is a fixed function of the *level* of the underlying. It was calibrated once, to today's surface, and it does not move when the market does. Write it in log-space as $\ell(x) = \sigma_{\mathrm{loc}}(e^x)$. By Lemma 10.6,

$$\sigma_{\mathrm{imp}}(K,F) = \ell\left(\frac{\ln F + \ln K}{2}\right).$$

Both quantities we want are derivatives of this single expression, taken in different directions.

For the skew, hold F fixed and differentiate in $\ln K$, then set $K = F$:

$$s = \left.\frac{\partial \sigma_{\text{imp}}}{\partial \ln K}\right|_{K=F} = \frac{1}{2}\,\ell'(\ln F).$$

For the backbone, set $K = F$ *first* — the at-the-money option is a different contract at each level of the forward — giving $\sigma_{\text{ATM}}(F) = \ell(\ln F)$, and then differentiate:

$$b = \frac{\partial \sigma_{\text{ATM}}}{\partial \ln F} = \ell'(\ln F).$$

Dividing, $b = 2s$. □

The factor of two comes from a genuinely simple place. The skew moves only the strike, so it moves the midpoint by half as much; the backbone moves the forward and the strike together, so it moves the midpoint by the full amount. Chapter 9 finds the same factor: its rule of two, that the local volatility curve extracted from a smile is twice as steep as the smile itself, is this statement read backwards. The two derivations are independent, and `local_volatility_has_twice_the_slope_of_the_implied_s` in `quant/src/localvol.rs` checks the result numerically against a Dupire formula that shares none of the algebra.

Remark (Why this is bad, in numbers)**.** Put the three predictions side by side. Take a one year at-the-money option with $\sigma_{\text{ATM}} = 25\%$ and a skew of $s = -6$ volatility points per unit of log-moneyness, which is an ordinary equity index figure. Suppose the market rises 10%.

	backbone b	change in σ_{ATM}
sticky delta	0	0.00 points
sticky strike	s	-0.6 points
local volatility	$2s$	-1.2 points

Equity index markets behave much closer to the first row than the third. The local volatility model does not merely get the magnitude wrong; it predicts the at-the-money volatility falls twice as fast as even the more aggressive of the two rules of thumb, and it does so because it was fitted to today's skew. The steepness was not chosen. It was forced on the model by the requirement to match the smile.

Now feed that into (10.1). With vega around $0.4 \times S$ per unit volatility, an error of 1.2 volatility points per 10% move is a delta error of roughly $0.4 \times 0.012/0.10 = 0.048$, about five percent of the underlying's notional, on every option in the book, in the same direction. That is not a rounding error; it is a systematic position nobody chose to take.

Remark (No vega)**.** A second failure follows from the same source and can be stated quickly. Because there is no independent volatility factor, the model is complete in the sense of chapter 4: the underlying and the money market replicate everything, so there is no vega to hedge and nothing to hedge it with. A desk whose business is trading volatility needs volatility to be a risk factor, and here it is not one.

10.4 The Backbone the Market Actually Has

The backbone above is a modelling choice presented as a menu. A rate's volatility is taken proportional to some power of its level,

$$\sigma(r) = \alpha\, r^{\beta}, \tag{10.5}$$

and three values of β have names. At $\beta = 0$ the model is normal, so a yield at one per cent moves in basis points exactly as violently as a yield at ten. At $\beta = 1$ it is lognormal and moves ten times as violently. At $\beta = \frac{1}{2}$ it is the square-root middle that keeps the rate positive without letting the volatility scale fully. SABR carries the same exponent under the same letter, and it is usually not calibrated at all — it is set to one of the three by convention and the remaining parameters absorb the difference.

These notes can do better than convention here, because the underlying is observable. The curve is quoted every business day and has been for sixty-four years.

Remark (Why a physical measure estimate is admissible)**.** The objection to measuring β from history is the usual one: hedging happens under the pricing measure and a time series is drawn under the physical one, so the estimate appears to answer a different question.

It does not, and chapter 6 says why. Girsanov changes the drift of a process and leaves its diffusion coefficient exactly where it was, so (10.5) — which is a statement about the diffusion coefficient and nothing else — is the same equation under both measures. The premium lives in α's drift and in the level, not in the exponent. So β is one of the few quantities a pricing model needs that can be read off history without a change of measure.

Calculation 10.8 (The exponent, from sixty-four years of the ten-year yield)**.** Cut the history into non-overlapping months, take the realised volatility of daily changes within each and the average level across it, and regress one log on the other. Overlapping windows would quadruple the apparent sample without adding information and make the standard error a fiction.

Over the whole series the level ranges from 0.59% to 15.2%, a factor of twenty-six, and it is that range which identifies the exponent at all — a few months of data cannot see it, however many days they contain. The result is

$$\beta = 0.27 \pm 0.04,$$

on 769 monthly windows.[1] Every one of the three conventional choices is rejected: normal by 6.8 standard errors, square root by 5.8, and lognormal by 18.5.

Structure (What the number is saying, and what it is not)**.** The measured exponent is closer to normal than to anything else on the menu, and that is the empirical content of a convention the rates market adopted for a different reason. Normal quoting spread after 2008 because lognormal volatility is undefined at a negative rate and rates went negative; the argument was that the alternative had stopped working, not that it was right. The exponent says it was closer to right all along.

Two limits on the claim, both of which the figure shows better than the number does. The scatter is wide: a single month's realised volatility is a poor estimate, and the regression is extracting

[1] `measured`.

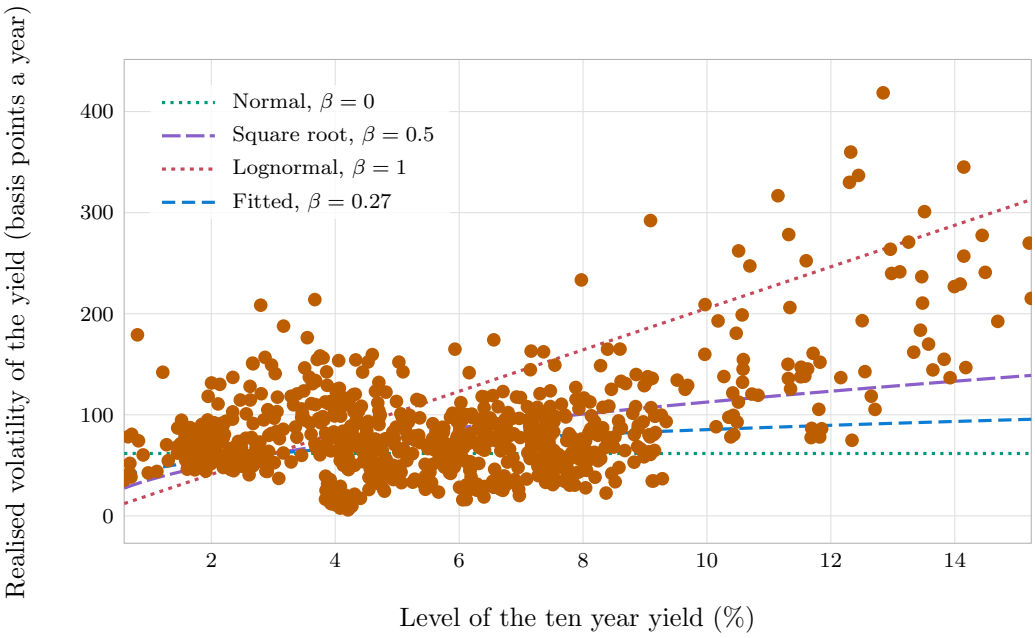

Figure 10.1: Realised volatility of the ten-year yield against its level, one point per non-overlapping month since 1962, with the three conventional exponents and the fitted one drawn through the sample's centre so that they are compared by slope rather than by height. The cloud is wide, which is why sixty-four years are needed and why the standard error matters more than the point estimate. What the picture shows is that the lognormal line is far too steep to pass through it — a yield at twelve per cent was not ten times as volatile as a yield at one, but roughly twice.
Drawn by `backbone`. Board of Governors of the Federal Reserve System (US), H.15 Selected Interest Rates, 10-year Treasury constant maturity (DGS10), retrieved from FRED, Federal Reserve Bank of St. Louis, 1962-01-02 to 2026-09-21 (constant maturity, bond equivalent yield, business days). Retrieved from https://fred.stlouisfed.org/graph/fredgraph.csv?id=DGS10.

a slope from a cloud rather than tracing a curve. And the two ends of the range come from different policy regimes. Every window above ten per cent falls between October 1979 and November 1985, which includes the three years the Federal Reserve targeted reserves rather than the funds rate, and every window below two per cent falls after 2011, when the rate was held near zero by decision.[2] The level that identifies the exponent is therefore confounded with the mechanism that set it, and the exponent is what remains after averaging over both: averaging over a regime is not the same as controlling for it.

10.5 The Forward Smile

The *forward smile* is the smile of an option on the ratio $S_{t+\tau}/S_t$, struck at the money at a future date t. Cliquets, ratchets and most structured equity payoffs are strings of them, so a model that gets the forward smile wrong misprices a whole product family without any vanilla price revealing it.

It helps to keep two term structures of skew apart, because the chapter turns on the difference.

- The *maturity* term structure: how today's skew varies with the expiry T. This is quoted, so we can read it off the market.
- The *forward* term structure: how the skew of a fixed tenor τ varies with the start date t. Nothing quoted pins this down, so the model supplies it.

The market says something definite about each. The first *decays*, roughly as $1/\sqrt{T}$. The second is roughly *stationary*: a one-year skew a year forward looks much like today's one-year skew, which is what one would expect of a market whose behaviour does not depend on the date in the calendar.

The question of this section is whether one model can do both.

10.5.1 A Local Volatility Model With No Calendar In It

Start with the plainest local volatility model that has a skew at all, constant elasticity of variance,

$$dS_t = \sigma_0 \left(\frac{S_t}{S_0}\right)^{\beta-1} S_t \, dW_t, \qquad \beta < 1, \tag{10.6}$$

in which volatility rises as the level falls, which is the equity shape. Simulating it gives both term structures at once.[3] At $\sigma_0 = 25\%$ and $\beta = 0.7$, neither one decays: the five-year skew stays close to the three-month skew, and the one-year forward skew starting in four years stays close to today's.

Structure. Both flat readings are the same fact. Equation (10.6) is *time homogeneous* — its coefficient depends on the level and not on the date — so the model looks identical from every starting date, and its forward smile is its spot smile. There is nothing in it that could distinguish one calendar date from another, so nothing that could make either term structure slope.

[2] `estimation::the_extremes_of_the_level_come_from_two_policy_regimes`, on the same non-overlapping windows the regression uses.

[3] `localvol::Cev`, `the maturity ratio`, and `the forward-starting ratio`.

So CEV gets the forward term structure right and the maturity term structure wrong. It cannot be calibrated to a real surface at all.

10.5.2 What Calibration Puts In

Now calibrate. The one empirical fact we need is the $1/\sqrt{T}$ decay, so take the simplest surface with that shape,

$$\sigma_{\mathrm{BS}}(k,T) = \sigma + \psi(T)\,k, \qquad \psi(T) = \frac{\psi_1}{\sqrt{T+T_0}}, \tag{10.7}$$

with $k = \ln(K/F)$ and a short-maturity floor T_0, since no real skew is infinite at zero maturity. Nothing below depends on the smile being straight in k; only on ψ falling with T.

First, run Dupire's formula of chapter 9 on it. At zero rates that formula reads

$$\sigma^2_{\mathrm{loc}}(T,K) = \frac{2\,\partial C/\partial T}{K^2\,\partial^2 C/\partial K^2},$$

and writing $w(k,T) = \sigma^2_{\mathrm{BS}}T$ for the total implied variance and expanding to first order in the skew, the at-the-money slope of the local volatility comes out as

$$\left.\frac{\partial \sigma_{\mathrm{loc}}}{\partial k}\right|_{k=0} \;=\; \psi_1\,\frac{\frac{3}{2}T + 2T_0}{(T+T_0)^{3/2}}. \tag{10.8}$$

Two limits stand out. At $T = 0$ it is $2\psi_1/\sqrt{T_0}$, exactly twice the implied skew: the rule of two, recovered. At large T it is $\frac{3}{2}\psi_1/\sqrt{T}$, so the factor has fallen from 2 to $\frac{3}{2}$. Both ends, and the monotone passage between them, agree with a numerical Dupire to within five percent.[4]

Second, read (10.8) as a statement about *calendar* time. It is the local volatility at date T, and its skew falls as that date advances: by ten years it is a third of today's. The decay we fed in across the maturity axis has come back out along the time axis. The calibrated model is no longer time homogeneous, and could not be, because the surface it was asked to match is not.

Third, price the forward starting option. A short-dated option beginning at t diffuses through the slice of the surface at t, and there the rule of two applies in reverse: its implied skew is half the local skew at t. So

$$\psi_{\mathrm{fwd}}(t) \;\approx\; \tfrac{1}{2}\,\psi_1\,\frac{\frac{3}{2}t + 2T_0}{(t+T_0)^{3/2}} \;\longrightarrow\; \frac{3\psi_1}{4\sqrt{t}} \qquad \text{for large } t. \tag{10.9}$$

Remark (What that costs)**.** Equation (10.9) decays as $1/\sqrt{t}$, so doubling the forward start date multiplies the forward skew by $1/\sqrt{2}$. With an equity-shaped surface — $\sigma = 20\%$, $\psi_1 = -0.10$, T_0 one month — the forward skew is a fifth of today's after one year, an eighth after four, and under a tenth after ten. A cliquet is then priced off skews that have very nearly vanished, and because it is a string of them the error compounds along the string rather than cancelling.

Comparing the two subsections, a local volatility model has only one dial for both term structures. Dupire converts decay across maturities into decay along the date axis, so matching the observable one determines the unobservable one — and determines it wrongly.

[4] `localvol::SkewSurface`.

10.5.3 Whether the Trade Is Forced

It is fair to ask whether any model can escape that, or whether decaying skew simply implies a decaying forward skew. It does not, and the cleanest demonstration uses the model of §10.8,

$$dS_t = \sqrt{v_t}\, S_t\, dW_t, \qquad dv_t = \kappa(\theta - v_t)\, dt + \eta\sqrt{v_t}\, dZ_t, \qquad dW_t\, dZ_t = \rho\, dt, \tag{10.10}$$

taken here only for the measurement; what its parameters do is §10.8's business.

The point is what (10.10) does *not* contain: any function of t. Every parameter is a constant, so the model is as time homogeneous as CEV was, and by the same argument its forward smile cannot depend on the start date once v_t has reached its stationary distribution. Yet its skew still decays with expiry — because the decay is produced by the variance mean reverting *over the option's own life*, which is a longer journey for a longer option, rather than by any parameter that knows the date.

Simulating (10.10) at $v_0 = \theta = 4\%$, $\kappa = 2$, $\eta = 0.3$ and $\rho = -0.7$ confirms both halves.[5]

	skew at 5y expiry (share of the 3m skew)	1y skew starting in 4y (share of today's)
market	about 22%	about 100%
CEV, equation (10.6)	near 100%	near 100%
calibrated local volatility	fitted exactly	11%
Heston, equation (10.10)	20%	97%

The market row uses the $1/\sqrt{T}$ rule, which gives $\sqrt{0.25/5} = 22\%$, and stationarity for the forward column. Heston's 20% landing beside it is a better agreement than the model deserves in general — its long-dated skew decays as $1/T$ rather than $1/\sqrt{T}$, and §10.9 has a separate complaint about its short end — but over this range the shape is right, and the forward column is the one that matters here.

Structure. The separation is possible because the two models locate the skew's decay in different places. Local volatility has one state variable, so the only place a term structure can live is in the coefficient's dependence on t — and a coefficient that depends on t is precisely what a forward starting option is exposed to. Stochastic volatility has a second state variable, and can put the term structure in the *dynamics of that variable* instead, where a forward starting option, which begins after the variance has settled, never sees it.

This is the sharper form of the completeness argument above. The second Brownian motion is not merely something to hedge vega with; it is somewhere to keep a term structure that would otherwise have to be written into the calendar.

One caveat on the mechanism, since it is the stationary distribution and not time homogeneity alone that does the work. Start Heston hot, at $v_0 = 9\%$ against $\theta = 4\%$, and the forward skew is *not* flat in the first year: today's one-year skew is diluted by a variance that has not yet reverted,

[5] `heston::Heston`. Spot and forward smiles come from the same routine, so no scheme difference stands between the two columns.

and the skew has to travel before it settles. By one year it has, since $1/\kappa$ is six months, and successive start dates then agree to within a few percent. A model calibrated on a day when volatility is far from its long-run level therefore has a genuine forward term structure over the first few mean-reversion times, and only after that does stationarity take over.

10.6 Making Volatility Random

The diagnosis suggests the cure, and the forward smile above has already used it: give volatility its own source of randomness, and the skew no longer has to be manufactured by tying volatility to the level. What remains is choosing a model that does this, and the two classical ones answer different questions. SABR is the rates market's quoting model. It is built for one expiry's smile, with four parameters that each do a recognisable job, and it turns the backbone into an explicit dial, which is what this chapter is about. Heston, in §10.8, gives up some of that fit for a mean-reverting variance, which is what buys a term structure. SABR comes first because the separation of skew from backbone is easiest to see there.

Definition 10.9 (SABR)**.** The SABR model is

$$\begin{aligned} dF_t &= \alpha_t F_t^{\beta}\, dW_t, \\ d\alpha_t &= \nu \alpha_t\, dZ_t, \qquad dW_t\, dZ_t = \rho\, dt, \end{aligned}$$

with $\alpha_0 = \alpha$, and $\beta \in [0, 1]$.

The forward is a martingale, as it must be under the measure of chapter 6, and its volatility α_t is itself a driftless lognormal process. Four parameters, and the reason the model is used on trading desks is that each one does a recognisable job:

- α sets the *level* of the smile.
- ν, the volatility of volatility, sets its *curvature*. With $\nu = 0$ the volatility is deterministic and the smile has no smile in it.
- ρ, the correlation, *tilts* it. Negative ρ means the volatility rises as the forward falls, which makes low strikes expensive.
- β also tilts it — and this is the interesting one, because it tilts it for a different reason.

What makes SABR usable at all is that Hagan and coauthors found an asymptotic formula for the implied volatility, so a vanilla can be priced without any simulation. At the money it reduces to something short:

$$\sigma_{\text{ATM}} \approx \frac{\alpha}{F^{1-\beta}} \left(1 + \left[\frac{(1-\beta)^2}{24} \frac{\alpha^2}{F^{2-2\beta}} + \frac{1}{4} \frac{\rho \beta \nu \alpha}{F^{1-\beta}} + \frac{2 - 3\rho^2}{24} \nu^2 \right] T \right). \tag{10.11}$$

The leading term is the whole of the intuition and the bracket is a correction that matters at longer expiries. The full expression, for all strikes, is `Sabr::implied_vol` in `quant/src/sabr.rs`.

Remark (It is an expansion, and expansions expire)**.** Hagan's formula is asymptotic in the time to expiry, not exact. For short and moderate expiries it is very accurate. For long ones it degrades, and at low strikes it can imply a negative probability density — which by chapter 9 is an arbitrage, sitting inside the formula the market uses to quote. This is well known and is patched in practice, either by replacing the wings below some strike or by using an arbitrage-free version of the model. The industry's standard smile formula is not arbitrage-free, and it matters to know exactly where it fails.

Now look at β properly, because it is the parameter that controls the thing this chapter is about. Set $\nu = 0$ for a moment, so volatility is deterministic. The leading term of (10.11) gives

$$\sigma_{\text{ATM}}(F) = \alpha F^{\beta-1}, \qquad \text{so} \qquad \frac{\partial \sigma_{\text{ATM}}}{\partial \ln F} = (\beta - 1)\,\sigma_{\text{ATM}}.$$

So β *is* the backbone. At $\beta = 1$ the volatility of the forward is proportional to the forward, the model is lognormal, and the at-the-money volatility does not move at all when the market does — exactly sticky delta. At $\beta = 0$ the volatility is an absolute amount, so a rise in the forward means the same absolute movement is a smaller relative one, and the at-the-money volatility falls hard.

And ρ produces skew without producing backbone. So we have two dials that both tilt today's smile and disagree completely about tomorrow.

10.7 The Smile Does Not Determine the Dynamics

This is the punchline.

Fix a market: an at-the-money volatility of 25%, an ordinary downward skew, and a vol-of-vol of 0.35. Now fit the model with β held at various values, letting α and ρ adjust to reproduce that level and that skew. Every fit succeeds — there is always a correlation that reproduces any given skew — and the fitted smiles are nearly identical.[6]

β	ρ	backbone b	b/s	σ_{ATM} after a 10% rise
1.0	−0.34	0	0	25.00%
0.7	−0.13	−0.00075	1.25	24.30%
0.5	+0.02	−0.00125	2.09	23.84%
0.0	+0.37	−0.00251	4.19	22.72%

All four rows price today's options the same. They disagree by more than two volatility points about where the at-the-money volatility goes after a move that markets make routinely, and therefore they disagree about the delta to put on today.

Remark (A check falling out of the table)**.** The row at $\beta = 0.5$ has $\rho \approx 0$, so its volatility is very nearly a deterministic function of the forward — it is essentially a local volatility model. Its backbone-to-skew ratio comes out at 2.09.

[6] Every column of the table below, at every row: `the_smile_dynamics_backbone_table_holds_at_every_beta`.

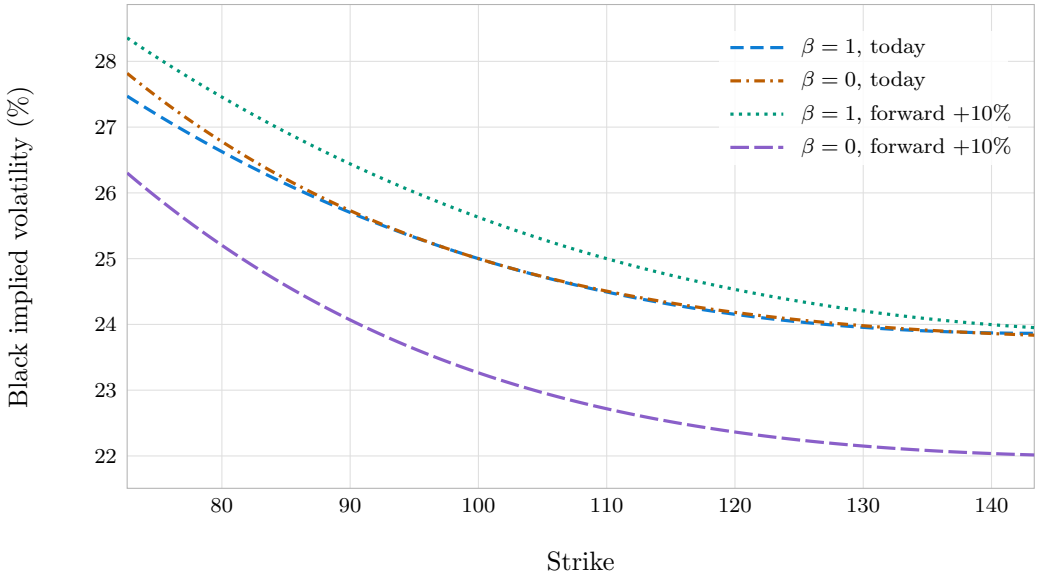

Figure 10.2: Two SABR models fitted to the same at-the-money volatility and the same skew. Today, their smiles lie on top of one another to within a fifth of a volatility point across the whole quoted range, so no set of vanilla quotes could distinguish them. Move the forward up by ten percent and they are more than two volatility points apart at the money. The parameter separating them, β, is invisible in today's prices and is the one that decides the hedge.

Drawn by `Sabr::implied_vol`, `Sabr::backbone` and `calibrate_alpha_rho`.

That is Theorem 10.7 arriving from a completely different direction. We derived $b = 2s$ from a short-maturity harmonic mean; here it emerges from Hagan's asymptotic expansion of a stochastic volatility model that happens to have no stochastic volatility left in it. Two independent routes to the same factor of two is the kind of agreement that makes a result trustworthy. The check is the_smile_does_not_determine_the_dynamics.

So how is β chosen, if the smile will not choose it? Not from vanilla prices, because they are silent. In practice it is fixed by judgement about the market — close to 1 for an equity index, lower for rates, near 0 where rates can be negative and an absolute volatility is the only sensible one — or it is inferred from the historical relationship between the at-the-money volatility and the level, which is precisely a measurement of the backbone. It is then usually held fixed, and α, ρ and ν are re-marked daily.

Remark (The general lesson)**.** Calibration is not the same as identification. A model can be fitted perfectly and still be undetermined in the direction that matters, because the instruments used to fit it are blind to that direction. Vanilla options are a set of one-date distributions, so they constrain the marginals and nothing else — which is Gyongi's theorem again, now doing damage rather than favours.

The practical consequence: if a parameter matters for your product and is invisible in your calibration set, then either the calibration set has to be enlarged — with forward starting options, or barriers, or historical dynamics — or the parameter has to be marked by hand and treated as a model risk, with a number attached.

10.8 Heston, and What a Term Structure Costs

SABR has one weakness that follows from its own definition: α_t is a driftless lognormal process, so it has no mean reversion and no long-run level. Over a long horizon it wanders arbitrarily far, and the model has nothing to say about a term structure of volatility. In practice this is handled by fitting a separate set of SABR parameters to every expiry, which prices every slice correctly and leaves the slices with no relationship to one another — fine for interpolating vanillas, useless for anything spanning two dates.

The other classical choice fixes exactly this.

Definition 10.10 (Heston)**.**

$$\begin{aligned} dS_t &= \mu S_t\,dt + \sqrt{v_t}\,S_t\,dW_t, \\ dv_t &= \kappa(\theta - v_t)\,dt + \eta\sqrt{v_t}\,dZ_t, \qquad dW_t\,dZ_t = \rho\,dt. \end{aligned}$$

The variance v_t is a mean-reverting square-root process: it is pulled towards a long-run level θ at a speed κ, and the $\sqrt{v_t}$ in its own diffusion keeps it non-negative, since the randomness switches off as it approaches zero. Those two features are what buy a term structure — θ sets where volatility goes in the long run, κ how fast it gets there — and they are what SABR lacks.

Heston is tractable for a reason that chapter 14 gives for any model being solvable at all. Its characteristic function

$$\phi(u) = \mathbb{E}\big[e^{iu\ln S_T}\big] = \exp\big(A(u,T) + B(u,T)v_0\big)$$

is available in closed form: the exponent is *affine* in the state, and finding A and B reduces to a pair of ordinary differential equations rather than a partial one. Option prices then follow by a single numerical integral. This is the affine structure that chapter 14 makes explicit and chapter 12 puts to work on interest rates.

The trade one is making between the two is roughly: SABR gives a better fit to one expiry's smile with parameters a trader can interpret, and Heston gives a consistent story across expiries at the cost of a stiffer fit to any one of them. Neither fits the whole surface exactly, and for a desk that has to reprice the market to the penny, that is a problem — which is where the next chapter starts.

10.9 The Short-Dated Skew, and Rough Volatility

Before leaving stochastic volatility there is one failure the two models share, and it is not a matter of degree. It concerns what happens to the skew as the expiry shrinks.

Measure the at-the-money skew s of §10.2 at a range of expiries. In SABR — and in Heston, and in classical stochastic volatility generally — it settles down to a constant as $T \to 0$. That is not an artefact of Hagan's expansion; it is a property of the model class. The skew comes from the correlation between the forward and its volatility, and over a short horizon the volatility simply has not had time to move far enough for that correlation to produce more than a bounded tilt.

The market disagrees, and not subtly. Observed at-the-money skews grow as the expiry shrinks, roughly as a power law, and short-dated options are far more skewed than any classical model will produce.

The fix that emerged in the last decade is to make the volatility *rough*.

Definition 10.11 (Rough volatility, informally)**.** Let the log variance be driven by a fractional Brownian motion with Hurst parameter H, or by its Riemann-Liouville version, which the first calculation below defines and which is the one used in practice,

$$\frac{dS_t}{S_t} = \sqrt{v_t}\, dW_t, \qquad \log v_t \text{ driven by } B^H,$$

with H small — estimates from realised volatility put it near 0.1, against the $H = 1/2$ of an ordinary diffusion.

A smaller H means a rougher path: one that reverses direction more violently and has less memory of its recent level. And the payoff is exactly the quantity that was wrong. A model of this kind produces an at-the-money skew behaving like

$$s(T) \;\sim\; T^{H-1/2},$$

which for $H = 0.1$ is $T^{-0.4}$ — a power law that blows up as the expiry shrinks, which is what the market shows. Setting $H = 1/2$ returns the constant of the classical case, so the classical models are the corner of this family in which the volatility is exactly as smooth as the price.

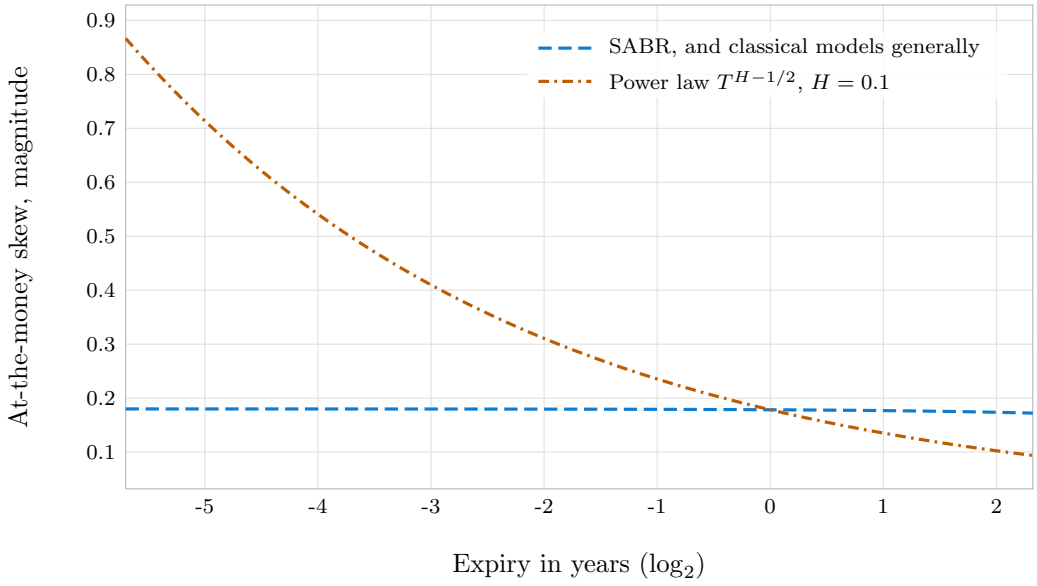

Figure 10.3: The at-the-money skew against expiry, from a week to five years. The SABR curve is computed; the power law beside it is the shape rough volatility predicts, drawn as a reference and anchored to agree with SABR at one year, so the comparison is about the term structure rather than the level. SABR moves from 0.180 to 0.172 across the whole range — it is, to the eye, a flat line. The power law spans nearly a factor of ten, and at a one week expiry it is five times as steep. No amount of recalibrating the classical model closes that gap, because the gap is in the shape.
Drawn by `Sabr::skew`.

Remark (Why this does not break anything)**.** Chapter 2 showed that fractional Brownian motion is not a semimartingale for $H \neq 1/2$, and that a *price* following one would admit arbitrage. To be clear, no such thing is happening here.

The roughness is in v, and v is not a traded asset. The price is still

$$S_t = S_0 + \int_0^t \sqrt{v_s}\, S_s\, dW_s,$$

an ordinary continuous semimartingale, and chapter 2's representation theorem never asked the volatility to be anything but adapted with $\int \sigma^2 ds$ finite. Every result in these notes therefore survives unchanged. Rough volatility is not an alternative to the framework; it is a choice of σ within it.

Remark (What it costs)**.** Fractional Brownian motion has memory, so the model is not Markov in any finite set of state variables. Chapter 12 will meet exactly this difficulty in a different guise — the Heath-Jarrow-Morton framework is also non-Markov, for a different reason — and the consequence is the same in both cases: no low-dimensional grid, no backward induction, and therefore real difficulty with anything callable.

The escape is also the same in both cases. Chapter 12 recovers a Markov state by insisting the volatility kernel decay exponentially, since $e^{-\kappa(T-u)}$ is the one shape that factorises into a maturity piece and a history piece. Rough volatility recovers one by approximating its fractional kernel with a *sum* of exponentials, each of which is Markov by that same argument, giving a model with several ordinary factors. The calculation below shows why that is possible, and what it does and does not reproduce. The trick that makes rates models computable is the trick that makes rough volatility models computable.

There is a further parallel. Chapter 12's excursion into affine models explains why certain models reduce a partial differential equation to ordinary ones. The rough Heston model retains an affine structure of exactly that kind, except that the function entering its exponent solves a fractional Riccati equation. That is a Volterra integral equation rather than an ordinary differential equation, so the machinery does not carry over unchanged; the second calculation below derives it and shows how the sum of exponentials restores the ordinary equations.

Calculation 10.12 (A rough kernel is a sum of exponentials, and how many it takes)**.** The rough process is the Riemann-Liouville one, $X_t = \int_0^t K(t-s)\, dW_s$ with the kernel $K(t) = t^{H-1/2}/\Gamma(H+\frac{1}{2})$. It is fractional Brownian motion with its past before time zero cut off: Mandelbrot and Van Ness write B^H as the same integral over $s \in (-\infty, t]$, with a compensating term $(-s)^{H-1/2}$ inside the kernel, and the part over $s < 0$ is smooth in t, so it changes nothing about the roughness. What the cut-off buys is a process that starts at zero, is adapted, and has variance $t^{2H}/\big(2H\,\Gamma(H+\frac{1}{2})^2\big)$ in closed form. This is the version rough volatility models use, and the "driven by B^H" of the definition above means it. For $H < \frac{1}{2}$ the Gamma integral $\int_0^\infty x^{s-1}e^{-xt}\,dx = \Gamma(s)\,t^{-s}$, taken with $s = \frac{1}{2} - H$, writes the kernel as a positive mixture of exponentials:

$$K(t) = \int_0^\infty e^{-xt}\,\mu(dx), \qquad \mu(dx) = \frac{x^{-H-1/2}\,dx}{\Gamma(H+\frac{1}{2})\,\Gamma(\frac{1}{2}-H)}.$$

Substituting, $X_t = \int Y_t^x\,\mu(dx)$ with $Y_t^x = \int_0^t e^{-x(t-s)}\,dW_s$, and each Y^x solves $dY^x = -x\,Y^x\,dt + dW$: an Ornstein-Uhlenbeck factor, all of them driven by the same Brownian motion. The memory of X is the infinite family of them.

Truncate μ to n atoms with weights c_i and rates x_i, so that $X^n = \sum_i c_i Y^i$. The vector $(Y^1, \dots, Y^n)$ is Markov, and X^n is a function of it. Cut the measure into geometric cells and give each cell one exponential, with the cell's mass as weight and its mean as rate, so that the total mass and first moment of μ are kept exactly. Only rates between $1/T_{\max}$ and $1/t_{\min}$ can matter: a smaller rate is constant over the longest horizon and one atom, on $[0, x_{\min}]$, stands for all of them, while a larger one has died before the shortest horizon. The measure μ is a pure power law, so it is self-similar, and a cell with a fixed ratio between its edges has the same relative error at every scale. Nodes at a fixed ratio r therefore need $n = \ln(T_{\max}/t_{\min})/\ln r$ of them: the number of atoms grows with the *logarithm* of the ratio of horizons, not with the ratio. Holding $r = 2.15$, twelve, twenty-four and forty-eight atoms cover horizon ratios of 10^2, 10^6 and 10^{14} with the same worst kernel error, about 1.6 per cent.[7]

The kernel converges quickly. At $H = 0.1$, twenty-four exponentials with rates from 0.1 to 10^7 a year reproduce the kernel to within about two per cent from a day to five years; eight leave it about ten per cent low and four, at worst, about half.[8]

The roughness converges slowly, and it is the part that matters. The variance of X_t^n is available in closed form,

$$\operatorname{Var} X_t^n = \sum_{i,j} c_i c_j \, \frac{1 - e^{-(x_i + x_j)t}}{x_i + x_j}, \qquad \text{against} \qquad \operatorname{Var} X_t = \frac{t^{2H}}{2H\,\Gamma(H + \tfrac{1}{2})^2},$$

and the exponent $2H$ in the second is what carries the roughness. The kernel is integrable at zero but its square is not bounded there, $K^2 \sim s^{2H-1}$, so a large share of the variance at time t was contributed at times *shorter* than any node's time scale. The part contributed before τ is $\tau^{2H}/(2H\,\Gamma(H + \frac{1}{2})^2)$, a share $(\tau/t)^{2H}$ of the total, and a truncated kernel is smooth below $\tau = 1/x_{\max}$, so the share it misses scales like $(x_{\max}\, t)^{-2H}$ — at $H = 0.1$ and one week it is about 0.83 times that, from $x_{\max} = 2\,000$ to 10^9 a year — which is slow when $2H$ is small. Reaching a missing share ε needs $x_{\max}\, t_{\min}$ of order $\varepsilon^{-1/(2H)}$, about 10^5 for ten per cent at $H = 0.1$: an additive term in the number of atoms, set by the accuracy wanted, on top of the logarithm above.[9] Measured at one week: with the fastest rate at 2 000 a year, about eight a trading day, the local exponent of the variance is 0.33 against the true 0.2, so the process is visibly too smooth, and with the fastest rate at 10^7 it is about 0.21. The exponent of the kernel's integral, a smoother quantity, is already within about 0.03 of $H + \frac{1}{2}$ at the lowest ceiling.[10]

So "behaves like a rough one over the horizons that matter" has a condition attached: the fastest node has to sit orders of magnitude above the reciprocal of the shortest horizon, which makes the factor system stiff, with time scales down to seconds, and below the fastest node's time scale the process is smooth whatever H is.

Calculation 10.13 (The fractional Riccati equation, and why the factor lift is affine)**.** Rough Heston is Heston's variance with the constant kernel of its integral form replaced by a fractional

[7] `the_error_is_the_same_at_every_scale_for_a_fixed_node_ratio`, and `rough::ExponentialSum::geometric`, with `the_kernel_is_a_mixture_of_exponentials` checking the identity itself and `the_ordinary_factors_have_the_variance_the_kernel_gives` simulating the factors and matching the closed form below.

[8] `a_few_nodes_reproduce_the_kernel_over_the_horizons_that_matter`.

[9] `the_missing_variance_scales_like_the_fastest_rate_to_the_minus_two_h`.

[10] `the_roughness_exponent_needs_nodes_far_above_the_shortest_horizon`.

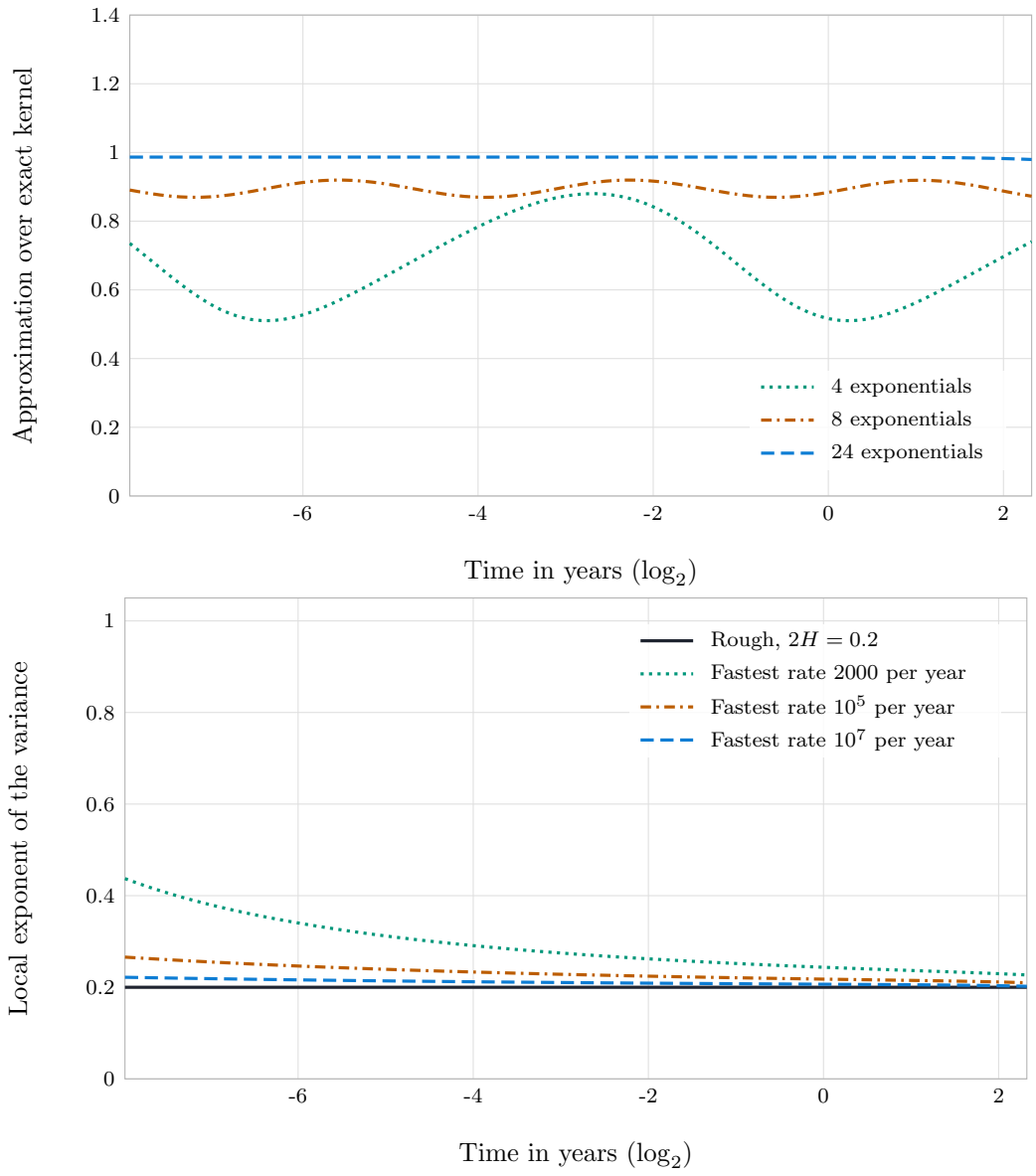

Figure 10.4: Above: the ratio of the sum-of-exponentials kernel to the exact rough kernel, $H = 0.1$, from a day to five years, for four, eight and twenty-four exponentials with rates from 0.1 to 10^7 a year. The ratio is one where the approximation is exact; twenty-four sits within about two per cent of it throughout. Below: the local exponent of the variance, $d \log \mathrm{Var} / d \log t$, which is $2H = 0.2$ for a rough process. Every approximation is smoother than that at short times, and the error retreats to shorter times only as the fastest rate is raised — the kernel is easy to reproduce and the roughness is not.

Drawn by `ExponentialSum::geometric`.

one. Heston's $dV = \lambda(\theta - V)\,dt + \nu\sqrt{V}\,dB$ reads $V_t = V_0 + \int_0^t \lambda(\theta - V_s)\,ds + \int_0^t \nu\sqrt{V_s}\,dB_s$, and putting the kernel K in front of both integrals gives

$$V_t = V_0 + \int_0^t K(t-s)\Big[\lambda(\theta - V_s)\,ds + \nu\sqrt{V_s}\,dB_s\Big], \qquad K(t) = \frac{t^{\alpha-1}}{\Gamma(\alpha)}, \quad \alpha = H + \tfrac{1}{2},$$

with $d\log S = -\frac{1}{2}V\,dt + \sqrt{V}\,dW$ and $dW\,dB = \rho\,dt$. This is the model of El Euch and Rosenbaum, who obtain it as the scaling limit of Hawkes-process models of order flow with a heavy-tailed kernel. Three things identify it. At $H = \frac{1}{2}$ the kernel is one and this is the Heston model of §10.8. The noise term is a rough process, since its kernel is the Riemann-Liouville kernel of the previous calculation with $\nu\sqrt{V_s}$ as the integrand, so V has roughness H. And the drift passes through the same kernel, which adds $\lambda\theta\int_0^t K = \lambda\theta\, t^\alpha/\Gamma(\alpha+1)$ to the mean; using one kernel for both is what keeps V an affine Volterra process, and so the exponent below affine. This is a different model from the definition above, in which $\log v$ is driven by B^H: here the variance itself is the square-root process.

Replace K by an exponential sum $K_n(t) = \sum_i c_i e^{-x_i t}$ as in the previous calculation, and set $U_t^i = \int_0^t e^{-x_i(t-s)}\big(-\lambda V_s\,ds + \nu\sqrt{V_s}\,dB_s\big)$. Then

$$V_t = g(t) + \sum_i c_i U_t^i, \qquad g(t) = V_0 + \lambda\theta\int_0^t K_n, \qquad dU^i = (-x_i U^i - \lambda V)\,dt + \nu\sqrt{V}\,dB,$$

so the state $(\log S, U^1, \dots, U^n)$ is Markov and V is affine in it. The process $M_t = \mathbb{E}[S_T^a \mid \mathcal{F}_t]$ is a martingale, and for $M = e^Y$ Itô's lemma gives $dM = M\,(dY + \frac{1}{2}\,d\langle Y\rangle)$, so M is a martingale exactly when the drift of Y plus half its quadratic variation vanishes. Because the drift is affine in the state and the squared diffusion coefficient V is too, the exponentials of affine functions are preserved by the generator, and an exponent of that form can satisfy the condition with coefficients that depend on $T - t$ alone. So look for $M_t = \exp\big(a\log S_t + \sum_i \psi_i(T-t)\,U_t^i + \varphi(T-t)\big)$, with $\psi_i(0) = \varphi(0) = 0$ so that $M_T = S_T^a$, and require that condition. Since $\log S_0 = 0$ and $U_0^i = 0$, the moment is $M_0 = e^{\varphi(T)}$. Write $\Psi = \sum_i \psi_i$. The terms multiplying V collect into

$$F(a,\Psi) = \tfrac{1}{2}(a^2 - a) + (a\rho\nu - \lambda)\Psi + \tfrac{1}{2}\nu^2\Psi^2,$$

and matching the coefficient of each U^i and the remaining constant gives

$$\psi_i' = -x_i\psi_i + c_i\,F(a,\Psi), \qquad \psi_i(0) = 0, \qquad \mathbb{E}[S_T^a] = \exp\Big(\int_0^T F\big(a,\Psi(\tau)\big)\,g(T-\tau)\,d\tau\Big).$$

That is n ordinary equations, coupled only through Ψ, exactly the structure of chapter 14. Zero drift makes M a local martingale, and it is the moment only if it is a true one, which needs the Riccati solution to exist on $[0,T]$ so that no moment explosion intervenes: this restricts a and T, and holds for $a \in [0,1]$. Solving each by variation of constants, $\psi_i(\tau) = \int_0^\tau c_i e^{-x_i(\tau-s)}F(a,\Psi(s))\,ds$, and summing,

$$\Psi(\tau) = \int_0^\tau K_n(\tau - s)\,F\big(a,\Psi(s)\big)\,ds,$$

which is a Volterra equation with the kernel K_n. Let $K_n \to K$ and it becomes

$$h(\tau) = \int_0^\tau K(\tau - s)\,F\big(a,h(s)\big)\,ds, \qquad \text{that is} \qquad D^\alpha h = F(a,h),$$

the fractional Riccati equation of El Euch and Rosenbaum, with $\mathbb{E}[S_T^a] = \exp\big(\int_0^T F(a, h(\tau))\, g(T-\tau)\, d\tau\big)$ and $g(t) = V_0 + \lambda\theta\, t^\alpha/\Gamma(\alpha+1)$. The lift is this equation with the memory turned into n state variables.

Three checks, none of which uses the lift's own algebra.[11] At $H = \frac{1}{2}$ the equation is $h' = F(a, h)$, whose closed form is Heston's, and the solver reproduces it and the exponent to 10^{-6}.[12] At $a = 1$ the source term $F(1, 0)$ vanishes, so $h = 0$ and $\mathbb{E}[S_T] = 1$ at every H: the price is a martingale.[13] And simulating the factors themselves, with no Riccati equation anywhere in the code, reproduces the lifted moment to within one per cent.[14]

What roughness changes is visible in the number. At $H = 0.1$ the Volterra solver gives $h(1) = -0.1036$ against -0.1026 at $H = \frac{1}{2}$, and near $\tau = 0$ the rough solution moves like $\tau^{H+1/2}$ rather than like τ. The lift approaches it as the sum is refined: with 12, 24, 48 and 96 exponentials it gives -0.0997, -0.1026, -0.1033 and -0.1035, the error falling by about a factor of four at each doubling.[15] The two routes differ in cost. The Volterra equation is solved by a predictor-corrector whose work grows with the square of the number of steps, since the whole history enters every step; the lift costs n times the steps, at the price of the stiff fast factors of the previous calculation.

References

- Hagan, P. S., Kumar, D., Lesniewski, A. S., & Woodward, D. E. (2002). Managing smile risk. *Wilmott Magazine*, 1, 84–108.
- Berestycki, H., Busca, J., & Florent, I. (2002). Asymptotics and calibration of local volatility models. *Quantitative Finance*, 2(1), 61–69.
- 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(2), 327–343.
- Hagan, P. S., Kumar, D., Lesniewski, A., & Woodward, D. (2014). Arbitrage-free SABR. *Wilmott*, 2014(69), 60–75.
- Gatheral, J., Jaisson, T., & Rosenbaum, M. (2018). Volatility is rough. *Quantitative Finance*, 18(6), 933–949.
- Bayer, C., Friz, P., & Gatheral, J. (2016). Pricing under rough volatility. *Quantitative Finance*, 16(6), 887–904.
- El Euch, O., & Rosenbaum, M. (2019). The characteristic function of rough Heston models. *Mathematical Finance*, 29(1), 3–38.
- Abi Jaber, E., & El Euch, O. (2019). Multifactor approximation of rough volatility models. *SIAM Journal on Financial Mathematics*, 10(2), 309–349.

[11] `roughheston::RoughHeston::fractional_riccati` and `roughheston::RoughHeston::lifted`, at $\lambda = 0.3$, $\theta = V_0 = 0.02$, $\nu = 0.3$, $\rho = -0.7$, $a = \frac{1}{2}$, one year.

[12] `at_one_half_the_fractional_equation_is_hestons_riccati`.

[13] `the_price_is_a_martingale_at_every_hurst_parameter`.

[14] `the_lifted_moment_matches_a_simulation_of_the_factors`.

[15] `the_factor_lift_converges_to_the_fractional_riccati`.

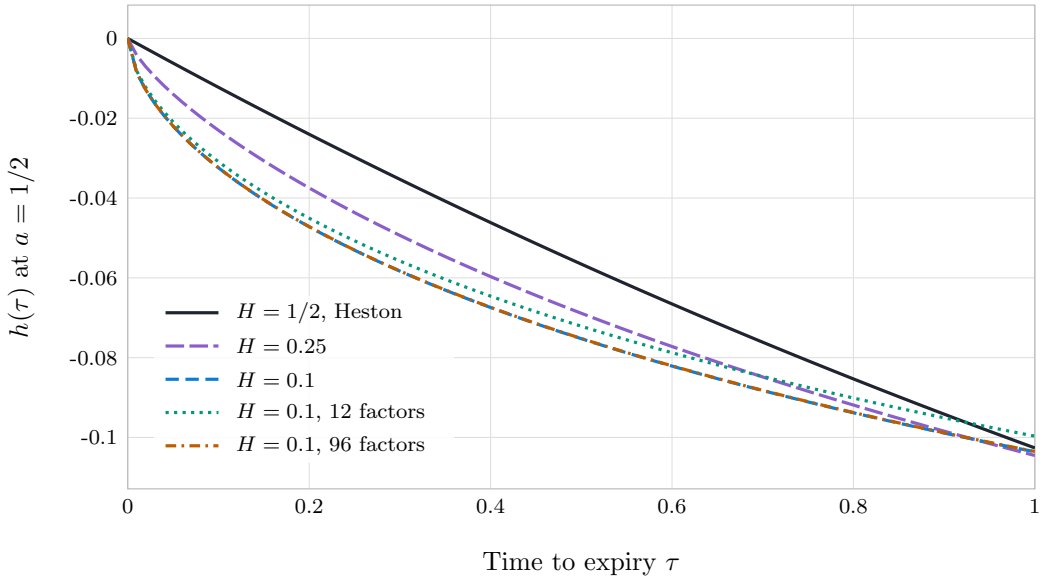

Figure 10.5: The solution h of the Riccati equation at $a = \frac{1}{2}$ against time to expiry, for the classical model, $H = 0.25$ and $H = 0.1$, and for the factor lift of the $H = 0.1$ case with 12 and 96 exponentials. Roughness bends the curve at short times, where the classical solution is close to a straight line. Twelve factors fall visibly short of the rough curve at longer times, and ninety-six lie on top of it: the lift is the same equation with its memory unrolled into state variables.
Drawn by `RoughHeston::fractional_riccati` and `RoughHeston::lifted`.

Chapter 11

Local-Stochastic Volatility

We combine the two models of the previous chapters. Local volatility fits every European option exactly but moves the smile wrongly; stochastic volatility moves it sensibly but does not fit. We derive the condition under which a model can do both at once, find that it defines the model implicitly rather than explicitly, and describe how that circularity is broken in practice.

11.1 Taking the Good Half of Each

Two chapters have left us with two models and a complaint about each.

Local volatility fits the market's European options perfectly, because Dupire's formula constructs it to. Its dynamics are wrong: the backbone comes out at twice the skew, which is outside the range the market actually behaves in, so the delta is systematically wrong.

Stochastic volatility has sensible dynamics, because the backbone and the skew are controlled by different parameters and can be set independently. It does not fit: SABR with four parameters cannot pass exactly through a few dozen quoted strikes and expiries, and Heston with five does no better. For a desk that has to reprice the market to the penny — because it is marking a book against those very quotes — a fit that is close is not a fit.

The idea is to keep the fit of the first and the dynamics of the second by multiplying them together.

Definition 11.1 (Local-stochastic volatility)**.** A local-stochastic volatility model is

$$\begin{aligned} \frac{dS_t}{S_t} &= (r-q)\,dt + L(t,S_t)\,\sqrt{v_t}\,dW_t, \\ dv_t &= \alpha(t,v_t)\,dt + \beta(t,v_t)\,dZ_t, \qquad dW_t\,dZ_t = \rho\,dt, \end{aligned}$$

where v is whatever stochastic variance process one likes — Heston's, say — and L, a deterministic function of time and level, is called the *leverage function*.

The instantaneous volatility is now a product of two things. The factor $\sqrt{v_t}$ is random and carries the dynamics: the vol-of-vol, the correlation, the mean reversion, everything that decides how the smile moves and what a forward smile looks like. The factor $L(t, S_t)$ is deterministic and carries the fit: it is a dial at each point of the plane, to be turned until the model's European prices agree with the market's.

The division of labour is the whole design. The question is whether it works — whether there really is a choice of L making the fit exact, whatever the stochastic part is doing. There is, and chapter 9 has already given us the tool to find it.

11.2 The Calibration Condition

Theorem 11.2 (The leverage function)**.** *A local-stochastic volatility model reproduces the market's European option prices at all strikes and expiries if and only if*

$$L^2(t,K)\,\mathbb{E}\big[\,v_t \,\big|\, S_t = K\,\big] \;=\; \sigma_{loc}^2(t,K), \tag{11.1}$$

where σ_{loc} is the Dupire local volatility of the market's own surface. Equivalently,

$$L(t,K) = \frac{\sigma_{loc}(t,K)}{\sqrt{\mathbb{E}\big[\,v_t \,\big|\, S_t = K\,\big]}}. \tag{11.2}$$

Proof. Chapter 9 established two facts and this is their intersection.

Gyongi's theorem says that a process with random volatility has the same one-dimensional distributions — and therefore the same European option prices — as the local volatility model whose squared volatility is the conditional expectation of the true squared volatility. Here the instantaneous volatility of S is $L(t,S_t)\sqrt{v_t}$, so the mimicking local volatility is

$$\tilde\sigma^2(t,K) = \mathbb{E}\Big[L^2(t,S_t)\,v_t \,\Big|\, S_t = K\Big] = L^2(t,K)\,\mathbb{E}\big[v_t \,\big|\, S_t = K\big],$$

where $L^2(t,S_t)$ comes out of the conditional expectation because, conditional on $S_t = K$, it *is* the number $L^2(t,K)$ — it is a deterministic function of the thing being conditioned on.

Dupire's formula says that a set of European prices determines exactly one local volatility surface, namely σ_{loc}. So the model matches the market if and only if its mimicking local volatility is the market's, which is (11.1). □

Dupire's local volatility is not the model's volatility. It is an average of the model's volatility, taken over all the ways the market could have arrived at the level K at time t. The leverage function is therefore a *correction factor*: at each point, take the local volatility the market demands, and divide by the volatility the stochastic part is already delivering there on average. Whatever is left over is what L must supply.

Example 11.1 (The two extremes)**.** Two sanity checks, both of which should be reassuring.

If the variance is not random at all — say $v_t \equiv 1$ — then $\mathbb{E}[v_t \mid S_t = K] = 1$ and (11.2) gives $L(t,K) = \sigma_{\text{loc}}(t,K)$. The leverage function is the local volatility surface, and the model is the local volatility model of chapter 9. Nothing has been gained and nothing lost.

If instead the stochastic volatility model already fits the market on its own, then by the same theorem its own mimicking local volatility is already σ_{loc}, so $\mathbb{E}[v_t \mid S_t = K] = \sigma^2_{\text{loc}}(t, K)$ and $L \equiv 1$. The leverage function switches itself off when it is not needed.

Between the two, L measures exactly the part of the market's smile that the stochastic volatility model failed to produce by itself. A well-chosen stochastic part leaves L close to one and smooth; a badly chosen one leaves L doing all the work, which puts us back in chapter 9 with extra steps.

The one thing in (11.2) that is easy to read past is the word *conditional*. It would be much more convenient if the denominator were $\mathbb{E}[v_t]$, the plain average variance, which one could compute once and be done with. It is not, and the difference is not small. The following example is small enough to compute by hand and shows why.

Example 11.2 (Where the conditioning bites)**.** Take the crudest possible stochastic volatility: at time zero a coin is tossed, and the forward is lognormal for the rest of its life with volatility either 15% or 35%, each with probability one half. So v is random but constant along each path, and $\mathbb{E}[v] = \frac{1}{2}(0.15^2 + 0.35^2) = 0.0725$, whose square root is 26.9%.

Now ask for $\mathbb{E}[v \mid S_T = K]$ at $T = 1$, with the forward at 100. Observing where the underlying ended is evidence about which coin came up, and Bayes' rule turns that evidence into a posterior. Writing ϕ_{lo} and ϕ_{hi} for the two lognormal densities,

$$\mathbb{E}\big[v \mid S_T = K\big] = \frac{v_{\text{lo}}\,\phi_{\text{lo}}(K) + v_{\text{hi}}\,\phi_{\text{hi}}(K)}{\phi_{\text{lo}}(K) + \phi_{\text{hi}}(K)}.$$

Evaluating:

K	60	80	100	125	160
$\mathbb{P}(\text{high} \mid S_T = K)$	0.98	0.51	0.30	0.51	0.96
$\sqrt{\mathbb{E}[v \mid S_T = K]}$	34.7%	27.1%	22.9%	27.1%	34.4%

Read the middle row first. If the forward has barely moved, it was probably a quiet path, so the low volatility state is more likely than it was a priori. If the forward has moved a long way, it was almost certainly the noisy one.

The bottom row is what (11.2) divides by, and it ranges from 22.9% to 34.7% — while the unconditional figure is a flat 26.9% everywhere. Using the unconditional average would leave the leverage function wrong by four volatility points at the money and eight in the wing, in opposite directions. The model would then miss the market by roughly that much, which is a hundred times any bid-offer spread.

Notice also what the conditional expectation is shaped like: a hump, low in the middle and rising in both wings. So L, which divides by its square root, is high in the middle and low in the wings. That is the general shape, and it has a plain reading — the stochastic part already supplies plenty of volatility in the tails, because that is what a random volatility does, so the leverage function has less work to do there and more in the middle.

Remark (Why the mixing matters)**.** It is tempting to conclude that since L can always be found, the choice of the stochastic part does not matter. The opposite is true, and the reason is the point of the whole construction.

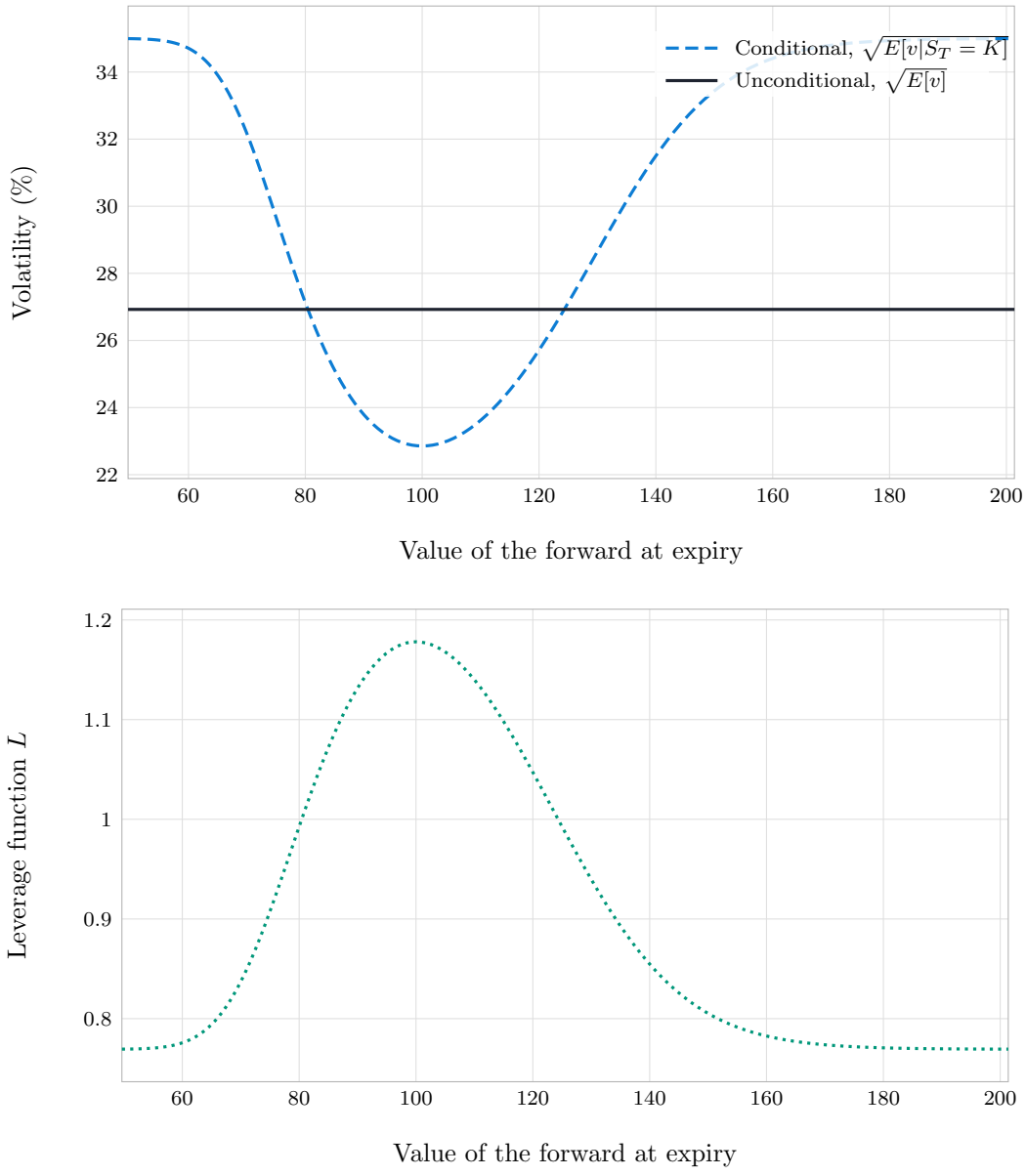

Figure 11.1: Above: the conditional volatility $\sqrt{\mathbb{E}[v \mid S_T = K]}$ against the unconditional $\sqrt{\mathbb{E}[v]}$ that it is tempting to use instead. Below: the leverage function this forces, against a flat target local volatility. The two differ by four volatility points at the money and eight in the wing, in opposite directions, which is the difference between a model that reprices the market and one that misses it by a hundred bid-offer spreads.

Drawn by `TwoRegime::conditional_variance` and `TwoRegime::leverage`.

Everything about the model's *dynamics* — the backbone, the forward smile, the vega — comes from the stochastic part, because L is deterministic and contributes no randomness at all. And the fit is exact regardless. So the stochastic part is now a free choice that is invisible in the calibration and decisive for everything else, which is chapter 10's lesson repeating one level up. Desks parametrise it explicitly with a *mixing weight*: dial it to zero and the model degenerates to pure local volatility, dial it up and the vol-of-vol takes over, and the vanilla prices do not move either way. What moves is every exotic in the book.

11.3 The Circularity, and How It Is Broken

Equation (11.2) looks like a formula for L. It is not, quite, and the reason deserves a clear look.

The right-hand side contains $\mathbb{E}[v_t \mid S_t = K]$, a conditional expectation taken under the law of the process (S, v). But the law of S depends on L, because L is in its diffusion coefficient. So L appears on both sides: it is defined in terms of a distribution that it itself determines.

Remark (What kind of equation this is)**.** An equation whose coefficients depend on the law of its own solution is called a McKean-Vlasov, or mean-field, stochastic differential equation. It is a genuinely different object from an ordinary SDE. The usual existence and uniqueness theorems do not apply, and for local-stochastic volatility in particular the question of whether a solution exists for a given market surface is, at the time of writing, not settled in general. Models are calibrated and traded anyway.

There are two standard ways to break the circle, and both work by stepping forward in time, so that at each step the conditional expectation is computed from a distribution that is already known.

The forward equation approach

Write $\psi(t, S, v)$ for the joint density of (S_t, v_t). It satisfies a two-dimensional Fokker-Planck equation whose coefficients involve L, and the conditional expectation we need is a ratio of integrals of it:

$$\mathbb{E}\big[v_t \mid S_t = K\big] = \frac{\int v\, \psi(t, K, v)\, dv}{\int \psi(t, K, v)\, dv}.$$

So march the density forward on a grid. At each time step, the density is known from the previous step; compute the conditional expectation from it by the ratio above; set L on that time slice by (11.2); use that L to take the density one step further. The circularity is broken because $L(t, \cdot)$ is only ever needed at a time when $\psi(t, \cdot, \cdot)$ has already been computed.

This is accurate and it is fast in two dimensions. It does not survive a large number of factors, because the grid does not.

The particle method

The alternative, due to Guyon and Henry-Labordere, replaces the density with a cloud of simulated paths.

1. Start N particles at (S_0, v_0). Take N large — tens of thousands.
2. At each time step t, having the particles' current positions (S_t^i, v_t^i):
   (a) estimate the conditional expectation from the cloud itself, by a kernel-weighted average over particles whose S is near K,
   $$\widehat{\mathbb{E}}\big[v_t \mid S_t = K\big] = \frac{\sum_{i=1}^N \delta_h(S_t^i - K)\, v_t^i}{\sum_{i=1}^N \delta_h(S_t^i - K)},$$
   where δ_h is a smoothing kernel of bandwidth h;
   (b) set $L(t,\cdot)$ on this time slice by (11.2);
   (c) evolve every particle one step using that L.
3. The leverage function is the collection of the slices.

The usual first choice for the kernel is Gaussian,

$$\delta_h(x) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{x^2}{2h^2}\right), \tag{11.3}$$

so a particle sitting exactly at K counts fully, one a bandwidth away counts about six tenths as much, and one three bandwidths away counts for nothing that matters. Compactly supported kernels — Epanechnikov, $\delta_h(x) \propto (1 - x^2/h^2)^+$, or the quartic — are common in production for the plain reason that they let the sum skip every particle outside the window. The estimator is Nadaraya-Watson kernel regression, the standard nonparametric way of reading a conditional expectation off a scatter of points; nothing about it is specific to finance.

The bandwidth cannot be a constant, because the cloud spreads as it goes. A common choice scales as

$$h_t \;\propto\; \sigma S_0 \sqrt{t}\, N^{-1/5},$$

the $\sqrt{t}$ tracking the width of the cloud and the $N^{-1/5}$ being the rate at which a kernel bandwidth should shrink as the sample grows — slowly, because a bandwidth that shrinks too fast leaves too few neighbours to average over.

Structure (A grid moves the values, a cloud moves the nodes)**.** The two methods are not different in *what* they compute. Both need the law of (S_t, v_t) at each date, and both march it forward. They differ only in how they hold it, and chapter 3 already supplied both ways.

The forward equation approach carries the density itself and pushes it with the adjoint $\mathcal{L}^*$, at nodes the modeller fixes in advance. The particle method carries a *sample* and pushes each member with the process whose generator is $\mathcal{L}$; the empirical measure

$$\hat{p}_t = \frac{1}{N} \sum_{i=1}^N \delta_{(S_t^i,\, v_t^i)}$$

converges to the same density the grid was computing. So the particles are not an approximation to something other than the density. They *are* the density, sampled instead of tabulated: a grid

fixes where the nodes are and lets the values on them move, while a cloud fixes the values — every particle is worth $1/N$, always — and lets the nodes move.

That is the whole of the difference, and it is what decides which one survives a third factor. A grid has to cover wherever the process might go, so its cost grows exponentially in the number of factors, and most of that cost is spent on regions the process visits rarely. A cloud goes where the process actually goes, so it never spends anything on a region the process ignores, and its error stays $O(N^{-1/2})$ whatever the dimension. Trading accuracy per node for independence from dimension is the bargain every Monte Carlo method strikes; here it is the difference between a model with two factors and a model with more.

Notice that the particles interact. Each one's next step depends on where all the others currently are, through the conditional expectation — which is exactly the mean-field structure of the equation being solved, appearing in the numerical scheme as a set of particles that cannot be simulated independently. That is the price of the method and also the reason it is faithful to the problem.

Two practical points decide whether it works. The bandwidth h trades bias against noise in the usual way, and it must widen in the tails where particles are sparse. And in regions the process rarely visits, the denominator of (11.2) is estimated from a handful of particles, so there is noise; production implementations cap and floor it, and accept that the model's wings are the least trustworthy part of it — which is unfortunate, since the wings are what the exotics are usually sensitive to.

11.4 The Mixing Weight

Since the fit is exact whatever the stochastic part does, desks make the freedom explicit and give it a dial.

$$\begin{aligned} \frac{dS_t}{S_t} &= (r-q)\,dt + L_\lambda(t,S_t)\,\sqrt{v_t^\lambda}\,dW_t, \\ dv_t^\lambda &= \alpha(t,v_t^\lambda)\,dt + \lambda\,\beta(t,v_t^\lambda)\,dZ_t. \end{aligned} \tag{11.4}$$

Two things about (11.4). The parameter does not appear in the equation for S at all: it scales β, the diffusion of the *variance*, and nothing else. And L_λ carries a subscript because it is not chosen — it is whatever Theorem 11.2 returns once λ is fixed,

$$L_\lambda(t,K) = \frac{\sigma_{\text{loc}}(t,K)}{\sqrt{\mathbb{E}\big[\,v_t^\lambda \,\big|\, S_t = K\,\big]}}. \tag{11.5}$$

So λ and L are not two knobs. Turning the first determines the second, and the reason the fit survives is that the calibration condition rearranges to

$$L_\lambda^2(t,K)\,\mathbb{E}\big[\,v_t^\lambda \,\big|\, S_t = K\,\big] \;=\; \sigma_{\text{loc}}^2(t,K) \qquad \text{for every } \lambda, \tag{11.6}$$

whose right-hand side is the market's and contains no λ. That is the precise sense in which the weight tunes a balance: the product is pinned, and λ decides how it is divided between a factor that is random and a factor that is not.[1]

The two ends are then immediate. At $\lambda = 0$ the variance is deterministic, so the conditioning in (11.5) does nothing, $\mathbb{E}[v_t^0 \mid S_t = K] = v_t^0$, and $L_0^2 = \sigma_{\text{loc}}^2/v^0$ leaves the price equation reading $dS/S = \sigma_{\text{loc}}(t, S_t)\, dW_t$ exactly. At $\lambda = 1$ the stochastic part supplies as much of the smile as it can and L_1 is left correcting the remainder.

Calculation 11.3 (What the dial actually moves)**.** In the coin-toss model of *Where the conditioning bites* above, the vol-of-vol is the gap between the two volatilities, so λ closes that gap towards their common mean while holding the mean variance fixed — which isolates the randomness of the variance from its level, so that what moves is the split rather than the total.

Take a flat local volatility surface at 25%, so that any structure in L is the stochastic part's doing and not the market's. At $\lambda = 0$ the leverage function is flat at 0.93: with a deterministic variance there is nothing strike-dependent for it to do. At $\lambda = 1$ it runs from 1.09 at the money down to 0.71 in both wings.[2]

So the dial does not scale L; it changes its shape, from a constant to a hump with a range half its own height. That shape is the smile the stochastic part now generates by itself, being taken back out again — which is the division of labour of this chapter, made visible as a single function.

Recalibrating at each setting always succeeds, by Theorem 11.2.

- At $\lambda = 0$ the variance is deterministic, the conditional expectation in (11.2) is just its value, and L absorbs the whole surface. The model *is* local volatility: exact fit, backbone twice the skew, no vega.
- At $\lambda = 1$ the stochastic part produces most of the smile on its own and L is a modest correction near one. The dynamics are the stochastic model's.
- In between, the two are blended in a proportion the desk chooses.

Every setting of λ prices every vanilla option identically — to the penny, by construction. And the prices of the exotics move steadily across the range, often by several percent of premium between the ends. So λ is a parameter that changes the answer, does not change the calibration, and therefore cannot be determined by it.

This is the third time in three chapters that we have met the same shape: β in chapter 10, the choice of stochastic part here, and λ now. The right response is not to search harder for a way to pin it down from vanillas — Gyongi's theorem says there is none — but to treat it as what it is. Desks mark λ by policy, revalue the book at both ends of the range, and carry the difference as a reserve. A model risk that has been measured and provisioned for is a different thing from one that has been assumed away.

[1] `checked at every setting`.

[2] `measured`.

Remark (The forward smile, concretely)**.** The clearest place the difference shows is the forward smile.

Consider an option that will be struck at the money in one year and expire a year later — a forward starting option, the building block of a cliquet. Its value depends entirely on what the smile looks like in a year.

Under local volatility, conditional on the underlying reaching some level in a year, all remaining randomness is Brownian, and the smile from that point on is close to flat. Under a model with genuine vol-of-vol, the variance in a year is itself uncertain, and that uncertainty produces a smile then just as it does now. The prices differ by a lot, and the observable forward smiles in liquid markets are much closer to the second. Since L hides the difference in the vanillas, the forward starting option is one of the few instruments that could tell the two apart — which is exactly why it is used to calibrate λ where it trades.

11.5 What Has Been Bought

Close the loop back to the complaint that started the chapter.

The model reproduces every European option price exactly, by construction, through L. It reproduces them for *any* choice of the stochastic part, so that choice is left entirely free.

That freedom is then spent on the things chapter 10 showed vanilla options cannot see. The correlation ρ and the vol-of-vol set the backbone, and therefore the delta. The mean reversion of the variance sets the maturity term structure — how fast the skew decays as the expiry lengthens — and, when today's variance sits away from its long-run level, how quickly the forward smile settles to its stationary shape. Both bear on the price of anything forward starting, though chapter 10's measurement is the thing to keep in mind here: in a model with no calendar in it the forward smile does not flatten as the start date recedes, it settles. And because volatility is now a genuine risk factor, the model is incomplete in the sense of chapter 4 — there is a vega, it is real, and it can be hedged with other options rather than being an artefact.

This is the standard equity and foreign exchange exotics model, and the pattern is more general than the setting. When a model must match a set of prices and also behave sensibly, the productive move is often to split it into a part that is forced by the fit and a part that is free, and then to be explicit that the free part is a choice that the calibration did not make for you.

Chapter 12 applies the same idea to interest rates, where the state is a curve rather than a number, and where the analogue of the leverage function is a local volatility function of the model's own state variables.

References

- Guyon, J., & Henry-Labordere, P. (2012). Being particular about calibration. *Risk*, 25(1), 88–93.

- Ren, Y., Madan, D., & Qian, M. Q. (2007). Calibrating and pricing with embedded local volatility models. *Risk*, 20(9), 138–143.

- Lipton, A. (2002). The vol smile problem. *Risk*, 15(2), 61–65.

Chapter 12

Markovian Term-Structure Models

We return to interest rates carrying what the volatility chapters taught us. The Heath-Jarrow-Morton framework of chapter 8 is completely general and, for that reason, computationally useless: its state is the whole curve. We derive the one restriction on the volatility that collapses it to a finite state, obtain the quasi-Gaussian models, and take a mathematical excursion into why models of this shape can be solved at all.

12.1 Why HJM Is Not Enough

Chapter 8 derived the Heath-Jarrow-Morton drift condition: under the risk neutral measure, an arbitrage-free evolution of the instantaneous forward curve must take the form

$$df(t,T) = \sigma(t,T)\left(\int_t^T \sigma(t,s)\,ds\right)dt + \sigma(t,T)\,dW_t. \tag{12.1}$$

This is a complete answer to the question chapter 8 asked. Choose any volatility structure whatsoever and the drift is determined; there is no further freedom and no further constraint. As a piece of theory it could hardly be better.

As a thing to compute with it is close to unusable. The state of the model at time t is the entire function $T \mapsto f(t,T)$. Nothing smaller will do: to take the next step, (12.1) needs $\sigma(t,\cdot)$ integrated over the whole remaining curve, and in general σ may depend on the curve's current shape. So the process is Markov in an infinite dimensional state and in nothing less.

For pricing a European payoff this is survivable, though not because simulation somehow escapes the dimension. It does not. Discretise the curve onto a few hundred maturities and every one of them is a state variable that has to be stored and stepped, because (12.1) needs $\sigma(t,\cdot)$ integrated over the whole remaining curve before it can take a step, and if σ depends on the curve's shape then the previous step's answer is what supplies it. A simulated HJM path is a simulated *surface*.

What makes that bearable is the shape of the cost rather than its size. Carrying three hundred tenors costs three hundred times a one-factor model per step, which is linear in the dimension; and

chapter 19's measurement is that a Monte Carlo error is set by the payoff's variance and the path count and does not grow with the number of state variables at all. So a large state multiplies the cost of each path and does not multiply the number of paths needed. A European payoff also never asks what anything is worth at an intermediate date: one sweep from today to expiry, average at the end, and the state's dimension is bookkeeping.

Remark (When even the stepping is unnecessary)**.** If $\sigma(t,T)$ is deterministic, (12.1) can be integrated in closed form:

$$f(T,T') = f(0,T') + \int_0^T \sigma(u,T') \left(\int_u^{T'} \sigma(u,s)\, ds \right) du + \int_0^T \sigma(u,T')\, dW_u,$$

in which the drift term is a number and the stochastic integral is Gaussian with a deterministic variance. The whole curve at T is then a jointly Gaussian vector that can be drawn in one shot, with no time stepping at all. It is the dependence of σ on the curve that forces the step-by-step simulation, and with it the need to carry the surface.

Backward induction is where the dimension stops being bookkeeping. A Bermudan swaption must be valued by comparing, at each exercise date, the value of exercising against the value of continuing, and that comparison has to be made state by state — so it needs the value *at* a state rather than an average over states. And regression needs a small set of variables to regress against, which is exactly what an infinite dimensional state does not offer, and nothing in the model tells us which handful of curve summaries would do.

So the question of this chapter is: what must we give up for the curve to be a function of a handful of numbers?

12.2 What a Finite State Requires

Two separate things have to be true before the curve collapses onto a handful of numbers. The first is that the model be Markov in finitely many variables: there is a process $X_t \in \mathbb{R}^n$ that is Markov in the sense of chapter 3's definition of the Markov property, such that the whole curve at each date is a function of the date and X_t, with today's curve as data. The second is that its volatility depend on how long a rate has left to run rather than on the date it matures. The exponential comes from the two together.

Take $\sigma(t,T)$ deterministic for this section. The state-dependent case is what the model is eventually for, and it is returned to at the end.

12.2.1 Markov Is Exactly Finite Rank

Integrating (12.1), the curve at time t is

$$f(t,T) \;=\; \underbrace{f(0,T) + \int_0^t \alpha(u,T)\, du}_{\text{deterministic, so free}} \;+\; \underbrace{\int_0^t \sigma(u,T)\, dW_u}_{\text{the part needing a state}} \tag{12.2}$$

for every maturity $T \geq t$. With σ deterministic the drift is deterministic too, so the whole question is the second term: a continuum of random variables, one per maturity, each an integral over the whole past.

Lemma 12.1 (Finite state, finite rank, and a sum of products are the same thing)**.** *Let σ be deterministic and square integrable. The following are equivalent.*

(i) For every t, the curve $f(t,\cdot)$ is recoverable from $f(0,\cdot)$ and n random variables.

(ii) The functions $u \mapsto \sigma(u,T)$, as T varies, span a space of dimension n.

(iii) σ is a sum of n products and no fewer,

$$\sigma(u,T) \;=\; \sum_{i=1}^{n} \phi_i(u)\,\psi_i(T). \tag{12.3}$$

In that case the state variables are $X_t^i = \int_0^t \phi_i(u)\,dW_u$, and

$$f(t,T) \;=\; f(0,T) + \int_0^t \alpha(u,T)\,du + \sum_{i=1}^{n} \psi_i(T)\,X_t^i. \tag{12.4}$$

Proof. (i) $\Leftrightarrow$ (ii) is the Itô isometry. The map $h \mapsto \int_0^t h(u)\,dW_u$ is linear, and

$$\mathbb{E}\left[\left(\int_0^t h(u)\,dW_u\right)^2\right] = \int_0^t h^2(u)\,du,$$

so it is an isometry and in particular injective. A family of such integrals therefore spans a space of exactly the dimension spanned by the family of integrands. Recovering (12.2) for every T from n numbers is thus precisely the statement that the integrands span n dimensions.

(ii) $\Leftrightarrow$ (iii) is linear algebra with no probability in it. If the sections span n dimensions, choose a basis $\phi_1, \ldots, \phi_n$ of that span. Each section is in the span, so

$$\sigma(\,\cdot\,,T) = \sum_{i=1}^{n} \psi_i(T)\,\phi_i,$$

where the coefficients depend on which section we took, that is on T alone; and that is (12.3). Conversely (12.3) exhibits every section as a combination of the same n functions. Substituting into (12.2) and pulling $\psi_i(T)$ out of the integral, since it does not involve u, gives (12.4). □

Nothing was assumed about the shape of σ to get (12.3). The product structure is not an ansatz; it is what having finite rank *means*, and lemma 12.1 says that is exactly what a finite state is.

12.2.2 Finite Rank Alone Forces Nothing

It is tempting to stop here and conclude that a Markov model needs an exponential decay. It does not.

Example 12.1 (A one-state model with an arbitrary maturity shape)**.** Let $\sigma(t,T) = \psi(T)$ for any square-integrable ψ whatsoever — the volatility of the forward rate maturing at T is $\psi(T)$, and does not change as t advances. This is rank one, with $\phi_1 \equiv 1$, so lemma 12.1 gives a single state,

$$X_t = \int_0^t dW_u = W_t, \qquad f(t,T) = f(0,T) + \int_0^t \alpha(u,T)\,du + \psi(T)\,W_t.$$

The drift condition determines α from ψ as always, so the model is arbitrage free; it is Markov in one variable; and ψ is unrestricted. Take $\psi(T) = 1/(1+T)$ if a concrete one is wanted. Measuring the rank of the kernel confirms it is one, for that ψ and for anything else put in its place.[1]

So a Markov model can carry any maturity shape at all. What is wrong with example 12.1 is not its mathematics but its economics: the volatility attached to the rate maturing in 2055 is the same number in 2054 as it is today. A thirty year forward rate and a rate one day from fixing are given whatever volatility their maturity *dates* happen to have been assigned, with no relation between them. Real forward rates do not behave that way — what governs a rate's volatility is how long it has left to run.

Structure. This is chapter 10's complaint arriving in the rates market. There the trouble was a local volatility surface whose skew decayed along the calendar, so a forward starting option saw a flatter smile than a spot one; here it is a volatility whose shape is pinned to dates rather than to tenors, so a forward curve does not look like a spot curve. Both are the same failure: a parameter carrying calendar time where the market carries only elapsed time.

12.2.3 The Second Requirement

So we impose it, as a modelling demand and not as a theorem.

Definition 12.2 (Time-homogeneous volatility)**.** The forward rate volatility is time homogeneous if its dependence on maturity is a fixed shape in the time left to it, with calendar time entering only as an overall level,

$$\sigma(t,T) \;=\; \sigma_r(t,\omega)\,g(T-t), \tag{12.5}$$

for a *maturity profile* g and a level σ_r carrying whatever randomness there is.

The product is the content. To say that $\sigma(t,T)$ depends on the maturity "only through $T-t$" would say nothing, since every function of (t,T) is a function of $(t,T-t)$; what is demanded is that the *shape* in tenor not depend on the date. Example 12.1 fails the demand and (12.5) is the simplest way to meet it, which is all this chapter's model needs. Several levels, each riding on its own shape, are taken up in §12.2.7.

[1] the test that pins this.

This is chapter 10's time homogeneity with the level allowed to move. There a model was time homogeneous when its coefficient did not depend on the date t at all. Here the shape does not, and in the model this chapter builds the level depends on the date only through the state, which is the same demand. A deterministic $\sigma_r(t)$, which the rank arguments below use because it reduces the question to linear algebra, is a relaxation: chapter 10 would call it inhomogeneous, and it is why the shape and not the whole volatility is what has to be calendar free.

Now put the two requirements together. Take σ_r deterministic and non-vanishing. The sections of (12.5) in u, for fixed T, are $u \mapsto \sigma_r(u)\, g(T-u)$ on $[0,t]$. The factor σ_r is the same function for every T, so multiplying by it is one invertible linear map applied to all the sections at once, and it cannot change the dimension of their span. That leaves the sections $u \mapsto g(T-u)$.

Index these by how far ahead the maturity lies, $a = T - t \geq 0$, and write $b = t - u \in [0,t]$ for how long ago the noise being summed was added. Then

$$T - u = \underbrace{(T-t)}_{a} + \underbrace{(t-u)}_{b}, \tag{12.6}$$

and the section is $b \mapsto g(a+b)$: a function of the calendar variable, indexed by the maturity, which is the family lemma 12.1 counts. Collect them in one function of two variables,

$$G(a,b) \;=\; g(a+b),$$

which depends on its arguments only through their sum, and write $S_a g = G(a,\cdot)$ for the profile shifted by a. Since $G(a,b) = G(b,a)$, the other family, $a \mapsto g(a+b)$ indexed by b, is the same set of shifts, so nothing depends on which way round the sections are taken. This is not a trick to make the algebra work. It is what time homogeneity does to the kernel: the sections of a kernel that depends only on a difference *are* the shifts of one function, and the substitution merely names them.

So lemma 12.1 becomes, under definition 13.10: the model is Markov in n states if and only if the shifts of g span a space of dimension n,

$$V = \operatorname{span}\{\, S_a g : a \geq 0 \,\}, \qquad \dim V = n,$$

which is to say that G has rank n as a kernel in (a,b). Expanding in a basis $g_1, \ldots, g_n$ of V gives the form (12.3) again, now with the shift structure visible:

$$G(a,b) \;=\; g(a+b) \;=\; \sum_{i=1}^{n} c_i(a)\, g_i(b), \qquad X_t^i = \int_0^t \sigma_r(u)\, g_i(t-u)\, dW_u. \tag{12.7}$$

Two things about (12.7) are easy to misread. Setting $b = 0$ gives $g(a) = \sum_i c_i(a)\, g_i(0)$, so the sections evaluated at one point recover g; but that is a single linear functional, and says nothing about how many functions of b are needed. The dimension belongs to the family $\{S_a g\}$ taken whole. And setting $a = 0$ gives $g = \sum_i c_i(0)\, g_i$: the profile lies in V, and the $c_i(0)$ are its coordinates in the basis.

Structure (The shifts are a representation of the maturity axis)**.** The shifts compose, $S_a S_{a'} = S_{a+a'}$, so they are a representation of the additive semigroup $([0,\infty),+)$ on functions of b, and

the sections $S_a g$ are the orbit of g. V is the smallest shift-invariant space containing g, so finite rank says that g generates a finite dimensional representation. Every continuous representation of $([0,\infty),+)$ on a finite dimensional space is $S_a = e^{aA}$ for a matrix A, and then $g(x) = (S_x g)(0)$ is a matrix coefficient of e^{xA}.

That is why exponentials appear and nothing else does. The one dimensional representations are the characters $e^{-\kappa a}$; the larger ones are the Jordan blocks of A, whose matrix coefficients are polynomials times exponentials. A kernel that depends only on the sum of its arguments is a Hankel kernel, and this is the continuous counterpart of Kronecker's theorem that a Hankel kernel has finite rank exactly when its symbol is rational. The next two subsections do it by hand, for rank one and then for rank n.

12.2.4 Rank One Is the Exponential

Proposition 12.3. *Under definition 13.10, a model Markov in one state per factor has $g(x) = e^{-\kappa x}$, and no other profile will do.*

Proof. $\dim V = 1$ means V is spanned by g itself, so (12.7) reads $g(a+b) = c(a)\,g(b)$. Setting $b = 0$ gives $c(a) = g(a)/g(0)$, and normalising $g(0) = 1$ — the constant can be absorbed into σ_r — leaves

$$g(a+b) = g(a)\,g(b). \tag{12.8}$$

This is Cauchy's exponential equation, whose measurable solutions are $g(x) = e^{-\kappa x}$ and nothing else.[2] □

That is the model this chapter uses.

Definition 12.4 (Separable volatility)**.** The forward rate volatility is separable if it factorises as

$$\sigma(t,T) = \sigma_r(t,\omega)\, e^{-\int_t^T \kappa(u)\,du}, \tag{12.9}$$

with the first factor allowed to depend on time and on the state of the world, and the second a purely deterministic decay in the maturity.

Take κ constant for the rest of this chapter, so the decay is $e^{-\kappa(T-t)}$. Proposition 12.3 is what earns it: given time homogeneity, the exponential is not one convenient choice among many but the only one a single state permits.

Remark (Whether a varying mean reversion is allowed)**.** (12.9) writes $e^{-\int_t^T \kappa(u)du}$ and proposition 12.3 concludes $e^{-\kappa x}$. Let's see how these two sit together.

Write $G(s) = e^{-\int_0^s \kappa(u)du}$, so that the decay in (12.9) is $G(T)/G(t)$ and

$$\sigma(t,T) = \frac{\sigma_r(t,\omega)}{G(t)} \cdot G(T).$$

[2] Measurability is not a formality one can drop: without it the equation has pathological solutions built from a Hamel basis. It is not a real restriction on a volatility.

That is a function of t times a function of T, whatever κ does — an outer product, of rank one as a kernel in (t, T).[3] So a time-dependent mean reversion is permitted and costs nothing: the state is still (x, y) and everything derived below goes through with $e^{-\kappa(T-t)}$ read as $G(T)/G(t)$.

What it costs is time homogeneity, which is the proposition's hypothesis rather than its conclusion. The maturity profile at date t is $x \mapsto G(t+x)/G(t)$, and that depends on t as well as on x: the shape a rate's volatility follows as it ages is a different shape depending on when one starts watching. Read as a fixed profile in time to maturity it is not an exponential at all, and its shifts span a space of dimension well above one.[4]

Which makes this example 12.1 again, in the notation the chapter actually uses. One state, because the volatility is indexed by the maturity date; no fixed profile, because it is not indexed by time remaining. The economic objection carries across unchanged, and so does the practical reason desks do it anyway: a varying κ is one of the cheapest ways to fit a term structure of volatility exactly, and what is bought with the fit is paid for in forward volatility, which is then whatever the calibration happened to leave behind.

12.2.5 Rank n Is a Differential Equation

Theorem 12.5 (Quasi-exponentials, and nothing else)**.** *Suppose g is smooth and $\dim V = n$ is finite. Then g satisfies a linear differential equation of order n with constant coefficients,*

$$g^{(n)} + \alpha_{n-1} g^{(n-1)} + \cdots + \alpha_1 g' + \alpha_0 g = 0, \tag{12.10}$$

and consequently

$$g(x) = \sum_j p_j(x)\, e^{\lambda_j x}, \tag{12.11}$$

a finite sum of polynomials times exponentials, with the λ_j the roots of the characteristic polynomial of (12.10) *and $\deg p_j$ one less than the multiplicity of λ_j. Complex roots appear in conjugate pairs and contribute $e^{\mu x}\cos(\omega x)$ and $e^{\mu x}\sin(\omega x)$. Conversely every such g has $\dim V$ equal to the order of the equation it satisfies.*

Proof. V is closed under shifting: a typical element is $h(x) = \sum_k \lambda_k\, g(x + a_k)$, and $h(x+s) = \sum_k \lambda_k\, g(x + a_k + s)$ is again a combination of shifts of g, so again in V.

That makes V closed under differentiation. For $h \in V$,

$$h'(b) = \lim_{s \to 0} \frac{h(b+s) - h(b)}{s},$$

and every difference quotient on the right lies in V, because both $x \mapsto h(x+s)$ and h do and V is a linear space. A finite dimensional subspace is closed, so the limit lies in V too.

So $D = d/dx$ is a linear operator on an n dimensional space. By Cayley-Hamilton it satisfies its own characteristic polynomial p, of degree n, on all of V; and $g \in V$. That is (12.10), with $p(\lambda) = \lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_0$, and (12.11) is its classical solution. For the converse, the shifts of a solution of (12.10) are again solutions, so V sits inside a solution space of dimension n.[5] □

[3] `computed`.

[4] `computed`.

[5] Smoothness is assumed here. It is not needed: lemma 12.6 in §12.2.7 derives it from continuity alone, since the shifts of a continuous g spanning a finite dimensional space already forces g to be a quasi-exponential.

Read (12.10) back as a statement about the model: the *order of the differential equation the maturity profile obeys is the number of state variables*. Nothing else about g matters. Proposition 12.3 is the case $n = 1$, where (12.10) is $g' + \kappa g = 0$.

12.2.6 What Rank Two and Three Look Like

The cheapest way to see the general case is to expand $g(a + b)$ by hand. For $g(x) = x^k e^{-\kappa x}$ the binomial theorem gives

$$\underbrace{(a+b)^k e^{-\kappa(a+b)}}_{g(a+b)} \;=\; \sum_{j=0}^{k} \underbrace{\binom{k}{j} a^{k-j} e^{-\kappa a}}_{c_j(a)} \; \underbrace{b^j e^{-\kappa b}}_{g_j(b)}, \tag{12.12}$$

which is (12.7) exactly, with basis $g_j(b) = b^j e^{-\kappa b}$ for $j = 0, \dots, k$. So $x^k e^{-\kappa x}$ needs $k + 1$ states, and (12.10) agrees: $(D + \kappa)^{k+1} g = 0$, a root of multiplicity $k + 1$.

The case $k = 1$ is the one that matters in practice, since a volatility peaking at some tenor rather than decaying from the front is closer to what is quoted. Written out, (12.12) is

$$g(a+b) \;=\; \underbrace{a e^{-\kappa a}}_{c_0(a)} \cdot \underbrace{e^{-\kappa b}}_{g_0(b)} \;+\; \underbrace{e^{-\kappa a}}_{c_1(a)} \cdot \underbrace{b e^{-\kappa b}}_{g_1(b)},$$

so the hump is Markov in the two states of (12.7) built from $g_0(b) = e^{-\kappa b}$ and $g_1(b) = b e^{-\kappa b}$. The first is Hull-White's own state; the hump keeps it and adds one. This is an identity per increment rather than in a limit, so it holds pathwise and exactly.[6]

Everything else follows straight from the characteristic polynomial:

maturity profile $g(x)$	**its equation**	**states**
$e^{-\kappa x}$	$(D + \kappa)\, g = 0$	1
$x\, e^{-\kappa x}$	$(D + \kappa)^2 g = 0$	2
$x^2 e^{-\kappa x}$	$(D + \kappa)^3 g = 0$	3
$e^{-\kappa_1 x} + e^{-\kappa_2 x}$	$(D + \kappa_1)(D + \kappa_2)\, g = 0$	2
$e^{-\kappa x} \cos(\omega x)$	$\big(D^2 + 2\kappa D + \kappa^2 + \omega^2\big) g = 0$	2
$1/(1 + x)$	none	no finite number
$e^{-\kappa\sqrt{x}}$	none	no finite number

Computing the rank of $(a, b) \mapsto g(a + b)$ numerically returns exactly this column,[7].

The last two rows deserve a word, since a theorem excluding nothing would hardly need proving. Neither function is pathological: both are smooth, positive and decaying, and either would be a reasonable thing to fit to a volatility term structure.

[6] `the pathwise test`.

[7] `quasigaussian::realisation_dimension`.

Example 12.2 (A profile with no finite realisation)**.** Take $g(x) = 1/(1+x)$ and suppose some finite combination of its shifts vanishes,

$$\sum_{j=1}^{m} \lambda_j \frac{1}{1+a_j+b} = 0 \quad \text{for all } b \geq 0,$$

with the a_j distinct. The left side is a rational function of b with simple poles at $b = -(1+a_j)$, all distinct. Multiplying by $(1+a_1+b)$ and letting $b \to -(1+a_1)$ kills every term but the first and leaves $\lambda_1 = 0$; repeating kills them all. So no finite set of shifts is dependent, $\dim V$ is infinite, and there is no Markov realisation in any number of states.

Set that beside example 12.1, which used the same $1/(1+T)$ and got a model with *one* state. Nothing about the function changed. What changed is whether it was read as a shape in calendar time or a shape in time to maturity, and that alone moves the model from one state to infinitely many. It is as sharp a statement as this section has of where the exponential actually comes from: not from Markovianity, which tolerates any shape, but from insisting that the shape ride along with the maturity.

So being Markov in finitely many variables does not force separability. It forces *finite rank*, by lemma 12.1; time homogeneity then turns finite rank into a differential equation, by theorem 12.5; and separability is the case $n = 1$ of the two together. Multi-factor Cheyette models are the case $n > 1$, at n times the state and with nothing else changed.

Structure. This is chapter 14's subject arriving early. There the question is which models are solvable, and the answer is that a family of functions preserved by the dynamics turns a partial differential equation into a few ordinary ones. Here the dynamics is an infinite dimensional one on the curve, the family is V, and preservation is closure under shifting; the payoff is the same, an infinite dimensional object collapsing onto finitely many coordinates. Tractability keeps turning out to be an invariance, and the invariant object keeps turning out to be a small linear space — here small enough that its dimension is literally the number of state variables.

12.2.7 Several Levels

Definition 13.10 asks for one level riding on one shape. The demand it was standing in for is weaker: that the tenor profiles at every date come from a fixed, finite dimensional family of shapes, so that calendar time changes only how much of each shape is present,

$$\sigma(u, u+x) \;=\; \sum_{j=1}^{m} a_j(u)\, h_j(x). \tag{12.13}$$

A single Brownian motion with $\sigma(t,T) = a_1(t)\, e^{-\kappa_1(T-t)} + a_2(t)\, e^{-\kappa_2(T-t)}$ is an example. Its shape moves with t as the two levels move against each other, so it is not of the form (12.5), and the argument of the previous subsections does not apply to it: the levels multiply different shapes inside one section, so they no longer factor out, and the sections are not the orbit of one function. (With a separate Brownian motion per term the argument does go through, verbatim, with vector valued shifts.) Is the exponential still forced?

The shifts of one function are not available, but the invariance behind them is. A kernel that depends only on $T-u$ is unchanged by moving both arguments together, so $\sigma(u+v,T+v)$ carries the same shapes $h_j(T-u)$ as $\sigma(u,T)$, with different levels. That is enough. First a lemma, which is also what removes the smoothness assumption from theorem 12.5.

Lemma 12.6 (A finite dimensional span of translates is a quasi-exponential)**.** *Let f be continuous on $[0,\infty)$, let $0<\varepsilon\le t$, and suppose the functions $x\mapsto f(x-u)$ on $[t,\infty)$, for $u\in[0,\varepsilon]$, span a finite dimensional space W. Then f restricted to $[t,\infty)$ lies in W and is a quasi-exponential.*

Proof. W consists of smooth functions. For ρ smooth and supported in $(0,\varepsilon)$, the function $\rho*f=\int\rho(u)\,f(\cdot-u)\,du$ is smooth, and lies in W because a finite dimensional subspace is closed. Suppose a linear functional Λ on W annihilates every such $\rho*f$. Then $\lambda(u)=\Lambda\big(f(\cdot-u)\big)$ is continuous with $\int\rho\lambda=0$ for every ρ, so $\lambda\equiv0$ on $[0,\varepsilon]$, and Λ vanishes on the translates that span W, hence on W. So the functions $\rho*f$ span W, and every element of W is smooth.

W is closed under differentiation. Choose $u_1,\dots,u_d\in(0,\varepsilon)$ such that $e_k=f(\cdot-u_k)$ is a basis of W, and write $f(x-u)=\sum_k c_k(u)\,e_k(x)$ for $u\in[0,\varepsilon]$ and $x\ge t$. The c_k solve a linear system whose right hand side is a smooth function of u, so they are smooth, and $c_k(u_l)=\delta_{kl}$. Differentiating in u and in x and adding the results gives $\sum_k\big[c_k'(u)\,e_k(x)+c_k(u)\,e_k'(x)\big]=0$, and at $u=u_l$ this reads $e_l'=-\sum_k c_k'(u_l)\,e_k$.

So d/dx maps W into itself by some matrix M, $e'=Me$, and $e(x)=e^{M(x-t)}e(t)$ on $[t,\infty)$: every e_k is a quasi-exponential. Taking $u=0$, $f=\sum_k c_k(0)\,e_k$ lies in W, so it is one too. □

Theorem 12.7 (Several levels, and nothing else)**.** *Let σ be deterministic and continuous, and suppose*

(i) for every t it has finite rank as a kernel on $[0,t]\times[t,\infty)$, so that a finite state suffices, by lemma 12.1;

(ii) it has the form (12.13) *with continuous levels a_j that are linearly independent on every interval $[0,t]$.*

Then every h_j is a quasi-exponential. Conversely, if the h_j lie in a finite dimensional space of quasi-exponentials, (i) holds.

Proof. The independence in (ii) loses nothing: dependent levels can be recombined into fewer terms. It makes the h_j continuous, being combinations of sections of σ.

Fix $t>0$. By (i) there are functions with $\sigma(u,T)=\sum_{i\le n}\phi_i(u)\,\psi_i(T)$ for $u\in[0,t]$ and $T\ge t$. Independent functions can be interpolated, so there are nodes $v_1,\dots,v_m\in(0,t)$ with $\det\big(a_j(v_r)\big)\neq0$, and by continuity there is an $\varepsilon\in(0,t]$ small enough that $u+v_r<t$ and $\det\big(a_j(u+v_r)\big)\neq0$ for all $u\in[0,\varepsilon]$. Now evaluate at the point $(u+v_r,\,T+v_r)$, moving both arguments by v_r. By (ii) the shapes see only their difference, and by (i) the kernel factorises:

$$\sum_j a_j(u+v_r)\,h_j(T-u)\;=\;\sum_i\phi_i(u+v_r)\,\psi_i(T+v_r),\qquad r=1,\dots,m,$$

for $u\in[0,\varepsilon]$ and $T\ge t$. The matrix on the left is invertible, so each $h_j(T-u)$ is a combination of the functions $T\mapsto\psi_i(T+v_r)$, with coefficients depending on u. These span a space of dimension

at most nm, so the translates of h_j by $u \in [0, \varepsilon]$ span a finite dimensional space, and lemma 12.6 says h_j is a quasi-exponential on $[t, \infty)$. Since t was arbitrary, on $(0, \infty)$.

Conversely, if the h_j lie in a shift invariant space with basis $e_1, \dots, e_q$, then $h_j(T-u) = \sum_k c_{jk}(u)\, e_k(T)$, and (12.13) is a sum of q products in (u, T). □

So with deterministic levels, a finite state and a calendar-free finite family of shapes leave exactly

$$\sigma(t,T) \;=\; \sum_{i=1}^{q} a_i(t)\, e_i(T-t),$$

with $e_1, \dots, e_q$ a basis of a space of quasi-exponentials closed under shifting and arbitrary levels a_i. The single level of definition 13.10 is the case where every level is the same σ_r: expanding $g = \sum_i c_i(0)\, g_i$ in a basis of V gives $\sigma_r(t) \sum_i c_i(0)\, g_i(T-t)$, with the loadings $c_i(0)$ fixed and the levels locked together. Freeing them lets the shape move with the date, at no cost in the number of states.[8] Random levels are a different matter: sufficiency survives them, and necessity does not, as the next remark explains.

Remark (What survives when the level is stochastic)**.** Everything above assumed σ deterministic, and (12.9) explicitly permits a σ_r depending on the state — which is what makes the model useful, since it is where the smile comes from.

Sufficiency is untouched. Nothing in theorem 12.8 uses determinism: the integrals defining x, y and A are pathwise, and hold for any adapted σ_r. So a rank-n profile with an exponential shape and an arbitrary state-dependent level still gives a finite Markov state, which is the direction the chapter actually needs. The construction is safe.

Necessity is what the Itô isometry was buying, and that is where the argument breaks — though not because the isometry fails. It does not: $\mathbb{E}\big[(\int h\, dW)^2\big] = \mathbb{E}\big[\int h^2 du\big]$ holds for every adapted square integrable integrand, random ones included. What stops working is the use the proof makes of it.

For deterministic σ the family $\{u \mapsto \sigma(u,T)\}_T$ is a family of *fixed* vectors in $L^2[0,t]$, and the isometry embeds that space linearly and injectively into $L^2(\Omega)$. So the span of the random variables has exactly the dimension of the span of the functions, and both are spans over $\mathbb{R}$ — constant coefficients. That is the notion (12.4) needs, because there the curve is rebuilt by *deterministic* loadings $\psi_i(T)$ on n random variables. Linear algebra is the right tool because, with σ deterministic, the curve is Gaussian and everything in sight is linear.

Let σ_r depend on the state and neither remains true. The integrands are themselves random, so $\{\sigma(\cdot, T)\}_T$ is no longer a deterministic family in a fixed space; and a curve recoverable from n state variables is recoverable by a *measurable* function of them, which for a non-Gaussian curve has no reason to be linear. The span that matters is then over the state-measurable coefficients rather than over $\mathbb{R}$ — a module rather than a finite dimensional vector space — and a rank computed over $\mathbb{R}$ is not counting it. The isometry survives intact; what it no longer does is convert "n state variables" into "n linearly independent integrands". Lemma 12.1 loses its proof there.

The replacement is Lie-algebraic. Reparametrising the curve by time to maturity, $r_t(x) = f(t, t+x)$ — the Musiela form — turns (12.1) into a stochastic partial differential equation whose

[8] the finite rank test, and its converse.

drift contains the transport term $\partial/\partial x$, because holding a tenor fixed means sliding along the curve as time passes. The model has a finite dimensional realisation exactly when the Lie algebra generated by its drift and volatility vector fields is finite dimensional.

Applied to a *deterministic* volatility, that criterion reproduces theorem 12.5. Bracketing the volatility field against the transport term differentiates the maturity profile; bracketing again differentiates it again. Finite dimensionality of the algebra therefore says that the derivatives of g span a finite dimensional space — which is (12.10), and so still the quasi-exponentials. The elementary proof and the geometric one arrive at the same object, V closed under d/dx, by different roads.

Once σ_r depends on the curve, those brackets acquire further terms from the volatility field's own dependence on the state, and the criterion stops reducing to a condition on g alone.

- What the chapter uses is the *if* direction, and it holds regardless. An exponential profile with any adapted level whatever gives the two-state model of theorem 12.8.

- What is given up is the *only if*. For deterministic σ, lemma 12.1 is an equivalence, so we know that nothing outside the quasi-exponentials could have worked. With a stochastic level we no longer have that proof, and whether some cleverly state-dependent level could rescue a profile that is not quasi-exponential is a question nothing here settles. It has to be settled model by model, from the Lie algebra.

The deterministic characterisation is due to Bjork and Christensen, the state-dependent analysis to Bjork and Svensson, and the general geometry of invariant manifolds to Filipovic and Teichmann.

With that, we can find the state.

Theorem 12.8 (Cheyette, quasi-Gaussian state)**.** *Under* (12.9)*, define*

$$x_t = r_t - f(0,t), \qquad y_t = \int_0^t \sigma_r^2(u)\, e^{-2\kappa(t-u)}\, du.$$

Then (x, y) *is Markov, with*

$$dx_t = \big(y_t - \kappa x_t\big)\, dt + \sigma_r(t)\, dW_t, \qquad dy_t = \big(\sigma_r^2(t) - 2\kappa y_t\big)\, dt, \tag{12.14}$$

both starting from zero, and the entire curve is recovered from them by

$$P(t,T) = \frac{P(0,T)}{P(0,t)} \exp\Big(- G(t,T)\, x_t - \tfrac{1}{2} G^2(t,T)\, y_t \Big), \qquad G(t,T) = \frac{1 - e^{-\kappa(T-t)}}{\kappa}. \tag{12.15}$$

Proof. Integrate (12.1) from 0 to t with the separable volatility. The stochastic part is

$$\int_0^t \sigma_r(u) e^{-\kappa(T-u)}\, dW_u = e^{-\kappa(T-t)} \int_0^t \sigma_r(u) e^{-\kappa(t-u)}\, dW_u,$$

which is the factorisation the previous remark promised: the maturity has come outside, leaving a single history-dependent quantity. Call that quantity X_t.

For the drift, the inner integral in (12.1) is

$$\int_u^T \sigma_r(u)e^{-\kappa(s-u)}\,ds = \sigma_r(u)\,\frac{1-e^{-\kappa(T-u)}}{\kappa} = \sigma_r(u)\,G(u,T),$$

so the accumulated drift is $\int_0^t \sigma_r^2(u)e^{-\kappa(T-u)}G(u,T)\,du$. Since $\kappa G(u,T) = 1-e^{-\kappa(T-u)}$, the integrand is

$$\tfrac{1}{\kappa}\,\sigma_r^2(u)\Big[e^{-\kappa(T-u)} - e^{-2\kappa(T-u)}\Big],$$

and splitting both exponentials at t as before gives

$$\int_0^t \alpha(u,T)\,du \;=\; \frac{1}{\kappa}\Big[e^{-\kappa(T-t)}A_t \;-\; e^{-2\kappa(T-t)}y_t\Big], \qquad A_t = \int_0^t \sigma_r^2(u)\,e^{-\kappa(t-u)}\,du. \tag{12.16}$$

Note that *two* accumulators have appeared, A at rate κ and y at rate 2κ, so the state looks three dimensional at this point.

It is not. Setting $T=t$ in (12.16) and adding the stochastic part,

$$r_t = f(t,t) = f(0,t) + \frac{A_t - y_t}{\kappa} + X_t, \qquad \text{so} \qquad x_t = X_t + \frac{A_t-y_t}{\kappa}.$$

That is a relation among the three, and using it to eliminate A_t from (12.16),

$$\begin{aligned} f(t,T) - f(0,T) &= e^{-\kappa(T-t)}X_t + \frac{1}{\kappa}\Big[e^{-\kappa(T-t)}A_t - e^{-2\kappa(T-t)}y_t\Big] \\ &= e^{-\kappa(T-t)}\underbrace{\Big[X_t + \frac{A_t-y_t}{\kappa}\Big]}_{x_t} + \frac{y_t}{\kappa}\,e^{-\kappa(T-t)}\Big[1-e^{-\kappa(T-t)}\Big], \end{aligned}$$

in which A_t has cancelled. The bracket on the right is $\kappa\,G(t,T)$, so

$$f(t,T) = f(0,T) + e^{-\kappa(T-t)}\,x_t + e^{-\kappa(T-t)}G(t,T)\,y_t. \tag{12.17}$$

Two states, not three, because the short rate itself supplies the third relation.

Integrating (12.17) in T and exponentiating, using $P(t,T) = \exp(-\int_t^T f(t,s)ds)$ from chapter 7, gives (12.15): the $\int_t^T e^{-\kappa(s-t)}ds$ produces $G(t,T)$, and

$$\int_t^T e^{-\kappa(s-t)}G(t,s)\,ds = \frac{1}{\kappa}\int_0^{T-t}\big(e^{-\kappa v} - e^{-2\kappa v}\big)dv = \tfrac{1}{2}G^2(t,T)$$

produces the quadratic term, after cancelling against the initial curve.

Finally the dynamics. Differentiating the three pieces separately,

$$dX_t = -\kappa X_t\,dt + \sigma_r\,dW_t, \qquad dA_t = \big(\sigma_r^2 - \kappa A_t\big)dt, \qquad dy_t = \big(\sigma_r^2 - 2\kappa y_t\big)dt,$$

the last of which is already the second equation of (12.14). For the first, $d\big[(A_t-y_t)/\kappa\big] = (2y_t - A_t)\,dt$, so

$$dx_t = \big(-\kappa X_t + 2y_t - A_t\big)dt + \sigma_r\,dW_t = \Big(y_t - \kappa\big[X_t + (A_t-y_t)/\kappa\big]\Big)dt + \sigma_r\,dW_t,$$

and the bracket is x_t. Note that A has cancelled again: it is needed to write the curve down and not to step it forward. □

Three things about this deserve emphasis, because they are what make the model usable rather than merely finite.

First, y has no dW in it. It is not a risk factor and cannot be shocked; it is a running accumulator of variance, and its only job is to make the pair Markov. It is exactly the memory that (12.1)'s drift needs and that x alone cannot supply.

Second, (12.15) reconstructs *every* discount factor as an explicit function of two numbers. So every forward rate, every swap rate and every annuity is a closed-form function of (x, y): the state says exactly what the underlying is worth, which is the exercise value at any date. It does not say what an option is worth. That is an expectation over where (x, y) goes next, and with a state-dependent level even its law is not known in closed form; for a Bermudan the value is the larger of exercising and the expected continuation, a maximum inside an expectation, which has no closed form even for Gaussian dynamics. What the reconstruction buys is that the exercise value is free at every state and, because (x, y) is Markov, that the continuation value at each exercise date is a function of the date and two variables. That is what makes backward induction possible: a two-dimensional grid, or a regression on (x, y) in a Monte Carlo, and the exercise decision can be made state by state.

A payoff that depends on the curve at one date is a function of the state at that date, and by the Markov property its value at any earlier date is a function of (t, x, y) alone. Path dependence is exactly the failure of this: the value at an intermediate date depends on the past beyond (x, y), through an accrual count, a running maximum or a barrier flag. The remedy is to carry that summary as an extra state variable, as in chapter 3's remark Markov is a statement about a choice of state. It works when the summary updates from its previous value and the current state, and a payoff with no finite summary has no finite state. The same holds for a level with randomness of its own, such as the independent variance factor z added at the end of the chapter, which makes the state (x, y, z).

Third, today's curve enters only through the ratio $P(0,T)/P(0,t)$, which is read straight off chapter 7's bootstrap. The initial curve is matched exactly and automatically, with nothing to calibrate.

Example 12.3 (Climbing down the ladder to Hull-White)**.** The quickest way to believe Theorem 12.8 is to take the one case where we already know the answer.

Set $\sigma_r(t) = \sigma$, a constant. The equation for y in (12.14) has no randomness in it at all, so it is an ordinary linear differential equation, $\dot{y} = \sigma^2 - 2\kappa y$ with $y_0 = 0$. Solving,

$$y_t = \frac{\sigma^2}{2\kappa}\Big(1 - e^{-2\kappa t}\Big),$$

which is a known function of time and not a state variable at all. The model has collapsed to one dimension.

What is left is

$$dx_t = \big(y_t - \kappa x_t\big)dt + \sigma\, dW_t,$$

a Gaussian Ornstein-Uhlenbeck process with a time-dependent drift, and since $r_t = f(0,t) + x_t$, the short rate is

$$dr_t = \big(\theta(t) - \kappa r_t\big)dt + \sigma\, dW_t, \qquad \theta(t) = \frac{\partial f(0,t)}{\partial t} + \kappa f(0,t) + y_t.$$

This is exactly the Hull-White short rate equation from chapter 8's Hull-White Model section, with θ assembled from the initial curve and the model's own accumulated variance — which is what that section found there too, by a completely different route.

And (12.15) becomes

$$P(t,T) = \frac{P(0,T)}{P(0,t)} \exp\Big(-G(t,T)\big(r_t - f(0,t)\big) - \tfrac{1}{2}G^2(t,T)\, y_t \Big),$$

an exponential of something affine in r_t — the affine term structure chapter 8's Hull-White Model section observed and did not explain. The next section explains it.

Remark (The ladder)**.** The example is the bottom rung of a ladder of restrictions:

HJM (general, infinite state) $\longrightarrow$ impose exponential separability $\longrightarrow$ quasi-Gaussian (Markov in (x, y)) $\longrightarrow$ make σ_r deterministic $\longrightarrow$ Hull-White.

Each arrow is a restriction bought for a computational gain. Separability buys a finite state. A deterministic level buys a Gaussian one, and with it closed form prices, at the cost of the smile, which lives in σ_r's dependence on the state. Knowing which rung you are standing on is most of knowing what your model can and cannot do.

12.3 Excursion: Affine Models and Why They Solve

Formula (12.15) has a particular shape — an exponential of something linear in the state, plus a correction — and it is not a coincidence. It is an instance of a general structure that accounts for essentially every interest rate model with a closed form.

Definition 12.9 (Affine term structure model)**.** A model with state $X_t \in \mathbb{R}^n$ has an affine term structure if

$$P(t,T) = \exp\big(A(t,T) + B(t,T)^\top X_t\big)$$

for deterministic functions A and B.

The question is which state processes produce this, and the answer is clean.

Theorem 12.10 (Duffie and Kan)**.** *Suppose the short rate is affine in the state, $r_t = \delta_0 + \delta^\top X_t$, and the state is a diffusion whose drift and* covariance *are both affine in the state:*

$$dX_t = \big(a + bX_t\big)dt + \Sigma(X_t)\, dW_t, \qquad \big[\Sigma(x)\Sigma(x)^\top\big]_{ij} = c_{ij} + d_{ij}^\top x.$$

Then the model has an affine term structure, and A and B solve a system of ordinary differential equations — Riccati equations — in the maturity.

Proof. The bond price satisfies the pricing equation of chapter 5 with this model's generator — discounted, it is a martingale — which written out for a multi-factor state is

$$\frac{\partial P}{\partial t} + (a + bx)^\top \nabla P + \frac{1}{2}\operatorname{tr}\big(\Sigma\Sigma^\top \nabla^2 P\big) - r\,P = 0, \qquad P(T,T) = 1.$$

Substitute the guess $P = \exp(A + B^\top x)$. Every derivative brings down a factor of the exponential, which cancels throughout, leaving

$$\dot{A} + \dot{B}^\top x + (a + bx)^\top B + \frac{1}{2} B^\top \big(c + d^\top x\big) B - \delta_0 - \delta^\top x = 0,$$

where the dots are derivatives in t. Now the key step: this must hold for *every* value of x, and every term is either constant in x or linear in it. A polynomial vanishing identically has vanishing coefficients, so the constant and linear parts must separately be zero:

$$\begin{aligned} \dot{A} + a^\top B + \tfrac{1}{2} B^\top c\, B - \delta_0 &= 0, \\ \dot{B} + b^\top B + \tfrac{1}{2} B^\top d\, B - \delta &= 0, \end{aligned}$$

with $A(T,T) = 0$ and $B(T,T) = 0$. These are ordinary differential equations in one variable, quadratic in the unknown — Riccati equations. □

Remark (What "solvable" actually means here)**.** The pricing problem for a general model is a partial differential equation in n space dimensions plus time, and solving it costs work that grows exponentially in n. What the affine structure achieves is to reduce that PDE to a system of $n + 1$ *ordinary* differential equations, which cost essentially nothing and are often available in closed form.

The mechanism is the separation in the proof: because the coefficients are affine and the guess is exponential-affine, the state variable x appears only linearly and can be matched off, leaving equations for functions of maturity alone. The state has been eliminated from the problem entirely.

The condition is genuinely restrictive. The drift being affine is mild. The requirement that the *covariance* be affine in the state is what excludes almost everything: a volatility proportional to x makes the covariance proportional to x^2, and the match fails. This is why the tractable models are the Gaussian ones (Σ constant) and the square-root ones ($\Sigma \propto \sqrt{x}$, so the covariance is proportional to x) and essentially nothing else. Heston's variance process in chapter 10 is a square-root process for exactly this reason, and its closed-form characteristic function is Theorem 12.10 applied to the log-price rather than to a bond.

Structure (Why affine gives Riccati, in one line)**.** The system above is usually presented as what falls out of substituting an ansatz. See why the ansatz works, because the reason generalises and the algebra does not.

Apply the generator of chapter 3 to an exponential, $f(x) = e^{\psi x}$, for a diffusion whose drift is $b + \beta x$ and whose variance is $a + \alpha x$:

$$\mathcal{L} e^{\psi x} = \Big[\underbrace{b\psi + \tfrac{1}{2} a \psi^2}_{\text{constant}} + x \big(\underbrace{\beta\psi + \tfrac{1}{2} \alpha \psi^2}_{\text{coefficient of } x} \big) \Big] e^{\psi x}.$$

The generator has mapped an exponential-affine function to an *affine multiple* of itself. It does not say the family is mapped into itself: $(c_0 + c_1 x)\, e^{\phi + \psi x}$ is not of the form $e^{\phi' + \psi' x}$, so $\{e^{\phi + \psi x}\}$ is not preserved by $\mathcal{L}$. What has happened is better suited to the purpose.

Regard the family as a two-dimensional surface in function space, with coordinates (ϕ, ψ). Differentiating along each coordinate gives the tangent directions,

$$\frac{\partial}{\partial \phi} e^{\phi + \psi x} = e^{\phi + \psi x}, \qquad \frac{\partial}{\partial \psi} e^{\phi + \psi x} = x\, e^{\phi + \psi x},$$

so the tangent space at any point of the surface is precisely the set of affine multiples of that point. The display above therefore says that $\mathcal{L}u \in T_u\mathcal{S}$ at every point u of the surface $\mathcal{S}$: the vector field $\mathcal{L}$ on the space of functions is tangent to the *surface*, not merely to the space around it. That is not automatic. The directions $\partial_\phi u$ and $\partial_\psi u$ are tangent by construction, being derivatives along the surface; $\mathcal{L}u$ is a derivative in x, which points out of the surface unless the coefficients cooperate. With a variance quadratic in x the term $x^2 e^{\psi x}$ appears, and the field leaves it.

That is exactly the condition for the surface to be invariant under the flow $\partial_t u = \mathcal{L}u$, which is the object we actually care about — invariance under the semigroup, not under the generator. A flow tangent to a two-dimensional surface is a pair of ordinary differential equations in its coordinates, whatever the underlying problem was. And because both sides of $\partial_t u = \mathcal{L}u$ now lie in the same tangent space, they can be compared in its basis $\{1, x\} \cdot e^{\phi+\psi x}$, which is all the next line does. Chapter 14 takes this distinction as its starting point and makes it general.

Reading the bracket gives them directly. The constant term is ϕ' and the coefficient of x is ψ', so

$$\phi' = b\psi + \tfrac{1}{2}a\psi^2, \qquad \psi' = \beta\psi + \tfrac{1}{2}\alpha\psi^2,$$

and the second is quadratic in ψ — a Riccati equation — for the single reason that the *variance* was allowed to depend on the state. Vasicek has $\alpha = 0$, the quadratic term disappears, and the equation is linear; the square-root model has $\alpha \neq 0$ and it is not.

So "affine" is not a description of the coefficients so much as a statement about the generator, and the Riccati equation is not a coincidence but the shadow of a two-dimensional invariant family. Chapter 14 asks which other families work.

Example 12.4 (Solving the Riccati equations for Vasicek)**.** The abstract statement is more convincing once the equations have actually been solved once, and in one dimension they can be.

Take $n = 1$, $r_t = x_t$, and $dx_t = \kappa(\bar{x} - x_t)dt + \sigma\, dW_t$: the Vasicek model. In the notation of Theorem 12.10, $a = \kappa\bar{x}$, $b = -\kappa$, the covariance is the constant σ^2 so $c = \sigma^2$ and $d = 0$, and $\delta_0 = 0$, $\delta = 1$. Writing $\tau = T - t$ so the equations run forward in τ from zero, they become

$$\begin{aligned} \frac{dB}{d\tau} &= -\kappa B - 1, & B(0) &= 0, \\ \frac{dA}{d\tau} &= \kappa\bar{x}B + \tfrac{1}{2}\sigma^2 B^2, & A(0) &= 0. \end{aligned}$$

Because $d = 0$ the first equation is *linear* — the quadratic term of the Riccati equation has vanished — and it integrates immediately:

$$B(\tau) = -\frac{1 - e^{-\kappa\tau}}{\kappa},$$

which is $-G$ from (12.15), arrived at yet again. The second equation is then a plain integral of known functions, giving A in closed form.

So a bond price in the Vasicek model costs two evaluations of an exponential. This is what Theorem 12.10 is worth: the same problem posed as a partial differential equation would have to be solved numerically on a grid, for every maturity, every time the parameters moved.

Notice also exactly where the linearity came from. It came from $d = 0$, that is, from the covariance not depending on the state — the Gaussian case. In the square-root case $d \neq 0$, the

quadratic term survives, and B solves a genuine Riccati equation. It still has a closed form, which is why the square-root process is the other tractable one, but the algebra is a page rather than a line.

Remark (Quadratic Gaussian models)**.** There is one more family that solves, and it looks at first like a counterexample. Take X to be an ordinary Gaussian process — constant Σ — but make the short rate *quadratic* in it:

$$r_t = X_t^\top Q X_t + \delta^\top X_t + \delta_0.$$

Guessing $P = \exp(A + B^\top x + x^\top C x)$ and repeating the argument of Theorem 12.10, the exponential again cancels and one matches the constant, linear *and quadratic* coefficients in x, obtaining ordinary differential equations for A, B and a matrix Riccati equation for C. The model solves.

It is not really an exception. Adjoin the products $X_i X_j$ to the state; then r is affine in the enlarged state, and the enlarged state's drift and covariance are affine in it too, because X is Gaussian. Quadratic Gaussian models are affine models wearing a smaller state.

What one buys is a smile. With Q non-zero, the rate is a quadratic function of a Gaussian variable, so it is not Gaussian, and it is bounded below if Q is positive definite — which gives a distribution with genuine skew and curvature while keeping closed-form bonds.

12.4 Where the Smile Lives

Return now to the quasi-Gaussian state equations (12.14). Nothing so far has said what σ_r is, and this is where the volatility chapters come back.

The model is called quasi-Gaussian because it is Gaussian whenever σ_r is deterministic, and departs from Gaussian only through the state dependence of σ_r.

σ_r	name	what it gives
$\sigma(t)$	Hull-White	Gaussian, closed forms, no smile
$\sigma(t)\big[a(t) + b(t)x\big]$	linear quasi-Gaussian	skew; a sets the level and b is the skew dial
$\sigma(t)\big[a(t) + b(t)x + c(t)x^2\big]$	quadratic quasi-Gaussian	skew and curvature: a genuine smile
$\sqrt{z_t}\,\big[a + bx + cx^2\big]$	stochastic-vol quasi-Gaussian	vol-of-vol, forward smile, real vega

The state x is a one-dimensional diffusion with volatility $\sigma_r(x)$, so it is a local volatility model — of the state rather than of the underlying, but the same object chapter 9 studied. Two results carry straight over.

Gyongi's theorem says the distribution of x_T depends on the volatility only through its value at each level, so the shape of σ_r as a function of x is precisely what shapes the distribution, and therefore the smile. Nothing else about the model can affect it.

And chapter 10's midpoint rule says how: the implied volatility at a strike is σ_r averaged over the journey from the forward to that strike. Averaging preserves shape and halves slopes. So a σ_r that is

- constant averages to a constant, and the smile is flat;

- linear in x averages to something linear, and the smile is a straight tilt at half the slope;
- quadratic in x averages to something quadratic, and the smile acquires curvature.

The quadratic term is the first shape that survives averaging as a *bend*, which is why it is the first one that produces a smile rather than a skew.

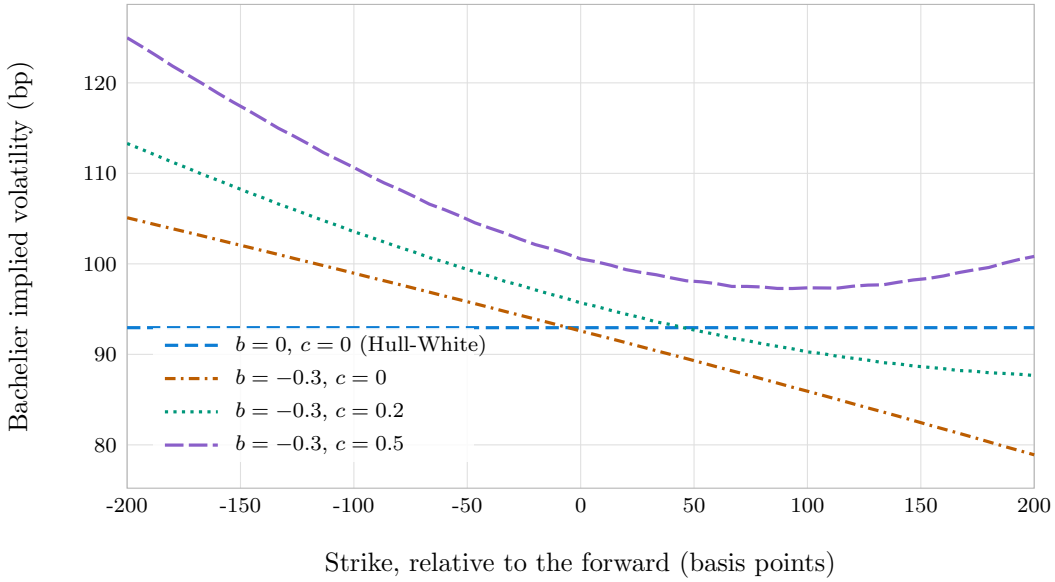

Figure 12.1: The smile a quadratic σ_r produces. Each curve solves the model rather than approximating it: the Fokker-Planck equation for the state is integrated forward and the options priced against the resulting density, so the curvature shown is the curvature the model has. With $b = c = 0$ the model is Hull-White and the Bachelier smile is exactly flat. The linear term tilts it without bending it. The quadratic term bends it, and more so as c grows.
Drawn by `QuasiGaussian::sigma_r`, `QuasiGaussian::density` and `QuasiGaussian::smile`.

Two details stand out.

The first is that the bent smile is not centred at the money. Its lowest point does not sit at a strike equal to the forward, and the averaging rule says exactly where it sits. Averaging $\sigma_r(u) = \sigma(1 + bu + cu^2)$ over the journey from the forward out to u,

$$\frac{1}{u}\int_0^u \sigma\big(1 + bz + cz^2\big)\,dz \;=\; \sigma\left(1 + \frac{b}{2}\,u + \frac{c}{3}\,u^2\right),$$

so the tilt is halved and the bend is divided by three. The smile is therefore least at

$$u_{\min} = -\frac{3b}{4c}, \tag{12.18}$$

against $-b/(2c)$ for σ_r itself: the vertex moves *outward* by half again, because averaging damps the quadratic term harder than the linear one and so leaves the tilt relatively stronger.[9]

[9] `the test that pins this`.

The plot spans ±200 basis points, which at these parameters is ±0.89 of a $\sigma\sqrt{T}$ move. The $b = -0.3$, $c = 0.5$ curve bottoms out at $u_{\min} = 0.45$, a hundred basis points above the forward, and that is on the picture. The $b = -0.3$, $c = 0.2$ curve bottoms out at $u_{\min} = 1.125$, some two hundred and fifty basis points above — off the right edge, so what is drawn is only the descent towards it.

Remark (Which mean, given that chapter 10 says harmonic)**.** The average taken above is the arithmetic one, and the short-expiry theorem of chapter 10 gives the implied volatility as the *harmonic* mean of the local volatility along the path. The two are not the same rule, so here is why the simpler one is being used and what it costs.

Write $\sigma_r = \sigma(1 + \varepsilon)$ with $\varepsilon(u) = bu + cu^2$. Expanding the reciprocal,

$$\left\langle \tfrac{1}{1+\varepsilon} \right\rangle^{-1} \;=\; 1 + \langle\varepsilon\rangle - \mathrm{Var}(\varepsilon) + O(\varepsilon^3),$$

so the harmonic mean is the arithmetic mean less the variance of the perturbation: the two agree to first order in (b, c) and part company only at second. And the vertex is a first-order statement: it is where the slope of $\langle\varepsilon\rangle$ in the strike vanishes, $\frac{d}{du}\langle\varepsilon\rangle = \frac{b}{2} + \frac{2c}{3}u = 0$, and the variance that separates the two rules is second order and moves it only at the next order, so (12.18) is what either rule gives at leading order.

What the difference is worth can be measured rather than bounded, since the density is solved exactly. Across the parameter sets used here the arithmetic rule places the vertex within 4.4% of the solved minimum and the harmonic rule within 3.4%.[10] The harmonic rule is better everywhere and by very little, and both leave a residual larger than the gap between them — because the theorem is a short-maturity limit and these are five year options. The choice of mean is not what the approximation is losing.

The consequence for a desk is that b and c do not act independently. The tilt and the bend share one vertex, and (12.18) places it by the *ratio* b/c rather than by either coefficient alone, so fitting a skew and a curvature is one fit rather than two.

The second is the units. The dials multiply $u = x/\text{scale}$ rather than x itself, and the scale is the typical move $\sigma\sqrt{T}$. That number is not chosen for convenience; it is the spread of the state.

Calculation 12.11 (Where $\sigma\sqrt{T}$ comes from)**.** Hold σ_r at σ for a moment, so x is the Hull-White state $dx = (y - \kappa x)dt + \sigma\, dW$. It is Gaussian, and integrating its variance gives

$$\mathrm{sd}(x_T) \;=\; \sigma\sqrt{\frac{1 - e^{-2\kappa T}}{2\kappa}} \;\longrightarrow\; \sigma\sqrt{T} \qquad \text{as } \kappa T \to 0. \tag{12.19}$$

So $\sigma\sqrt{T}$ is how far the state typically travels by T, in the limit of no mean reversion, and u counts standard deviations of the state. At $\sigma = 100$ basis points and $\kappa = 0.03$ the approximation is worth 99% of the exact spread at one year and 93% at five, falling to 68% at thirty — where mean reversion has begun to bite and the true spread is heading for its ceiling of $\sigma/\sqrt{2\kappa}$, about 410 basis points, while $\sigma\sqrt{T}$ goes on growing.

[10] `measured`.

With that, the level is

$$\sigma_r(x) \;=\; \sigma\left[1 + b\,\frac{x}{s} + c\,\frac{x^2}{s^2}\right], \qquad s = \mathrm{sd}(x_T) \;\approx\; \sigma\sqrt{T}, \tag{12.20}$$

with s the spread (12.19) and $u = x/s$. Then $u = 1$ means a one standard deviation move of the state, and b and c are dimensionless: b is how much the volatility tilts over one typical move, c how much it bends over one. Those are questions with comparable answers at every expiry, which is exactly what coefficients in fixed rate units would not give. A c multiplying x^2 in basis points describes a gentle bend where the state reaches a hundred of them and a violent one where it reaches three hundred, so the same number would mean something different at one year and at ten — a statement about the parametrisation rather than about the model. Normalising by the distance the state actually covers removes that. In rate units the coefficients of (12.20) are b/s on x and c/s^2 on x^2, so the same $c = 0.5$ at $\sigma = 100$ basis points is $0.5/100^2 = 5 \times 10^{-5}$ per basis point squared at one year, where $s \approx 100$, and $0.5/316^2 = 5 \times 10^{-6}$ at ten, where $s \approx 316$: a tenfold difference in the bend for the same dial.

It is a convention for reading the dials rather than a claim that one pair of dials fits every expiry. The model carries a single s in (12.20), fixed rather than recomputed as the state moves, and (12.19) is the number to set it to for the expiry being fitted. Fitting another expiry re-sets s, and with it the rate-unit coefficients b/s and c/s^2, while b and c keep their meaning.

Remark (Then why not simply use a local volatility model?)**.** Dupire's construction fits every quoted option exactly and needs no dials at all, so the resemblance above invites an obvious question. The answer is that the two are not competing for the same job.

They model different objects. A local volatility model gives the marginal distribution of one underlying. This gives the entire curve — every forward rate, every bond, every swap rate, at every date — out of the same two numbers. One can certainly fit a Dupire surface to the ten year swap rate, but it says nothing about the two year, nothing about the same ten year rate seen from a future date, and nothing that guarantees the two fits are consistent with any single arbitrage-free model. Here they are one model by construction.

The volatility is indexed differently. Dupire's $\sigma_{\mathrm{loc}}(t, K)$ is a function of calendar time and strike. Here $\sigma_r(x)$ is a function of the *state*, with no t in it and no strike. That is the same distinction chapter 10 turned on: a shape written in the calendar decays along the calendar, and a shape written in the state does not. What forward smile term structure this model has therefore comes from its own state settling down — y starts at zero and climbs towards $\sigma^2/2\kappa$ over a mean reversion time — and not from a dated coefficient. It is the same caveat chapter 10 attached to starting Heston away from its long-run variance.

The state is small, and that is the entire point of the chapter. Two numbers fit on a grid, so backward induction works and a Bermudan can be priced. Dupire's surface is not the obstacle there; the obstacle is that a local volatility model of a swap rate is not a model of the curve the callable depends on.

And the fit is worse. This is the cost. Three parameters fit the *shape* of a smile; they do not reprice a strip of quotes to the penny, and chapter 10's local-stochastic volatility construction exists precisely because desks want both. What is bought for that is stability: a shape with three dials cannot develop the local ripples that differentiating a noisy surface twice in strike produces, which is the instability chapter 9 warns about.

The last row of the table adds an independent variance factor, usually a square-root process $dz = \theta(1 - z)dt + \eta\sqrt{z}\,dZ$, chosen to be square-root for the reason the previous section gave. The parallel with chapter 11 is exact: the polynomial in x is the leverage function, carrying the fit, and z is the stochastic part, carrying the dynamics. The same division of labour, applied to a curve.

Remark (The quadratic term will try to explode)**.** The standard existence theorem asks for a diffusion coefficient of *at most* linear growth: locally Lipschitz coefficients obeying $|\sigma(x)| \le C(1 + |x|)$ give a unique strong solution for all time. A quadratic coefficient grows faster than that, so the theorem does not apply — the growth bound is the hypothesis being broken, not the conclusion being met.

Losing a sufficient condition is not by itself losing the solution. For a driftless diffusion $dX = \sigma(X)dW$, Feller's test says the process reaches infinity in finite time with positive probability exactly when

$$\int^{\infty} \frac{x}{\sigma^2(x)}\,dx \;<\; \infty,$$

and the two cases sit on either side of that. With $\sigma(x) = x$ the integrand is $1/x$ and the integral diverges, which is why geometric Brownian motion never explodes. With $\sigma(x) = x^2$ it is x^{-3} and the integral converges: the process does reach infinity in finite time, and the model's moments go with it.

A quadratic σ_r can also go negative, which is meaningless.

Production implementations therefore cap and floor σ_r, or replace the raw polynomial by a smooth interpolation between an absolute and a proportional regime. This is not a numerical detail bolted on afterwards; a quadratic-volatility model without it is not well posed.

12.5 More Than One Factor

Everything above used a single Brownian motion, and that carries a consequence which is easy to miss and expensive to ignore.

Look again at (12.17). With one factor, the whole curve is driven by the single state x_t, so every point of it moves in the same direction on any given day, by an amount $e^{-\kappa(T-t)}$ that depends only on maturity. The two year rate and the thirty year rate are *perfectly correlated*. The curve can shift and it can steepen as it shifts, but it cannot do the two independently, and it can never twist.

That is empirically false. A principal component analysis of a year of Treasury curve changes finds three shapes, of which a one-factor model can represent only the first.

Calculation 12.12 (How the components were found, and what a loading is)**.** The figure is a principal component analysis. The same decomposition returns in chapter 13 as the low-rank form imposed on a correlation matrix.

The data. One row per business day, one column per quoted tenor, and the entries are daily *changes* rather than levels. So the observations are days and the features are tenors — thirteen of them — and a single observation is one day's move of the whole curve, a vector $M \in \mathbb{R}^{13}$.

The decomposition. Subtract each tenor's mean change and form the covariance $C = \mathbb{E}[MM^{\top}]$, thirteen by thirteen and symmetric positive semi-definite. Any such matrix diagonalises with

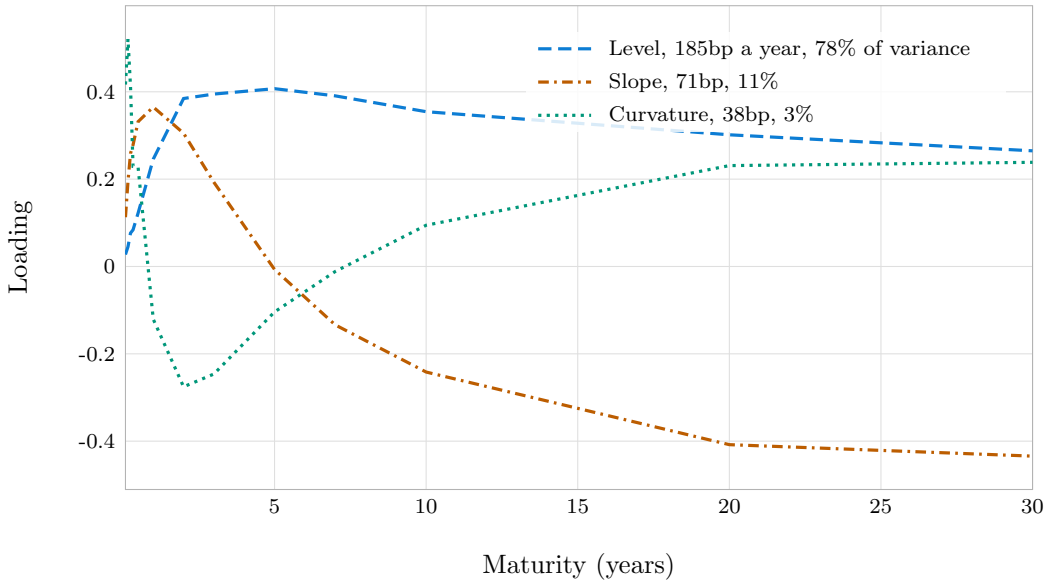

Figure 12.2: The three leading components of daily changes in the US Treasury par curve, over the year to the date shown. The first moves every tenor the same way and is the only shape (12.17) can produce — the one-factor loading $e^{-\kappa(T-t)}$ is a positive decay across maturities, so it lives entirely in this component. The second crosses zero once: the short end and the long end move in opposite directions, which is a steepening, and a one-factor model asserts it does not happen. The third crosses twice, the wings against the belly. Read the legend for what each costs to omit: the second and third carry a seventh of the variance of the first between them, which sounds ignorable and is not, because a hedged book has removed its exposure to the first.

Drawn by `curve_factors`. US Department of the Treasury, daily par yield curve rates, 2026-01-02 to 2026-09-22 (par yield, semiannual coupon, actual/actual). Retrieved from https://home.treasury.gov/interest-rates-data-csv-archive.

orthonormal eigenvectors,

$$C \;=\; \sum_{j} \lambda_j \, v_j v_j^{\top}, \qquad v_j^{\top} v_k = \delta_{jk}, \tag{12.21}$$

and the computation is nothing more than reading off the leading three: take the top eigenvector by power iteration, subtract $\lambda_1 v_1 v_1^{\top}$, and repeat.[11]

What that says about the curve. Equation (12.21) is exactly the statement that the day's move can be written

$$M_i \;=\; \sum_{j} \sqrt{\lambda_j}\; v_{j,i}\, W_j, \tag{12.22}$$

with W_j uncorrelated and of unit variance — multiply (12.22) out and its covariance is (12.21). So the analysis has rewritten thirteen correlated tenor moves as thirteen *independent* drivers, and the jth eigenvector is the vector of loadings of the tenors on the jth driver. That is the sense in which a curve has factors at all, and it is what the plotted shapes are.

Why three. Truncating (12.21) after d terms is the best rank-d approximation of C there is, in the sense that no other matrix of rank d is closer in Frobenius norm. So keeping three factors is not a convenient simplification but the optimal one at that rank, and the share the three account for is exactly what the approximation retains. Chapter 13 imposes the same truncation on a correlation matrix for a different reason — there it is the market's failure to identify all the entries rather than a wish to see the shapes — and inherits the same consequence, which the next paragraphs draw out: d factors admit exactly d shapes of curve move, and a payoff sensitive to a $(d+1)$th is priced as though it did not exist.

The omission is expensive for any product whose payoff depends on the shape of the curve rather than its level. The clearest case is a constant maturity swap spread option, which pays on the difference between a long and a short swap rate.

Perfect correlation does *not* make the spread deterministic. The two rates load on the state by different amounts — a short swap by nearly all of it, a long one by less, since the loading $e^{-\kappa(T-t)}$ has further to decay — so the difference still moves. What perfect correlation does is put a floor under it, and put the model on the floor.

Calculation 12.13 (How much a factor costs)**.** For two rates with volatilities σ_1 and σ_2 and correlation ρ, the spread has volatility

$$\sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\,\sigma_1\sigma_2} \;\longrightarrow\; \big|\sigma_1 - \sigma_2\big| \qquad \text{as } \rho \to 1, \tag{12.23}$$

and the right-hand side is the *minimum* of the left over all ρ. A one-factor model sets $\rho = 1$, since every rate is an increasing function of the same x, so it always delivers exactly that minimum.

Taking the loading of an n-year swap rate as the average of $e^{-\kappa u}$ over its life, $\beta(n) = (1 - e^{-\kappa n})/(\kappa n)$, at $\kappa = 3\%$:[12]

spread	$\rho = 1$ **(one factor)**	$\rho = 0.9$	**ratio**
2y against 10y	11%	44%	4.0
2y against 30y	32%	49%	1.5

[11] `risk::curve_factors`.

[12] `quasigaussian::spread_volatility`.

as a percentage of the two year rate's own volatility. So the one-factor model gives the two-year-against-ten-year spread about a quarter of the volatility a plausible correlation would, and prices the option accordingly. Not zero — a quarter.

Note where the damage is worst. It is not the widest spread but the narrowest, because there the two loadings are close, so $|\sigma_1 - \sigma_2|$ is nearly nothing while a correlation below one leaves two almost uncancelled volatilities behind. The 2s10s trade is both the standard one and the one a single factor handles worst.

And there is no dial. Once σ, b and c are fitted to swaptions, (12.23) is determined: the model has no parameter that moves the correlation, so it cannot be marked to a spread option's price even in principle. That is chapter 20's question answered in the negative — the payoff is a bet on decorrelation, and the model has not assumed a view on decorrelation so much as assumed it away.

The construction generalises without difficulty. With n Brownian motions, take $x_t \in \mathbb{R}^n$ and let y_t become an $n \times n$ matrix of accumulated covariances:

$$dx_t = \big(y_t \mathbf{1} - \kappa x_t\big)dt + \sigma_r^\top dW_t, \qquad \frac{dy_t}{dt} = \sigma_r^\top \sigma_r - \kappa y_t - y_t \kappa,$$

and the bond reconstruction keeps its shape with the scalar products becoming quadratic forms:

$$P(t,T) = \frac{P(0,T)}{P(0,t)} \exp\Big(- G(t,T)^\top x_t - \tfrac{1}{2} G(t,T)^\top y_t \, G(t,T) \Big).$$

The cost is dimensionality. The state is x together with the distinct entries of the symmetric matrix y, so $n + n(n+1)/2$ variables: two factors means five states, three means nine. Five is comfortable in a Monte Carlo with a regression on the states, and marginal on a grid; nine is Monte Carlo only. Since y carries no randomness of its own, its entries are often frozen at their expected paths in practice, which brings the effective dimension back down to n at some cost in accuracy.

What two factors buy is exactly the decorrelation the one-factor model cannot express, and with it the ability to price curve-shape products and to get a Bermudan's exercise boundary right — since the decision to exercise depends on the shape of the curve at that moment and not only on its level.

12.6 Calibration, and the Parameter Nobody Can See

Finally, how such a model is fitted — because the pattern of chapters 10 and 11 appears here a third time, and by now it should be recognisable.

- **The initial curve** is matched exactly and for free, by (12.15).
- $\sigma(t)$, the level, is bootstrapped to a strip of at-the-money European swaptions chosen to match the exotic's own exercise schedule.
- b **and** c, the skew and smile parameters, are fitted to the smile around those same swaptions.
- **The vol-of-vol** is fitted to how the smile's curvature varies with expiry, and is usually re-marked occasionally rather than daily.

- κ, the mean reversion, is not fitted to European swaptions at all.

That last point deserves a closer look. European swaptions are very nearly insensitive to κ: it can be moved a long way and the fit compensated by adjusting $\sigma(t)$, leaving vanilla prices essentially unchanged. But κ controls how strongly rates of different maturities move together, and therefore how much a Bermudan is worth over the European swaptions it could become.

So κ is the single largest model risk in a callable book, and it is invisible in the instruments used to calibrate. It is set instead from the term structure of rate correlations, historically or from a principal component analysis of the curve, or from constant-maturity-swap spread option prices, or by matching a Bermudan-to-European ratio the desk believes in.

If that sounds familiar, it should. It is β from chapter 10 and the mixing weight from chapter 11, in different clothing: a parameter that the calibration set cannot see and the product depends on. The recurring lesson of these last four chapters is that a model is not determined by fitting it, and the part that fitting leaves undetermined is usually the part that decides the answer.

Remark (The alternative is to stop insisting on a state)**.** Everything in this chapter was bought with one decision: keep the Markov state small enough to put on a grid. That decision is what constrains the volatility structure, what makes the vanilla fit approximate, and what leaves κ carrying the correlation that the calibration cannot see.

Chapter 13 takes the opposite decision. It models the quoted rates directly, so the calibration instruments price exactly and the correlation between rates is specified rather than inferred from a mean reversion — and it gives up the small state, so the exercise decision this chapter handles by backward induction becomes a regression on the simulated paths. The two chapters are the same trade seen from either end, and neither dominates: this one prices the exercise exactly and the vanillas approximately, that one the reverse.

References

- Bjork, T., & Svensson, L. (2001). On the existence of finite-dimensional realizations for nonlinear forward rate models. *Mathematical Finance*, 11(2), 205–243.

- Bjork, T., & Christensen, B. J. (1999). Interest rate dynamics and consistent forward rate curves. *Mathematical Finance*, 9(4), 323–348.

- Filipovic, D., & Teichmann, J. (2004). On the geometry of the term structure of interest rates. *Proceedings of the Royal Society A*, 460(2041), 129–167.

- Cheyette, O. (2001). Markov representation of the Heath-Jarrow-Morton model. Available at SSRN 6073.

- Ritchken, P., & Sankarasubramanian, L. (1995). Volatility structures of forward rates and the dynamics of the term structure. *Mathematical Finance*, 5(1), 55–72.

- Duffie, D., & Kan, R. (1996). A yield-factor model of interest rates. *Mathematical Finance*, 6(4), 379–406.

- Duffie, D., Pan, J., & Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, 68(6), 1343–1376.

- Andersen, L. B. G., & Piterbarg, V. V. (2010). *Interest Rate Modeling*, Volumes 1–3. Atlantic Financial Press.

Chapter 13

Market Models

We model the rates that are actually quoted rather than the instantaneous abstractions of the previous chapters, which makes the calibration instruments price exactly by construction and costs the low-dimensional state that made those chapters tractable. We build the Libor market model with its drift derived rather than quoted, do the same in swap rate coordinates, and then settle the choice between them — forwards and swap rates cannot both be lognormal, and how much that matters is a measurement rather than a preference. The answer reorders the modelling priorities: decorrelation is worth percentage points of volatility and the departure from lognormality is worth hundredths.

13.1 Modelling What Is Quoted

Chapter 8 gave the general framework for arbitrage-free curve dynamics and chapter 12 made it usable by forcing a low-dimensional Markov state. Both model the instantaneous forward rate $f(t,T)$, which is not a rate anybody trades. Every quoted instrument references an accrual over a real period — three months, six months — and the passage from the instantaneous object to the traded one is an approximation, small but present, in the pricing of every calibration instrument.

The market models invert that. They take as primitive exactly the rates the market quotes, and accept whatever dynamics follow. The reward is immediate: the calibration instruments price by the market's own formula, exactly, with nothing to approximate. The cost is the subject of most of this chapter.

Definition 13.1 (Tenor structure)**.** Fix dates $T_0 < T_1 < \cdots < T_N$ with accruals $\delta_i = \tau(T_i, T_{i+1})$ on the appropriate convention of chapter 7. The forward rates are

$$F_i(t) = \frac{1}{\delta_i}\left(\frac{P(t,T_i)}{P(t,T_{i+1})} - 1\right), \qquad i = 0, \dots, N-1, \tag{13.1}$$

so F_i is the rate that fixes at T_i and pays at T_{i+1}. The curve and the vector $(F_0, \dots, F_{N-1})$ carry the same information: (13.1) inverts to

$$P(t,T_k) = P(t,T_{\eta(t)}) \prod_{j=\eta(t)}^{k-1} \frac{1}{1+\delta_j F_j(t)}, \tag{13.2}$$

with $\eta(t)$ the index of the next date in the structure.

One observation about (13.1) does all the structural work in this chapter.

Lemma 13.2 (Each forward is a martingale under its own measure)**.** *F_i is a martingale under the T_{i+1}-forward measure, the measure of chapter 6 whose numeraire is $P(\cdot, T_{i+1})$.*

Proof. Rearranging (13.1),

$$\delta_i F_i(t)\, P(t, T_{i+1}) = P(t, T_i) - P(t, T_{i+1}),$$

and the right hand side is a portfolio of two traded assets — long a bond maturing at T_i, short one maturing at T_{i+1}. So $\delta_i F_i$ is the value of that portfolio divided by $P(\cdot, T_{i+1})$, which is the numeraire, and chapter 6's remark Prices in a numeraire are martingales in its measure showed that any traded asset divided by the numeraire is a martingale under the associated measure. □

Remark (Why that is more than a technicality)**.** Lemma 13.2 is not a property of a model. It holds in *any* arbitrage-free market, because it follows from (13.1) and nothing else. So a forward rate has a distinguished measure in which it has no drift, handed over by the definition, and the only remaining modelling choice is what its volatility is.

Choose that volatility to be deterministic and lognormal, $dF_i/F_i = \sigma_i(t)\, dW_i$, and a caplet on F_i — which pays $\delta_i(F_i(T_i) - K)^+$ at T_{i+1} — has value

$$P(0, T_{i+1})\, \delta_i\, \mathbb{E}^{T_{i+1}}\big[(F_i(T_i) - K)^+\big],$$

an expectation of a call payoff on a driftless lognormal variable, which is Black's formula exactly. Not approximately, and not after an adjustment. The market quotes caps in Black volatilities; this model's parameters *are* those volatilities.

That is the whole appeal: the model has been built so that its calibration is a change of notation. Every chapter before this one had to solve for parameters that reproduce quotes. Here the quotes are the parameters.

13.2 The Drift, and Why There Has To Be One

Lemma 13.2 gives each forward its own measure. A portfolio containing several of them has to be priced under *one*, and no measure makes them all driftless.

Theorem 13.3 (The terminal measure drift)**.** *Suppose $dF_i/F_i = \sigma_i\, dW_i$ under the T_{i+1}-forward measure for each i, with $d\langle W_i, W_j\rangle = \rho_{ij}\, dt$. Then under the terminal measure, with numeraire $P(\cdot, T_N)$,*

$$\frac{dF_i}{F_i} = -\sigma_i \sum_{j=i+1}^{N-1} \frac{\delta_j \rho_{ij} \sigma_j F_j}{1 + \delta_j F_j}\, dt + \sigma_i\, dW_i^N. \tag{13.3}$$

Proof. The change of measure from T_{i+2} to T_{i+1} is governed by the ratio of numeraires, and (13.1) says what that ratio is:

$$\frac{P(t,T_{i+1})}{P(t,T_{i+2})} = 1 + \delta_{i+1}F_{i+1}(t).$$

By chapter 6's theorem Change of numeraire, in drifts, moving from numeraire A to numeraire B adds to a process's drift the covariation of that process with $\log(B/A)$ — for a positive process, the covariation of its logarithm with $\log(B/A)$. Here $\log(B/A) = -\log(1+\delta_{i+1}F_{i+1})$, whose stochastic differential has volatility

$$-\frac{\delta_{i+1}\,dF_{i+1}}{1+\delta_{i+1}F_{i+1}} = -\frac{\delta_{i+1}F_{i+1}\sigma_{i+1}}{1+\delta_{i+1}F_{i+1}}\,dW_{i+1},$$

so moving F_i from its own measure to the T_{i+2} measure adds the drift

$$-\sigma_i\frac{\delta_{i+1}\rho_{i,i+1}\sigma_{i+1}F_{i+1}}{1+\delta_{i+1}F_{i+1}}\,dt.$$

Iterating from T_{i+1} out to T_N accumulates one such term per step, which is (13.3). □

Remark (Reading the drift)**.** Three features of (13.3) explain most of what a market model is like to work with.

The sum is empty for $i = N-1$. The last forward is a martingale under the terminal measure, since its own measure *is* the terminal measure. This is the cheapest available check on any implementation, and an error in the limits of the sum shows up here before anywhere else.[1]

The drift is state-dependent. F_j appears on the right hand side, so (13.3) is not a lognormal process with a known drift; it is a coupled system of N equations in which every forward's drift depends on the current level of all the later ones. There is no closed form for the joint law — the drift is what enforces no-arbitrage between rates that were each specified separately, which is chapter 8's drift condition arriving in discrete coordinates.

The sign is one-directional. Every term is negative, so under the terminal measure every forward but the last drifts down, and further the further it sits from the numeraire. A numeraire at the far end of the structure gives the near forwards large drifts, which are exactly the forwards a short-dated product depends on.

Structure (The drift is a change of coordinates, not a force)**.** The algebra can make it look like a modelling assumption and it is not one.

Each forward was specified as driftless in its own measure. Nothing about the market has been asserted beyond that. The drift appears purely because the forwards are being written in a common coordinate system, and a change of measure is a change of coordinates on the space of processes: the same process described from a different numeraire acquires a drift, in the way that a straight line acquires curvature when written in polar coordinates.

[1] `tested`, along with the drift growing with distance from the numeraire, and `a caplet repricing to Black` — which only works if the drift and the deflator cancel, and they cancel only if both are right.

13.3 The Numeraire Is a Choice, and Jamshidian's Is the Better One

The terminal measure is the easiest to derive and among the worst to simulate in. Two problems, and the first is fatal rather than inconvenient.

A payoff with cashflows beyond T_N cannot be priced at all, because there is no numeraire past the end of the structure. That sounds like an artefact of choosing N too small, and it is not: a callable structure whose exercise generates a swap extending past the last modelled date requires the bond $P(t,T)$ for $T > T_N$, which (13.2) cannot supply. Extending N moves the problem rather than solving it, because the terminal forwards then have no volatility data to calibrate against.

The second is that the drifts of the near forwards are the largest, by the sign argument above, so the discretisation error is concentrated exactly where a short-dated product lives.

Definition 13.4 (The spot Libor measure)**.** Let B be the value of a unit invested at T_0 and rolled at every date in the structure at the rate fixing then:

$$B(T_k) = \prod_{j=0}^{k-1} \big(1 + \delta_j F_j(T_j)\big), \tag{13.4}$$

interpolated between dates by the bond $P(t, T_{\eta(t)})$. Jamshidian's *spot Libor measure* is the measure associated with B as numeraire.

This is a discretely rebalanced money market account, and it is the natural numeraire for a market model in the same way that the continuously compounded account is natural for a short rate model — with the difference that (13.4) is built entirely from rates the model already has, so it requires nothing that is not modelled.

Remark (What changes)**.** Repeating the change of numeraire argument with B in place of $P(\cdot, T_N)$ gives

$$\frac{dF_i}{F_i} = \sigma_i \sum_{j=\eta(t)}^{i} \frac{\delta_j \rho_{ij} \sigma_j F_j}{1 + \delta_j F_j}\, dt + \sigma_i\, dW_i^B, \tag{13.5}$$

and the differences from (13.3) are exactly the ones wanted. The sum now runs *forward* from the front of the structure to i, so it involves only forwards that have not yet fixed — and it is empty for the front forward, which is therefore the one with no drift, rather than the far one. The sign is positive, the numeraire never expires, and a product with cashflows at many dates is priced by discounting each on the same path of B rather than by deflating everything to a single far date.

The cost is that B is path-dependent: (13.4) depends on the rates that fixed along the way, so a path has to be carried to evaluate it. That is not an obstacle to a Markov state. B does not enter the dynamics (13.5), so the vector of forwards is Markov on its own, and a value at t needs only the ratio $B(t)/B(T)$, in which the fixings up to t cancel, so it is a function of the current forwards alone. For a simulation the path is being generated anyway, and the numeraire costs nothing. What the model gives up is a *small* state, not a Markov one, as the next section says.

13.4 No Small State

That giving-up is the central cost of the whole construction.

Chapter 12 went to considerable trouble to force the curve's evolution into a low-dimensional Markov state, because a low-dimensional state is what permits a partial differential equation, and a partial differential equation is what permits early exercise to be handled by backward induction. The market model has no such state. Its state is the whole vector $(F_{\eta(t)}, \dots, F_{N-1})$ — forty numbers for a ten year quarterly structure, eighty for twenty years — and chapter 19 measures what a grid does in forty dimensions.

So a market model is a simulation model, necessarily and not by preference. Everything follows from that:

- European payoffs are straightforward: simulate to the expiry, average the payoff over paths.
- Early exercise requires the continuation value at each exercise date as a function of a forty-dimensional state, which must be estimated by regression on the simulated paths, and the choice of regressors is a modelling decision that no calibration constrains. Chapter 19 is where that difficulty is treated; here the point is that the model's structure forces it.
- Greeks come from differentiating a simulation, with the accuracy problems chapter 19 sets out, rather than from reading a derivative off a grid.

Structure (The trade that this chapter is)**.** Set the two families side by side and the trade is clean.

A short rate or quasi-Gaussian model has a small Markov state, so it admits a partial differential equation, backward induction, and exact early exercise — and it prices its calibration instruments approximately, with a volatility structure constrained by the requirement that the state stay small. Chapter 12 measures what that constraint costs.

A market model prices its calibration instruments exactly and imposes no constraint on the volatility structure — and it has no small state, so early exercise becomes a regression problem whose error is not controlled by anything the calibration sees.

Neither is better. The choice is which error you would rather have, and that is decided by the product: the exact-calibration model is right when the answer is dominated by the vanilla prices, and the small-state model is right when it is dominated by the exercise decision. Chapter 25 returns to this, because it is the clearest instance in the book of a compromise that was forced by compute and is now a genuine choice.

13.5 The Same Curve in Swap Rate Coordinates

Nothing in the construction was specific to forward rates. The argument was: find a quantity that is a traded asset divided by a numeraire, declare it lognormal in that numeraire, and the corresponding option prices by Black. Swap rates satisfy the same description.

Lemma 13.5 (A swap rate is a martingale under its annuity measure)**.** *The par swap rate $S_{a,b}$ of chapter 7 is a martingale under the measure whose numeraire is the annuity $Ann_{a,b}(t) = \sum_{k=a+1}^{b} \delta_{k-1} P(t, T_k)$.*

Proof. The par rate is defined by the fixed leg matching the floating leg, which gives

$$S_{a,b}(t)\,\mathrm{Ann}_{a,b}(t) = P(t,T_a) - P(t,T_b),$$

again a portfolio of two bonds. So $S_{a,b}$ is a traded asset divided by the annuity, and the annuity is a positive portfolio of bonds and hence an admissible numeraire. □

Declaring $dS_{a,b}/S_{a,b} = \sigma_{a,b}\,dW$ under that measure makes a swaption — which pays $\mathrm{Ann}_{a,b}(T_a)(S_{a,b}(T_a)-K)^+$ — price by Black exactly, with the annuity as the discount factor. That is the swap market model, and for the *co-terminal* family $S_{a,N}$, $a = 0,\dots,N-1$, it plays the same role for swaptions that the Libor market model plays for caps.

Theorem 13.6 (Both cannot be lognormal)**.** *If every forward F_i is lognormal under its own measure, then the swap rates are not; and if the co-terminal swap rates are lognormal, the forwards are not.*

Proof. By (13.2) and Lemma 15.1,

$$S_{a,b} = \frac{P(t,T_a) - P(t,T_b)}{\sum_k \delta_{k-1} P(t,T_k)} = \sum_{i=a}^{b-1} w_i(F)\,F_i, \tag{13.6}$$

where the weights w_i are ratios of bonds and therefore themselves functions of the forwards. So a swap rate is a weighted sum of the forwards, and a weighted sum of lognormal variables is not lognormal — the family is not closed under addition, as it is under multiplication. The same argument run backwards inverts the conclusion. □

So the two models are not two descriptions of one thing. They are two incompatible specifications, and at most one of them can price both caps and swaptions by the market's formula. The choice of coordinates is a choice of which set of quotes to get exactly right.

13.6 Which Coordinates, and What the Answer Costs

Theorem 13.6 says there is an approximation. It says nothing about its size, and the size is what decides the question, so it has to be measured.

Calculation 13.7 (What a forward-based model costs on swaptions)**.** Take a flat curve at 4% on ten semiannual periods, every forward lognormal at 25% volatility, correlations decaying exponentially with the gap between maturities,

$$\rho_{ij} = e^{-\beta_\rho\,|T_i - T_j|}, \tag{13.7}$$

so β_ρ is a decay rate per year: at $\beta_\rho = 0$ every forward moves with every other, and a large β_ρ leaves forwards a few years apart nearly independent. The forwards have no smile whatever, by construction. Simulate, price swaptions on the five year swap starting in two years across strikes from 70% to 130% of the forward, and invert to Black implied volatility.[2]

[2] `measured`.

The induced smile spans under a tenth of a volatility point. At a hundred thousand paths the shape is not stable from one seed to the next. Running two million paths over several seeds a consistent shape does appear, implied volatility rising with strike by about three hundredths of a point across that strike range. Resolving it took eight million paths.

The relevant comparison is a swaption bid-offer, which is a large fraction of a volatility point. So the approximation is one to two orders of magnitude below the price at which the instrument can be traded.

Calculation 13.8 (And what the correlation is worth)**.** The same model, at the money, with the volatility of the swap rate read off its own option price rather than assumed.[3]

Every forward carries 25% volatility. The swap rate averaging them comes out near 22.8%, because (13.6) is an average of imperfectly correlated things and an average moves less than its parts. Increasing the correlation decay — decorrelating the forwards further — lowers it again, measurably.

So the correlation structure moves the at-the-money swaption volatility by whole percentage points, while the departure from lognormality moves the smile by hundredths of one.

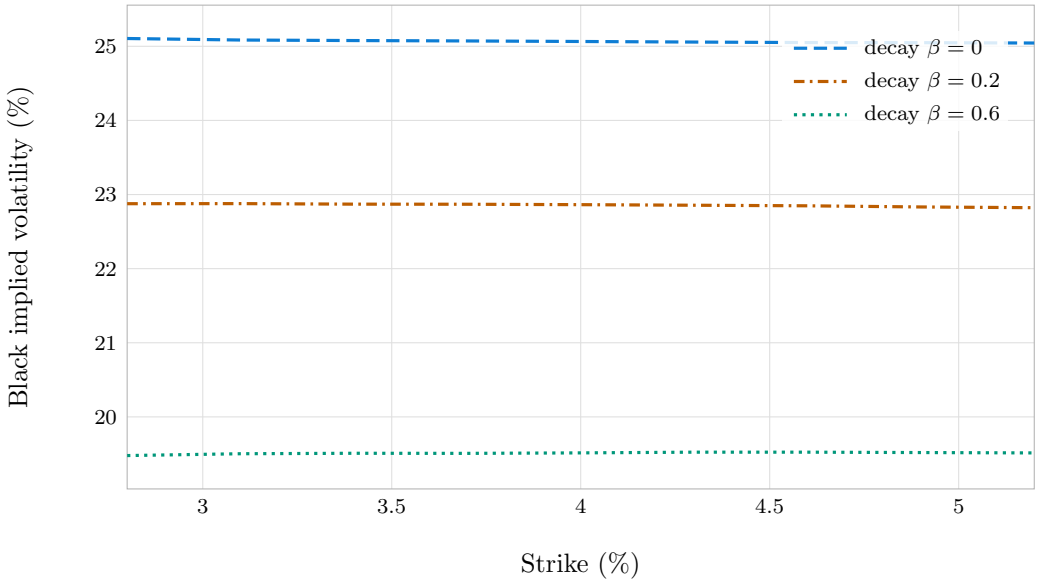

Figure 13.1: The swaption smile at three correlation decays, with every forward carrying the same 25% volatility so that nothing on the picture comes from the inputs. At $\beta_\rho = 0$ the forwards move together and the swap rate inherits their volatility exactly, which is the internal check: 25.1% against a 25% input. Decorrelating them takes the level to 22.9% and then 19.6%, while each curve stays flat across strikes to within a twentieth of a point. The vertical spread between the curves is what the correlation is worth; the tilt within each is what the departure from lognormality is worth, and the ratio of the two is why this chapter argues the correlation first.
Drawn by `Lmm::simulate` and `swaption_smile`.

[3] `measured`.

Remark (Two things these numbers do not say)**.** *They do not say the model produces a smile.* A tenth of a volatility point is not a smile mechanism; it is the residue of a coordinate mismatch, measured in a model built to have no smile at all — every forward is lognormal by construction. A lognormal market model cannot fit a market smile in either set of coordinates, which is what §13.9 is about. What the calculation says is narrower and more useful: the *approximation* incurred by choosing one coordinate system over the other is not the thing standing between the model and the market. The smile has to be put in separately, and doing so is orthogonal to the question this section is asking.

They do not say the correlation is unobservable. It is quoted, if indirectly. Because the swap rate's variance is the double sum (13.9), every swaption price is a statement about ρ, and the next section calibrates it that way: caps pin the diagonal and swaptions pin the off-diagonal. What makes the correlation the largest risk in the model is not that the market is silent about it but that the market does not say *enough* — a hundred or so liquid swaptions against the $N(N-1)/2$ entries of a full matrix — so the fit proceeds under a low-rank restriction that is an assumption rather than a measurement.

Structure (The measurement reorders the priorities)**.** Put Calculation 13.7 and Calculation 13.8 together and they answer a question that is usually argued about on grounds of taste.

The thing a swap market model buys — exact pricing of its own calibration instruments — is worth hundredths of a volatility point. The thing both models must get right, and which neither gets for free, is the correlation between rates, and that is worth percentage points. The two differ by two orders of magnitude, so a decision made on the first while ignoring the second is a decision made on the wrong axis.

That does not make the choice of coordinates irrelevant, but it relocates the argument. It is not about calibration accuracy. It is about which coordinates make the *payoff* simple:

- A Bermudan swaption's exercise decision at each date compares a swap rate with a strike. In co-terminal swap rate coordinates that is a comparison of one state variable with a number; in forward coordinates it is a comparison of a nonlinear function of many state variables with a number, and the regression that estimates the continuation value has to learn that function. So the swap market model makes the exercise boundary simple, which is where the uncontrolled error of the previous section lives.

- A product touching several tenors, or a book that must be priced consistently across caps, swaptions and exotics, cannot use co-terminal swap rates at all: the family is tied to one terminal date, and a second product with a different terminal date needs a different, mutually inconsistent, model. The forward family spans everything by (13.2).

Which gives the criterion. Use swap rate coordinates for a book of Bermudan swaptions on one tenor structure, where the exercise decision is the whole difficulty and the calibration is a read-off. Use forward coordinates for anything that has to span tenors, and pay a cost that Calculation 13.7 bounds at a fraction of a bid-offer. And in either case spend the modelling effort on the correlation, because Calculation 13.8 is where the money is.

13.7 Calibration

What has to be chosen is the volatility function $\sigma_i(t)$ for each forward and the correlation ρ_{ij} between them. The instruments available are caps, which price each forward on its own, and swaptions, which price combinations.

A caplet on F_i depends on σ_i only through

$$\int_0^{T_i} \sigma_i(t)^2 \, dt, \tag{13.8}$$

the accumulated variance to its own expiry, and on no correlation at all. A swaption depends on the whole matrix, since by (13.6) the swap rate's variance is

$$\sigma_{a,b}^2 \approx \frac{1}{S_{a,b}^2} \sum_{i,j} w_i w_j F_i F_j \rho_{ij} \int \sigma_i \sigma_j \, dt. \tag{13.9}$$

So caps pin the diagonal and swaptions pin the off-diagonal. That is a clean identification statement of the kind chapter 20 asks for, and it is the reason both instrument sets are needed: caps alone leave the correlation entirely free, which by Calculation 13.8 is the parameter that matters most.

Remark (Time-homogeneity is not a convenience)**.** Equation (13.8) constrains only the integral, so infinitely many $\sigma_i(\cdot)$ fit the same caplet. The standard resolution is to insist that the volatility depend on time to maturity rather than on calendar time,

$$\sigma_i(t) = \phi_i \, g(T_i - t), \tag{13.10}$$

with a shared shape g and a per-forward scaling ϕ_i absorbing the fit.

This is usually presented as a parsimonious parametrisation. It is more than that, and chapter 10 has already made the argument in another setting. A calendar-time volatility fitted to today's caps will generally have σ_i collapsing as t approaches T_i in whatever way the fit requires, and that collapse is a statement about the volatility structure the model will have *tomorrow*. A model whose implied volatility surface a year from now is nothing like today's cannot be used for anything whose value depends on the surface at a future date — which is precisely what a callable is. Time-homogeneity is the discipline that stops the calibration from spending the future to fit the present, and it is the same discipline that chapter 10 demanded of a stochastic volatility model.

Remark (The correlation matrix has too many entries)**.** A full ρ has $N(N-1)/2$ parameters, which for a forty-forward structure is seven hundred and eighty, against perhaps a hundred liquid swaption quotes. The market does not identify it, and chapter 20's warning applies exactly: a fit that succeeds does not mean the parameters were determined.

What is done is to impose a low-rank form, $\rho = LL^\top$ with L of d columns. This is the truncated spectral decomposition of chapter 12's calculation How the components were found, and what a loading is: writing $\rho = \sum_j \lambda_j v_j v_j^\top$ and keeping the largest d terms gives L with columns $\sqrt{\lambda_j}\, v_j$, and no matrix of rank d approximates ρ better. It is the same reduction that calculation made for a different reason and which chapter 24's calculation What a one-factor report cannot see justifies — three or four factors carry almost all the variance. The parametrisation used above, $\rho_{ij} = e^{-\beta|T_i - T_j|}$,

is the one-parameter version: it has the right qualitative shape, decaying with separation, and one number to fit.

What carries over from chapter 12's calculation How the components were found, and what a loading is is the algebra: a symmetric positive semi-definite matrix, its spectral decomposition, and a truncation that is optimal at its rank. What differs is three things. That calculation decomposes a *covariance* estimated from a history of curve moves, so its eigenvectors are a measurement and the truncation discards information that was in the data. Here the matrix is a *correlation*, because the volatilities are carried separately and calibrated from caps, and it is not estimated from anything — it is a parametric shape, or the output of a fit to swaptions. So the truncation is not discarding a measurement; it is the reason the fit is identified at all, since the full matrix has more entries than there are prices.

And a correlation matrix does not survive truncation as cleanly as a covariance does. Keeping d terms of $\sum_j \lambda_j v_j v_j^\top$ leaves a matrix whose diagonal entries are below one, since each is $\sum_{j \le d} \lambda_j v_{j,i}^2$ and the discarded terms were positive. A correlation matrix with a diagonal below one is not a correlation matrix, so the rows of L are rescaled to unit length afterwards. That restores the diagonal and perturbs every off-diagonal entry slightly, and the perturbation — not the truncation — is what a low-rank correlation actually costs.

One thing does carry across unchanged. A correlation between diffusions is a property of the diffusion coefficients, which chapter 6's structure A measure change moves the drift and cannot touch the volatility showed a change of measure leaves alone, so a shape estimated from history is an admissible input to a pricing model in a way that a drift estimated from history would not be. That is why desks are willing to take the *shape* of the correlation from a historical decomposition and fit only its overall level to swaption prices — and it is the same argument that licensed measuring a backbone from a time series in chapter 10.

The choice of d then has a consequence that is easy to miss. A d-factor model can produce at most d independent shapes of curve move, so a payoff sensitive to a $(d+1)$th shape is priced as though that shape did not exist. This is chapter 12's calculation How the components were found, and what a loading is arriving in market model coordinates, and the resolution is the same: choose d by what the payoff is exposed to, not by what fits.

13.8 Simulating It

Three practical points, each of which is a place the implementation can be wrong while appearing to work.

Use logarithms. Simulating $\log F_i$ keeps each forward positive by construction and makes the diffusion term exact, leaving discretisation error only in the drift.

The drift must be frozen, and that is the scheme's only error. Equation (13.5) depends on the current forwards, so a step requires the drift at some point within it, and using its value at the start is the standard choice. The resulting bias is not detectable by looking at the output; it is detectable by the caplet test, which compares a simulated price against a closed form that shares none of the simulation's machinery. That test is the reason to trust the implementation, and no amount of internal consistency substitutes for it.

Check the martingales. Every forward has a measure in which it is driftless, and under the simulated measure the front forward in (13.5) and the last in (13.3) are driftless outright. A simulation whose martingales drift has a bug, and the check costs one line.

Remark (Where the real error is)**.** None of the above is the dominant error in a Bermudan price. That is the regression for the continuation value.

The continuation value is a function on a forty-dimensional state space, estimated from paths. Whatever basis is chosen, it is a low-dimensional approximation of a high-dimensional function, and its error is a bias in the exercise decision — exercising when continuing was better, or the reverse. The bias is one-signed in a way that flatters the price, because a suboptimal exercise rule always produces a lower value, so the simulated price of a Bermudan is biased *low* by an amount nobody can compute directly.

Which returns to the trade of this chapter. The small-state models of chapter 12 handle the exercise decision exactly and the calibration approximately. This model handles the calibration exactly and the exercise decision approximately, and the approximation in the exercise decision is the larger of the two for exactly the products this model exists to price.

13.9 What Is Missing, and It Is the Smile

Everything above had deterministic volatility. So every caplet in the model has one Black volatility per expiry, and the market quotes a different volatility at every strike. The model as constructed cannot fit the market at all across strikes: it does not fit the smile badly, it has no smile.

The historical repair was to make the volatility depend on the level of the forward — a displaced diffusion, $d(F_i + \alpha)/(F_i + \alpha) = \sigma_i \, dW_i$, or a CEV exponent. This produces a skew and it is exactly what chapter 9 calls a local volatility model, applied to each forward. So it inherits, precisely, the pathology chapter 10 measures: the skew is glued to the level of rates, the backbone is deterministic, and forward smiles flatten as expiry recedes.

Why not mark each swap rate as SABR?

There is an obvious construction to try, and it repays working through because it is both used and more limited than it looks.

The swaption cube is marked in SABR: for every expiry and tenor, a desk carries four parameters that reproduce that smile. Lemma 15.1 says each co-terminal swap rate is a martingale under its own annuity measure. So declare each one to follow SABR dynamics there. Every swaption then prices by the market's own formula with the market's own parameters, and the calibration is not a fit at all — it is a copy of the marks.

Three things are wrong with concluding that the model is therefore right.

A perfect fit instrument by instrument is not a joint model.

The issue is counting, and it bites on the cube rather than on the co-terminal family. The $n - 1$ co-terminal rates are a coordinate system for a curve with $n - 1$ degrees of freedom — the overall level divides out of every one of them — so giving each its own diffusion is exactly

determined and not over-determined. A desk does not mark $n-1$ rates. It marks a cube, every expiry against every tenor, which is many more rates than the curve has dimensions; and since all of them are smooth functions of the same bonds, Itô expresses the diffusion of each in terms of the others. Most of the cells are therefore consequences rather than choices, and marking every one of them independently is Theorem 13.6 again with more instruments: the marks can be mutually inconsistent, and generically are.

The parameters are constants, so the smile moves wrongly. This is the more serious objection and it is chapter 10's, transplanted. SABR's β — the backbone exponent, not the correlation decay β_ρ of (13.7) — fixes the backbone deterministically, and ρ and ν do not move at all, so the model's smile a year from now is determined by today's marks in a way the market's smile does not respect. For a European swaption that is a second-order complaint, because the price depends on the terminal distribution and the terminal distribution was fitted. For a callable it is first order: the exercise decision is priced off the smile *at the exercise date*, so fitting every smile today and moving them incorrectly is being wrong about the only quantity the construction exists to compute. Fitting the cube exactly buys the vanillas, and the vanillas were already quoted.

And it leaves free the parameter that matters most. A SABR specification per swap rate says nothing whatever about the correlation between different swap rates, or between their volatilities. Those are extra assumptions, unconstrained by any of the marks that were copied — and Calculation 13.8 measured that the correlation is worth percentage points of volatility while the smile fit is worth hundredths. So the construction spends all its structure on the cheap axis and leaves the expensive one to be assumed.

Remark (One further reason the fit is not as exact as it appears)**.** Hagan's SABR formula is an asymptotic expansion, not a solution, and it is known to admit arbitrage at low strikes — the implied density it corresponds to can go negative. So a market marked in SABR is marked with a convention that is not itself arbitrage-free, and a model that reproduces those marks exactly reproduces the arbitrage along with them.

The honest verdict is that this is a good *marking* device and not a model of a callable. Carrying the cube as a field of SABR parameters is an excellent way to interpolate a surface, reprice vanillas quickly, and communicate a smile in four numbers. What it does not supply is either of the two things a Bermudan needs — how the smile evolves, and how the rates move together — and those are precisely the two it leaves unspecified.

Which points at what the fix looks like. Rather than a separate SABR per rate, attach a *common* stochastic volatility process to the whole family, scaling the loadings of every rate together. Then the smile dynamics are specified once and jointly, the correlation structure is explicit rather than implied by a collection of marginals, and the model has as many smile parameters as the market can identify instead of four per cell of a cube. That is the stochastic-volatility market model, and it is what the remark below is about.

The stochastic-volatility market model

Definition 13.9 (Market model with a common variance factor)**.**

$$\begin{aligned}\frac{dF_i(t)}{F_i(t)} &= \mu_i(t)\,dt \;+\; \sqrt{V_t}\,\sigma_i(t)\,dW_t^i,\\ dV_t &= \theta\,(1-V_t)\,dt \;+\; \eta\sqrt{V_t}\,dZ_t, \qquad V_0 = 1,\end{aligned} \tag{13.11}$$

with $dW^i dW^j = \rho_{ij}\,dt$, and μ_i the no-arbitrage drift of Theorem 13.3 in whichever measure is being simulated.

Four things about (13.11) are the whole of its design.

One factor, not one per rate. A single V multiplies every forward's volatility, so the smile dynamics are one specification rather than a field of them. That is the repair to the first objection above: the swap rates can no longer be given mutually inconsistent smiles, because they are not given smiles at all — they inherit one.

The variance is dimensionless. It starts and reverts to one, so $\sigma_i(t)$ keeps its old meaning as the level of volatility and η adds only randomness around it. Setting $\eta = 0$ returns the deterministic model exactly.[4]

The drift is not made harder. The drift is a sum of products of *two* volatilities — it comes from a covariation — so it carries V where each diffusion carries $\sqrt{V}$, and the derivation of Theorem 13.3 goes through with V_t multiplying the same sum. Nothing has to be rederived; a factor is threaded through.

And the split of labour is clean. $\sigma_i(t)$ fits the caplet levels as before, η and θ fit the caplet smile's curvature and how it flattens with expiry, and ρ_{ij} still fits the swaption grid. The instruments that identified the deterministic model identify this one, with the smile parameters answering to the strike dimension nobody was using.

Calculation 13.10 (What the factor is worth, against what the coordinates were worth)**.** The same ten-period structure as Calculation 13.7, every forward at 25%, correlation decay $\beta_\rho = 0.2$, priced across strikes from 70% to 130% of the forward swap rate.

At $\eta = 0$ the smile spans four hundredths of a volatility point, which is Calculation 13.7 again. At $\eta = 0.6$ it spans four tenths, and at $\eta = 1.2$ about one and a half points — stable across seeds and path counts.[5]

So one extra state variable moves the smile by more than an order of magnitude beyond anything the choice of coordinates could reach, and into the range where swaptions actually trade. The at-the-money level drifts down slightly as η grows, from 22.9% to 21.6%, which is the usual convexity of averaging over a random variance and is absorbed by recalibrating σ_i.

A single common factor is a strong assumption.

Every smile moves together. One V cannot decorrelate the two year's smile from the ten year's, and a product sensitive to that — a spread option on volatility, or a callable whose exercise depends

[4] `checked to machine precision`.

[5] `measured`.

on how the front and back of the surface move apart — is priced as though the question did not arise. This is chapter 12's calculation How the components were found, and what a loading is — "d factors give d shapes" — arriving in the volatility dimension rather than the rate one, and the repair is the same: more factors, at the usual cost.

And (13.11) as written has a symmetric smile. The skew has to come from somewhere else — a displaced diffusion or a CEV exponent on each forward, or a correlation between Z and the W^i — and which of those is chosen is exactly chapter 10's backbone question, asked again in a market model. The measurement of chapter 10 applies unchanged: the exponent is observable, and it is not any of the conventional choices.

Remark (Why that matters more here than for a vanilla)**.** Chapter 10 showed that a local volatility model fits today's smile exactly and gets the future smile wrong. For a European option that is a second-order complaint: the price depends on the terminal distribution, which is fitted.

For a callable it is a first-order complaint. The holder's decision at a future exercise date compares the swap against the value of continuing, and the value of continuing is the value of the remaining option — which depends on the volatility surface *at that date*. The forward smile is not a diagnostic here; it is an input to the price. A model that gets it wrong prices the exercise decision wrong, which is the one quantity the whole apparatus of this chapter was built to get right.

So the answer to whether a market model needs stochastic volatility is yes, and the argument is not the usual one about fitting the smile better. It is that a deterministic-volatility model, whether it has a skew or not, has a wrong forward smile, and a callable is priced off the forward smile. The stochastic volatility versions — a variance process multiplying the loadings, in the manner of chapter 11 — are what fix it.

With one ordering caveat, which Calculation 13.8 supplies. The smile is second order in size to the correlation structure, so a model with rich volatility dynamics and a one-factor correlation is worse than the reverse. Get the number of factors and their correlation right first; then the smile. Both, now that compute forces neither to be sacrificed — which is chapter 19's argument arriving at a specific recommendation.

References

- Brace, A., Gatarek, D., & Musiela, M. (1997). The market model of interest rate dynamics. *Mathematical Finance*, 7(2), 127–155.

- Jamshidian, F. (1997). Libor and swap market models and measures. *Finance and Stochastics*, 1(4), 293–330.

- Miltersen, K. R., Sandmann, K., & Sondermann, D. (1997). Closed form solutions for term structure derivatives with log-normal interest rates. *Journal of Finance*, 52(1), 409–430.

- Rebonato, R. (2002). *Modern Pricing of Interest-Rate Derivatives: The LIBOR Market Model and Beyond*. Princeton University Press.

- Andersen, L., & Andreasen, J. (2000). Volatility skews and extensions of the Libor market model. *Applied Mathematical Finance*, 7(1), 1–32.

- Piterbarg, V. (2005). Stochastic volatility model with time-dependent skew. *Applied Mathematical Finance*, 12(2), 147–185.

- Andersen, L., & Piterbarg, V. (2010). *Interest Rate Modeling*. Atlantic Financial Press.

Chapter 14

Solvable Models

Chapter 3 ended on a promise: tractability is a property of the generator. This chapter cashes it. A model is solvable when $\mathcal{L}$ preserves a small family of functions, so that an evolution on an infinite-dimensional space collapses onto a finite one — and the three classical families, exponential-affine, exponential-quadratic and polynomial, are three answers to that one question. We separate that collapse from a different one it is often confused with, being Markov, since a model can have either without the other. We take the Riccati equation apart to see why it is solvable rather than merely quadratic. And we look at the spectral route, which is how physics solves such equations and which works here for exactly the diffusions whose eigenfunctions are the classical orthogonal polynomials. The chapter ends with the recipe read backwards: how to look for a tractable class that nobody has written down.

14.1 Two Collapses, Often Confused

Pricing begins as an infinite-dimensional problem. The state is a path, the expectation is over path space. Every tractable model is the result of two separate reductions that are independent.

	Reduction	**Bought by**
Markov	path space $\to$ a finite state	a state whose future ignores its past
Solvable	a PDE $\to$ finitely many ODEs	$\mathcal{L}$ preserving a small family that the payoff class can be built from

The first turns a functional integral into a partial differential equation in a handful of variables. The second turns that equation into a system one can solve with a pen. They are not the same thing, and much confusion follows from treating them as one.

The second needs a definition, since the whole chapter is named for it and the word is used loosely in the literature — sometimes for a closed-form formula, sometimes for anything faster than simulation.

Definition 14.1 (Solvable)**.** A model is *solvable* for a class of payoffs if the price of any payoff in the class can be computed to arbitrary accuracy by

(i) solving finitely many ordinary differential equations, the number fixed by the model and not by the accuracy demanded; and

(ii) evaluating finitely many one-dimensional integrals.

Three features of definition 14.1 are deliberate.

It does not ask for a closed-form formula. Almost nothing is solvable in that sense — Heston is not, and neither is Vasicek's swaption price without the normal distribution function — and the distinction between an expression built from special functions and one built from a Riccati solution is a fact about which functions have names, not about the model.

It does say "the number fixed by the model". A grid also produces a system of ordinary differential equations, one per node, but the number grows as the grid is refined and grows exponentially in the dimension; refining a grid is buying accuracy with state variables. A solvable model buys accuracy with nothing, because the same two Riccati equations serve every accuracy and every strike.

And it is relative to a *class of payoffs*. Solvability is not a property of a model alone. The same model appears in both columns of the table below depending on what is asked of it, which is why chapter 19 exists.

Definition 14.1 is the general form, framed this way so that it can be set against a grid on the same axis. In a fixed income setting it has a concrete and more demanding reading, and that reading is what practitioners mean by the word.

Definition 14.2 (Solvable, as a rates desk means it)**.** A term structure model is solvable if

(i) the discount bond $P(t,T)$ is an explicit function of the state, for every t and T and at *every* value the state can take; and

(ii) European swaption prices follow in closed form or by one quadrature.

The first clause does the work, and the phrase "at every value the state can take" does all of it. Definition 14.1 asks for a number: today's price of one payoff. Definition 14.2 asks for a *field* — the whole curve, as a function on the state space, available wherever one cares to evaluate it. That is strictly more, and it is what a desk needs, for two reasons.

- *Assembling instruments.* A swap, a cap, a range accrual and an amortising callable are all built from discount factors at many dates. If $P(t,T)$ is a function of the state then so is every one of them, and no separate solution is needed for each.

- *Backward induction.* This is the decisive one. A Bermudan swaption needs, at every exercise date and every node, the value of exercising — which is a swap, and so needs the whole curve at that node. Chapter 12's reconstruction formula, which writes every bond as an explicit exponential in the two state variables, supplies it: that is exactly clause (i) being met, and it is why that chapter's model prices callables without a grid.

So the two definitions separate the two things solvability is wanted for. What makes a calibration cheap is fitting European prices without a grid: definition 14.1 states it for any class of payoffs, and definition 14.2 contains it as clause (ii). What makes an exotic possible is clause (i), which is all the definition for rates adds: evaluate the model anywhere and let backward induction do the rest. A model can have the first and lack the second — Heston does — and the failure is not a matter of degree, since no amount of extra computation turns a transform into a value function.

Structure. Read definition 14.1 against the semigroup of chapter 3. Pricing is evaluating $P_T f$, and $P_T = e^{T\mathcal{L}}$ is the exponential of an operator on an infinite-dimensional space. Nothing can be done with that directly.

A grid makes the space finite by force: replace the functions on $\mathbb{R}^d$ by their values at n^d nodes, and $\mathcal{L}$ by a matrix of that size. The exponential is then computable, and the cost of accuracy is the cost of a finer discretisation.

Solvability makes it finite by *restriction* instead. If some small family is carried into itself by the flow, then on that family P_T is an operator of low dimension exactly — no approximation — and the accuracy is limited only by how well the ODEs are integrated, which is a one-dimensional problem. Both routes reduce an infinite-dimensional exponential to a finite computation. One truncates the space and the other finds a piece of it that is already closed.

Remark (Markov is necessary and nowhere near sufficient)**.** Being Markov buys a great deal — it is the difference between a problem with a differential equation and a problem without one — and it buys nothing about closed forms.

The clearest case is chapter 9's local volatility model. It is perfectly Markov: the state is the spot, one dimension, and the pricing equation is an ordinary parabolic PDE. It has no closed-form solution for a general local volatility function, and nobody expects one. Markov gave the PDE; nothing gave the formula.

The clearest case in the other direction is what chapter 12 had to do. The Heath-Jarrow-Morton framework is not Markov in any finite state at all: the state is the entire forward curve, and the drift condition makes the evolution depend on the whole of it. Quasi-Gaussian models are the response, and they buy a finite Markov state by insisting the volatility structure decay exponentially — a restriction imposed for no reason except that it makes the first reduction possible. Rough volatility, in chapter 10, refuses the same restriction and pays for it by having no finite Markov state and therefore no PDE.

Model	**Markov**	**Solvable**	**What one gets**
Black-Scholes	yes	yes	a formula
Vasicek, Hull-White	yes	yes	bonds and swaptions in closed form
Heston	yes	yes	a characteristic function
Local volatility	yes	no	a PDE, solved on a grid
SABR	yes	no	an asymptotic expansion
Quasi-Gaussian	yes	partly	a low-dimensional grid
General HJM	no	no	a simulation
Rough volatility	no	no	a simulation, and an awkward one

Remark (Which one a difficulty belongs to)**.** Consult the table when a model is being hard to work with, because the two failures have different remedies. A model that is not Markov cannot

be put on a grid at all. It can still be priced by simulating paths, since a payoff that is a function of the path needs no conditioning on the past; what it loses is controlled early exercise, because the continuation value at an exercise date depends on the whole path so far, and a regression on hand-picked summary variables gives a lower bound whose error nothing in the calibration controls, which a dual upper bound can bracket. The fix — if there is one — is to enlarge the state until it is Markov, as chapter 12 does, or, for rough volatility, to approximate the kernel by a sum of exponentials, as in chapter 10's calculation A rough kernel is a sum of exponentials, and how many it takes. A model that is Markov but not solvable is merely slow, and the fix is chapter 19's: a better numerical method, or a machine.

Confusing them wastes time in a predictable direction. One cannot make rough volatility affine, and one does not need a new model class to price a local volatility barrier.

14.2 Invariant Families

Now the second reduction, properly.

Definition 14.3 (Invariant family)**.** A family $\mathcal{F}$ of functions is invariant if the evolution keeps a solution inside it: whenever $u(0,\cdot) \in \mathcal{F}$ and $\partial_t u = \mathcal{L}u$, we have $u(t,\cdot) \in \mathcal{F}$ for every t.

If $\mathcal{F}$ is finite dimensional and invariant, a function that starts inside it stays inside it, and its evolution is described by however many numbers it takes to say *which* member of $\mathcal{F}$ it is. An infinite dimensional flow has collapsed onto a finite dimensional one. That is the whole of solvability. Two examples make this concrete before anything is said in general — one where invariance is the easiest thing in the world to arrange, and one where it is not.

14.2.1 A Subspace: Polynomials

Take $\mathcal{F}$ to be the polynomials of degree at most n — a linear subspace of dimension $n+1$, with the obvious basis $1, x, x^2, \dots, x^n$. A member is specified by its coefficient vector, so "which member" is $n+1$ numbers, and since $\mathcal{L}$ is a linear operator its action on those numbers is a matrix.

Calculation 14.4 (The matrix, for an Ornstein-Uhlenbeck process)**.** Take $\mathcal{L} = -\kappa x\,\partial_x + \frac{1}{2}\sigma^2\partial_{xx}$. Applying it to a single power,

$$\mathcal{L}\,x^n \;=\; -\kappa n\,x^n \;+\; \tfrac{1}{2}\sigma^2 n(n-1)\,x^{n-2}, \tag{14.1}$$

which is a polynomial of degree n: the family is invariant, and visibly so. Writing that out on $1, x, x^2, x^3$ gives

$$\mathcal{L} \;\longleftrightarrow\; \begin{pmatrix} 0 & 0 & \sigma^2 & 0 \\ 0 & -\kappa & 0 & 3\sigma^2 \\ 0 & 0 & -2\kappa & 0 \\ 0 & 0 & 0 & -3\kappa \end{pmatrix},$$

reading columns as inputs. It is triangular, because (14.1) can lower a degree but never raise one — the drift takes one power off and gives one back, the diffusion takes two off. And its diagonal is $-\kappa n$, so those are the generator's eigenvalues on this subspace; they are the same numbers the spectral section below reaches by a different route.

Now use it. Let $m_n(t) = \mathbb{E}[x_t^n]$. Since $\frac{d}{dt}\mathbb{E}[f(x_t)] = \mathbb{E}[\mathcal{L}f(x_t)]$ by Dynkin, taking expectations of (14.1) gives

$$\frac{dm_n}{dt} = -\kappa n\, m_n \;+\; \tfrac{1}{2}\sigma^2 n(n-1)\, m_{n-2}, \tag{14.2}$$

a linear system, triangular, and therefore solvable one rung at a time. The first three rungs:

- $m_0 = 1$, since a probability is a probability.
- $m_1' = -\kappa m_1$, so $m_1 = x_0 e^{-\kappa t}$: the mean decays at κ.
- $m_2' = -2\kappa m_2 + \sigma^2$, so $m_2 = x_0^2 e^{-2\kappa t} + \dfrac{\sigma^2}{2\kappa}\big(1 - e^{-2\kappa t}\big)$, which is the familiar variance plus the squared mean.

Every higher moment follows by the same substitution, with no simulation, no grid, and no distributional assumption — (14.2) never mentions normality. Integrating it numerically and comparing against the exact Gaussian moments confirms both the matrix and the ladder.[1]

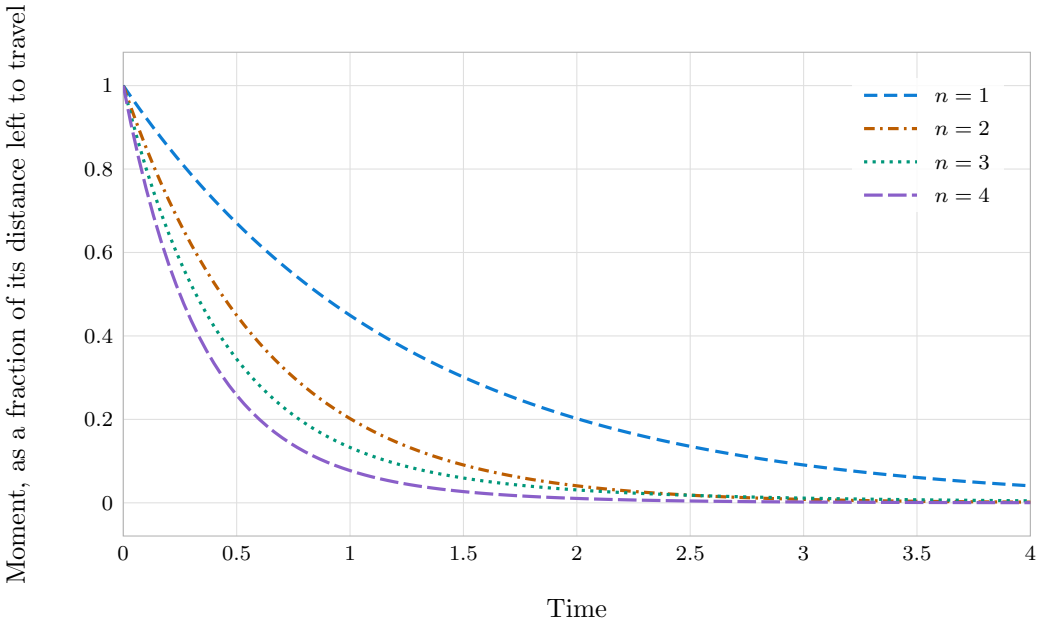

Figure 14.1: Each moment of an Ornstein-Uhlenbeck process relaxing towards its stationary value, as a fraction of the distance it had to travel. The curves are solved from (14.2) and nothing else — no simulation, no grid, no assumption of normality. The ordering is the point: the fourth moment settles first and the first moment last, because rung n of the ladder relaxes at κn, which is the diagonal of the matrix above. Read the other way, the slowest thing about a mean-reverting process is its mean.

Drawn by `polynomial_generator_matrix` and `polynomial_moments`.

[1] `generator::polynomial_moments`.

This is the pattern for a subspace generally: pick a basis, write $\mathcal{L}$ as a matrix, and the flow is $\dot{a} = Ma$ on coefficients, with P_t the matrix exponential e^{tM}. Linear, always, because $\mathcal{L}$ is linear and a subspace is where linearity has somewhere to act.

Remark (Two ways to be invariant)**.** Definition 14.3 is about the flow rather than about $\mathcal{L}$, and the distinction is not pedantry — it is what the rest of this section turns on.

For the polynomials just treated, a linear *subspace*, the two are the same thing. $\mathcal{L}$ maps $\mathcal{F}$ into itself if and only if $P_t = e^{t\mathcal{L}}$ does, because the exponential of an operator preserving a subspace preserves it too. That is why writing $\mathcal{L}$ as a matrix was enough.

The next family is a *manifold*, and there the two part company: the flow condition is genuinely weaker. Take the family of exponentials below: $\mathcal{L}e^{\phi+\psi x}$ is some function of x multiplied by $e^{\phi+\psi x}$, and that product is not an exponential-affine function at all. The family is not preserved by $\mathcal{L}$. What is required instead is that the multiplier be something the *parametrisation* can absorb, and the proposition below makes that exact.

Invariance is the hypothesis of the theorem *Invariance gives solvability* below, so it needs a test that can be run on a formula. For an exponential family the test is tangency, and it reduces to a question about spans.

Proposition 14.5 (Invariance of an exponential family)**.** *Write a family in exponential form with a fixed set of functions in the exponent,*

$$u \;=\; \exp\left(\sum_{i=1}^{n} \theta_i \, g_i(x)\right), \tag{14.3}$$

the θ_i being the parameters and the g_i fixed. The family is invariant under the flow of $\mathcal{L}$ if and only if

$$\frac{\mathcal{L}u}{u} \;\in\; \operatorname{span}\{g_1, \dots, g_n\} \quad \textit{for every member } u,$$

and when it holds, the θ_i' are the coordinates of $\mathcal{L}u/u$ in that span.

Proof. Differentiating in time,

$$\frac{\partial_t u}{u} \;=\; \partial_t \log u \;=\; \sum_{i=1}^{n} \theta_i'(t)\, g_i(x),$$

so the *only* multipliers the parameters are capable of producing are combinations of $g_1, \dots, g_n$. The flow $\partial_t u = \mathcal{L}u$ stays in the family exactly when some choice of θ_i' matches $\mathcal{L}u/u$, that is, when the multiplier lies in the span, and the θ_i' are then its coordinates, which is where the ODEs come from. If it lands outside the span no choice of θ_i' can match it, and the flow leaves the family at once. □

Geometrically this is tangency. Differentiating u along θ_i gives $g_i u$, so the tangent space to the family at u is $u \cdot \operatorname{span}\{g_1, \dots, g_n\}$, and the condition says $\mathcal{L}u \in T_u\mathcal{F}$, the check made for the affine family in the excursion on affine models in chapter 12. A vector field tangent to a surface generates a flow on it, which is why that flow is a set of ordinary differential equations in the coordinates.

For the exponential-affine family the exponent is $\phi \cdot 1 + \psi \cdot x$, so $g_1 = 1$, $g_2 = x$, and the absorbable multipliers are the affine functions of x. That is the whole origin of the word "affine" in this subject.

Example 14.1 (The rule applied to a quadratic exponent)**.** Proposition 14.5 is a calculation and not just a slogan, so run it on $\mathcal{F} = \{e^{q(x)}\}$ with q quadratic — the family behind the quadratic-Gaussian models. Here $g = (1, x, x^2)$, so quadratics are absorbable. Since $\partial_x e^q = q' e^q$ and $\partial_{xx} e^q = (q'' + (q')^2)e^q$,

$$\frac{\mathcal{L}u}{u} \;=\; \mu\, q' \;+\; \tfrac{1}{2} v \left((q')^2 + q'' \right),$$

and with q quadratic, q' is linear and q'' constant. Counting degrees: $\mu q'$ has degree $\deg \mu + 1$ and $v(q')^2$ has degree $\deg v + 2$. Both must be at most two, so

$$\deg \mu \le 1, \qquad \deg v = 0.$$

A quadratic exponent demands a *constant* variance — strictly more than the affine family asks of it, which is why the quadratic-Gaussian class is built on a Gaussian state. Give the variance even a linear term and $v(q')^2$ turns cubic, leaving the span; the multiplier then has a component no choice of ϕ', ψ' or γ' can match.[2]

14.2.2 A Manifold: Exponential-Affine Functions

Now the family that motivated the remark above. Take

$$\mathcal{F} = \Big\{ x \mapsto e^{\phi + \psi x} \;:\; \phi, \psi \in \mathbb{R} \Big\}.$$

It is two dimensional in the sense that two numbers name a member, but it is not a linear subspace: the sum of two exponentials is not an exponential. There is no basis, and no coefficient vector to move. What moves instead is the *label* (ϕ, ψ) — the family is a surface in function space, parametrised by two coordinates, and a flow that preserves the surface is a curve traced out in those coordinates. That is what "the induced flow is on the parameters" means, and the consequence is that nothing forces it to be linear, because (ϕ, ψ) are not coordinates in a basis and $\mathcal{L}$ is not acting on them by matrix multiplication.

Before solving anything, the invariance has to be checked, and by the remark distinguishing the two invariances it is the weaker condition that applies. Applying the generator to a member,

$$\mathcal{L}\, e^{\phi + \psi x} \;=\; \Big[\mu(x)\, \psi + \tfrac{1}{2} v(x)\, \psi^2 \Big]\, e^{\phi + \psi x}, \qquad v = \sigma^2, \tag{14.4}$$

using $\partial_x e^{\phi+\psi x} = \psi\, e^{\phi+\psi x}$ and $\partial_{xx} e^{\phi+\psi x} = \psi^2 e^{\phi+\psi x}$. The bracket is a function of x, so the right-hand side is *not* an exponential-affine function and $\mathcal{L}$ does not map the family into itself. By proposition 14.5 what matters is whether the bracket lies in $\operatorname{span}\{1, x\}$, and it does exactly when μ and v are affine. So the family is invariant in the sense of definition 14.3 precisely for affine drift and affine variance, which is the definition of an affine model, arrived at rather than assumed.

That is the structural claim. Chapter 12 already derived a Riccati equation for a particular model. The point here is not to derive it again but to say where its *shape* comes from: not from the model, and not from any nonlinearity in $\mathcal{L}$, but from the family being a manifold whose parameters sit in an exponent.

[2] both halves measured.

Calculation 14.6 (Where the Riccati equation comes from)**.** Write $u(t,x) = e^{\phi(t)+\psi(t)x}$ and ask that it satisfy $\partial_t u = \mathcal{L}u$. Dividing both sides by u makes the left

$$\frac{\partial_t u}{u} \;=\; \phi'(t) + \psi'(t)\,x,$$

and the right is (14.4).

Put $\mu(x) = a + bx$ and $v(x) = c + dx$, expand, and match the two powers of x separately:

$$\phi' = a\,\psi + \tfrac{1}{2}c\,\psi^2, \tag{14.5}$$

$$\psi' = b\,\psi + \tfrac{1}{2}d\,\psi^2. \tag{14.6}$$

Equation (14.6) is a Riccati equation — quadratic in the unknown — and (14.5) is then an integral, since its right side is known once ψ is. Two ordinary differential equations in place of a partial one.

Checking (14.4) against a numerically differenced generator confirms both the identity and that its right side really is affine in x, which is the step that lets the coefficients be read off one at a time.[3]

So "invariant" is one condition on $\mathcal{L}$ for a subspace — that it map $\mathcal{F}$ into itself — and a different, weaker one for a manifold, that $\mathcal{L}u/u$ fall in the exponent's span. Both examples have so far shown the condition *necessary*: if the flow stays in the family, the multiplier has nowhere else to be. That is the easy half and not the useful one. What a model needs is the converse — that the condition *delivers* a solution — and it does not follow for free, because an ODE system is not obviously solvable and a function satisfying a partial differential equation is not obviously the price of anything.

Theorem 14.7 (Invariance gives solvability)**.** *Let* $\mathcal{F} = \big\{\exp\big(\sum_{i=1}^{n}\theta_i g_i\big)\big\}$ *with the* g_i *linearly independent, and suppose that for every member*

$$\frac{\mathcal{L}u}{u} \;\in\; \operatorname{span}\{g_1,\dots,g_n\}, \tag{14.7}$$

so that the coordinates of the multiplier define a map $\theta \mapsto V(\theta) \in \mathbb{R}^n$*, locally Lipschitz. Then for each* $\theta(0)$ *there is a* $T^* > 0$ *and a unique*

$$\theta'(t) = V\big(\theta(t)\big), \qquad t \in [0, T^*), \tag{14.8}$$

and on that interval

$$\mathbb{E}\Big[\exp\Big(\textstyle\sum_i \theta_i(0)\,g_i(X_T)\Big)\;\Big|\;X_0 = x\Big] \;=\; \exp\Big(\textstyle\sum_i \theta_i(T)\,g_i(x)\Big), \tag{14.9}$$

provided the left side is finite. The model is therefore solvable, in the sense of definition 14.1, for every payoff whose transform is a member of $\mathcal{F}$*: the* n *equations of* (14.8) *are the ODEs, and* n *does not depend on the accuracy wanted.*

Note that (14.9) *delivers more than definition 14.1 asks. Its right-hand side is a function of* x *and not a number — the expectation from* any *starting state, at once, for the price of the same* n *equations. So an invariant family meets clause (i) of definition 14.2 as well, and that is why the affine construction is what rates modelling is built on rather than merely a way of pricing one European option.*

[3] the test.

Proof. Three steps.

Existence and uniqueness of the parameter flow. Condition (14.7) says the multiplier lies in the span; since the g_i are independent, its coordinates in that span are determined, which is what makes V a well-defined function of θ rather than a relation. With V locally Lipschitz, Picard-Lindelöf gives a unique solution of (14.8) on some interval around zero. It need not be all of $[0,\infty)$: V is typically quadratic, and a quadratic vector field can reach infinity in finite time. That is the explosion the Riccati section returns to, and T^* is where it happens.

The candidate satisfies the equation. Define $u(t,x) = \exp(\sum_i \theta_i(t) g_i(x))$ using that solution. For each t this is a member of $\mathcal{F}$, so (14.7) applies to it at every time; that it is also the value function is the content of the next step. Then

$$\frac{\partial_t u}{u} = \sum_i \theta_i'(t)\, g_i = \sum_i V_i(\theta)\, g_i = \frac{\mathcal{L}u}{u},$$

the last equality by the definition of V as the coordinates of $\mathcal{L}u/u$. So $\partial_t u = \mathcal{L}u$, with $u(0,\cdot) = \exp(\sum_i \theta_i(0) g_i)$. That initial condition is the payoff being priced, an exponential of the state. It is called a transform because a payoff of ordinary shape, a call for instance, is a superposition of such exponentials by Fourier inversion: the ODEs price each exponential, and one integral over the parameters prices the call. This step is pure bookkeeping — it is the reason the construction was set up this way — and it is also where the argument would break if the multiplier had a component outside the span, since then no choice of θ' could match it.

The candidate is the expectation. This is the step with content, and it does not follow from solving the equation: many functions satisfy a parabolic equation and only one is the conditional expectation. Here t in u is the time remaining to maturity, so the value function in calendar time is u with time reversed. Fix T and set $m(t,x) = u(T-t,x)$, the value at calendar time t of a claim maturing at T: $m(T,\cdot) = u(0,\cdot)$ is the payoff, and $m(0,x) = u(T,x)$ is the value from state x today. Since $\partial_t m = -\partial_t u$, Itô's lemma, legitimate as m is smooth in x and C^1 in t, gives

$$d\, m(t,X_t) = \Big(\partial_t m + \mathcal{L}m\Big)dt + \partial_x m \cdot \sigma(X_t)\, dW_t = \partial_x m \cdot \sigma(X_t)\, dW_t,$$

the drift $\partial_t m + \mathcal{L}m = -\partial_t u + \mathcal{L}u$ vanishing by the previous step. So $m(t,X_t)$ is a local martingale. If it is a *true* martingale, then $\mathbb{E}[m(T,X_T)] = m(0,X_0)$, which is $\mathbb{E}[u(0,X_T)] = u(T,X_0)$, that is (14.9), since $u(0,X_T)$ is the payoff. Integrability of the left side of (14.9) is what upgrades the local martingale to a martingale, by dominated convergence along a localising sequence. □

Structure (Move the point or move the label)**.** Read equation (14.9) slowly, because the two sides do different things.

On the left the state moves. It starts at x, runs to X_T, and the label $\theta(0)$ is held fixed while an expectation is taken over every path it might have taken. On the right the state does not move at all — it is still x — and the *label* has run from $\theta(0)$ to $\theta(T)$ instead. The same number is obtained by moving the point or by moving the label, and never by moving both.

Which of the two is cheap is the point. Moving the point means an expectation over paths: a partial differential equation in as many dimensions as the state has, or a simulation. Moving the label means the n ordinary differential equations of (14.8), and n is fixed by the family rather than by the dimension of the state or the length of the horizon.

The duality is not a coincidence of the algebra; it is what invariance says. The semigroup $P_T = e^{T\mathcal{L}}$ acts on functions, and $\mathcal{F}$ is invariant, so P_T maps $\mathcal{F}$ into itself. A map of a manifold into itself, read in a chart, is a map of coordinates — and the coordinates here are the θ. So (14.8) is not an approximation to the semigroup. It *is* the semigroup, written in the only coordinates the invariant family has. The remark above on the two ways to be invariant is what licenses that reading: the vector field has to be tangent to the family, or there would be no induced motion on the coordinates to write down.

Notice too that the label runs the opposite way to the state. It is $\theta(0)$ that sits inside the expectation, against the terminal state, and $\theta(T)$ that sits outside, against the initial one. That reversal is the signature of a pairing: the g_i pair a point with a label, and moving one of the two is the same as moving the other backwards. It is the same reversal that separates the forward equation, which pushes densities from today into the future, from the backward equation, which pulls payoffs from the future back to today — and this theorem is the backward equation with the function replaced by its coordinates.

The manoeuvre generalises well beyond this chapter. Substituting an exponential ansatz and finding that a linear equation for the function has become a nonlinear equation for the exponent is the pattern shared by the method of characteristics, by Hamilton-Jacobi theory, and by the large deviation asymptotics of these same models — and, in chapter 12, by the Riccati system that appeared there for what looked like an unrelated reason. In each the exponent's coefficients are coordinates dual to the state, and solving the problem means flowing them rather than the state.

Remark (Where the proviso bites)**.** The integrability proviso is not decoration. In an affine model the transform $\mathbb{E}[e^{\psi_0 X_T}]$ is finite only for ψ_0 in a strip, and outside it the moment does not exist — which is the same statement as the Riccati solution exploding before T. So the three failure modes of theorem 14.7 line up: the ODE reaching infinity, the moment being infinite, and the local martingale not being a martingale are one phenomenon, not three.

It matters in practice because the strip is where the damping parameter of a Fourier inversion has to live. A transform method evaluates the characteristic function at a complex argument, and choosing that argument outside the strip returns a number the model does not have. Implementations report this as an overflow in the Riccati solver; it is really a statement about which moments exist.

14.2.3 What Each One Actually Gives You

Both calculations end with an ordinary differential equation, but what one does with the answer differs, and the two entries in the table below are opaque without saying how.

Remark (The transform, and inverting it)**.** Solving (14.5) and (14.6) out to T, starting from $\phi(0) = 0$ and $\psi(0) = \psi_0$, gives

$$\mathbb{E}\big[e^{\psi_0 x_T}\big] \;=\; e^{\phi(T)+\psi(T)x_0}, \tag{14.10}$$

since $u(t,x) = \mathbb{E}[e^{\psi_0 x_T} \mid x_t = x]$ is the solution of the backward equation with that terminal condition. Setting $\psi_0 = iw$ makes the left side the *characteristic function* of x_T, and that is what "transform" means here: not a rearrangement of the problem but a specific function of one variable w, delivered in closed form by two ODEs, one w at a time.

A characteristic function determines a distribution, so it determines every European price. Recovering the price is an integral over w — Fourier *inversion*.

Calculation 14.8 (Inverting the transform)**.** The damping trick below is Carr and Madan's (1999) — not to be confused with their 2001 result behind chapter 15's static replication, an unrelated construction that happens to share authors.

Write $s_T = \ln S_T$, let $k = \ln K$ be the log-strike, and take zero rates so that

$$C(k) = \mathbb{E}\big[(e^{s_T} - e^k)^+\big].$$

The natural move is to Fourier transform in k, and it fails immediately: as $k \to -\infty$ the call price tends to S_0, not to zero, so C is not integrable and has no transform.

Damp it. Multiply by $e^{\alpha k}$ for some $\alpha > 0$. The left tail is now killed by the exponential, and the right tail was already fine, so $c(k) = e^{\alpha k}C(k)$ is integrable and

$$\zeta(w) = \int_{-\infty}^{\infty} e^{iwk}\, c(k)\, dk = \int_{-\infty}^{\infty} e^{zk}\, C(k)\, dk, \qquad z = \alpha + iw,$$

exists.

Exchange the order of integration. Writing q for the density of s_T and using that the payoff vanishes for $k > s$,

$$\zeta(w) = \int_{-\infty}^{\infty} q(s) \int_{-\infty}^{s} \big(e^s - e^k\big) e^{zk}\, dk\, ds.$$

The inner integral is elementary, and needs $\operatorname{Re} z = \alpha > 0$ for its lower limit to converge:

$$\int_{-\infty}^{s} e^s e^{zk} dk - \int_{-\infty}^{s} e^{(z+1)k} dk = e^{(z+1)s}\left(\frac{1}{z} - \frac{1}{z+1}\right) = \frac{e^{(z+1)s}}{z(z+1)}.$$

Recognise what is left. The outer integral is now $\int q(s) e^{(z+1)s} ds = \mathbb{E}[e^{(z+1)s_T}]$, which is the characteristic function at the complex argument $-i(z+1)$. So

$$\zeta(w) \;=\; \frac{\varphi\big(w - i(\alpha+1)\big)}{(\alpha + iw)(\alpha + 1 + iw)}, \tag{14.11}$$

with $\varphi(u) = \mathbb{E}[e^{ius_T}]$. Everything on the right is available from the Riccati equations, one w at a time.

Invert. Finally $c(k) = \frac{1}{2\pi}\int e^{-iwk}\zeta(w)\, dw$, and since C is real, $\zeta(-w) = \overline{\zeta(w)}$ and the two halves of the line fold together:

$$C(k) \;=\; \frac{e^{-\alpha k}}{\pi} \int_0^{\infty} \operatorname{Re}\Big[e^{-iwk}\, \zeta(w)\Big]\, dw. \tag{14.12}$$

One integral per strike, with an integrand in closed form. Applied to a lognormal, whose characteristic function is immediate, it returns the Black-Scholes price to ten decimals.[4]

[4] `checked here`, against a formula the routine knows nothing about.

Remark (What α has to satisfy, and why it is the same condition as before)**.** Two constraints have appeared, and the second is the interesting one.

The damping must be strictly positive, $\alpha > 0$, or the inner integral above diverges. Equivalently: at $\alpha = 0$ the denominator of (14.11) vanishes at $w = 0$, so the undamped transform has a pole at the origin. That pole is the statement that a call price is not integrable in log-strike.[5]

The numerator of (14.11) evaluates φ at $w - i(\alpha + 1)$, which is to say it evaluates

$$\mathbb{E}\big[S_T^{1+\alpha}\big].$$

So α is admissible only if the moment of order $1 + \alpha$ exists — and that is exactly the proviso of theorem 14.7. There the condition appeared as the integrability that upgrades a local martingale; here it appears as the strip in which a damping parameter may be chosen. They are one condition.

The practical consequence is that in a model whose moments fail above some order — Heston with a large enough correlation, or any model with a heavy right tail — choosing α beyond that order does not fail loudly. The Riccati solver returns something, the integral evaluates, and the answer is not a price.[6] Within the strip, the answer is insensitive to α, as it must be for a device that is not part of the problem.[7]

Remark (Moments, and what they will not give you)**.** The polynomial side delivers something more elementary and more limited. Equation (14.2) gives $\mathbb{E}[x_T^n]$ for as many n as wanted, exactly — the ladder is a finite triangular linear system, so its solution is a finite combination of exponentials in T with no discretisation error and no sampling error — and cheaply, since an $(n + 1)$-dimensional matrix exponential is the whole of the work.

What it does not give is a price. Moments do not determine a distribution in any usable numerical sense: reconstructing a density from its moments is severely ill-conditioned, and the error in a price built that way is not controlled by the error in the moments. So a polynomial family is the right tool when the moments *are* the answer — variance swaps and other quadratic payoffs, method-of-moments estimation, or the covariance structure a hedge ratio needs — and the wrong one when a strike is involved.

That asymmetry is the honest reason affine models are the ones on trading desks and polynomial models are not, despite the polynomial condition being the weaker of the two.

14.2.4 The Distinction, and What It Decides

The two calculations differ in one structural respect and it accounts for everything else.

Structure. A linear *subspace* has a basis, a member is a coefficient vector, and a linear operator acts on coefficient vectors by a matrix. So the induced flow is $\dot{a} = Ma$ and is linear, necessarily and always.

A *manifold* has no basis. A member is named by parameters, and those parameters enter the function nonlinearly — here they sit in an exponent. Applying $\mathcal{L}$ therefore does not act on them

[5] `the pole is visible in the scaling`: undamped, shrinking w by a thousand grows the transform by a thousand; damped, it does not grow at all.

[6] `demonstrated by capping the moments by hand`.

[7] `checked across a range`.

linearly; the second derivative brings ψ down twice, and the induced flow acquires a ψ^2. It is a vector field on the parameter space, and its degree is inherited from how the parameters were used to build the family, not from any nonlinearity in $\mathcal{L}$ — which remains a perfectly linear operator throughout.

So the shape of the family fixes the shape of the equations. Subspaces give linear systems; the exponential manifold gives Riccati. Neither is an accident and neither is a fact about the model, only about the family one chose to preserve.

Family	**Shape**	**Equations**	**What it computes**
$e^{\phi+\psi x}$	manifold	Riccati	the characteristic function, and so any European price, by one Fourier integral per strike
$e^{\phi+\psi x+x^\top C x}$	manifold	matrix Riccati	the same, with a quadratic state
polynomials of degree $\le n$	subspace	linear	every moment up to n, in closed form; no prices without one

14.3 Reading the Recipe Backwards

The definition can be used as a search rather than as a description, and this is how one would look for a class that is not in the table.

Calculation 14.9 (What preserving polynomials forces)**.** Suppose we want $\mathcal{L}$ to map polynomials of degree at most n into polynomials of degree at most n, for every n. Apply it to x^n:

$$\mathcal{L}x^n = n\,\mu(x)\,x^{n-1} + \tfrac{1}{2}n(n-1)\,\sigma^2(x)\,x^{n-2}.$$

The first term has degree $n-1+\deg\mu$ and the second $n-2+\deg\sigma^2$. For both to be at most n,

$$\deg\mu \le 1, \qquad \deg\sigma^2 \le 2.$$

So the drift must be affine and the variance quadratic — and that is the definition of a polynomial process, obtained rather than assumed.

Remark (The same question, asked of other families)**.** That calculation is a template. Choose the family one wants preserved, apply $\mathcal{L}$ to a general member, and demand the result stay inside. What comes back is a condition on μ and σ, and it is usually restrictive enough to name a model class.

Asking it of $e^{\psi x}$ gives affine drift and affine variance — chapter 12's structure note is that calculation. Asking it of $e^{\psi x+cx^2}$ gives the quadratic-Gaussian class. Asking it of polynomials gives the above. These are not three discoveries; they are one question asked three times.

It also says when to stop. If the condition that comes back is that μ and σ are constant, the family was too small to be interesting. If nothing comes back, it was too large.

14.4 Why the Riccati Equation Is Solvable

Chapter 12 derived $\psi' = \beta\psi + \frac{1}{2}\alpha\psi^2$ and solved it in the Vasicek case by noticing that the quadratic term vanishes. The general case is solvable too because it is the reason Riccati equations are solvable anywhere in mathematics.

Calculation 14.10 (A Riccati equation is a linear system in disguise)**.** Substitute $\psi = -\dfrac{2}{\alpha}\dfrac{w'}{w}$. Then

$$\psi' = -\frac{2}{\alpha}\left(\frac{w''}{w} - \frac{(w')^2}{w^2}\right) = -\frac{2}{\alpha}\frac{w''}{w} + \frac{\alpha}{2}\psi^2,$$

and substituting into the Riccati equation, the quadratic terms cancel identically, leaving

$$w'' + \beta w' = 0, \tag{14.13}$$

a *linear* second order equation with constant coefficients.

Structure (What the substitution means)**.** The cancellation looks like luck and is not. A Riccati equation is what a linear flow looks like after projection.

Take the linear system

$$\begin{pmatrix} u \\ v \end{pmatrix}' = M \begin{pmatrix} u \\ v \end{pmatrix}, \qquad M = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

and follow only the ratio $\psi = v/u$. The quotient rule gives two terms,

$$\psi' \;=\; \frac{v'u - vu'}{u^2} \;=\; \underbrace{\frac{v'}{u}}_{\text{numerator moves}} \;-\; \underbrace{\psi\,\frac{u'}{u}}_{\text{denominator moves}},$$

and both fractions become affine in ψ once the linear dynamics are divided through by u: $v'/u = c + d\psi$ and $u'/u = a + b\psi$. Substituting,

$$\psi' \;=\; (c + d\psi) \;-\; \psi\,(a + b\psi) \;=\; c + (d - a)\,\psi - b\,\psi^2. \tag{14.14}$$

There is the quadratic, and it comes from one place: the second term, in which the ratio multiplies its own denominator's growth rate. Were the denominator held fixed, $u' = 0$, the equation would be affine and there would be nothing to solve. The nonlinearity is the price of the denominator being alive.[8]

So ψ satisfies a Riccati equation — and conversely, every Riccati equation arises this way, since (14.14) can be matched to any given coefficients. The substitution above is that projection run backwards: it lifts a nonlinear flow on a line to a linear flow on a plane, where it can be solved by exponentiating a matrix.

Two consequences follow immediately. Riccati equations are solvable exactly when the linear system above them is, which is why affine models have closed forms at all. And a Riccati solution can blow up in finite time even though the linear system never does — the ratio v/u runs to infinity when u passes through zero, while nothing has happened upstairs. That is not a numerical failure

[8] the test that checks (14.14) against an integrated linear flow.

and cannot be fixed by a smaller step; it is a feature of the projection, and it is why affine model calibrations sometimes report an explosion at a maturity for no financial reason. The cleanest illustration is a rotation, $M = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$: the pair (u, v) walks around the unit circle forever, perfectly bounded, while $\psi = \tan t$ runs to infinity at $t = \pi/2$ and comes back from the other side.[9]

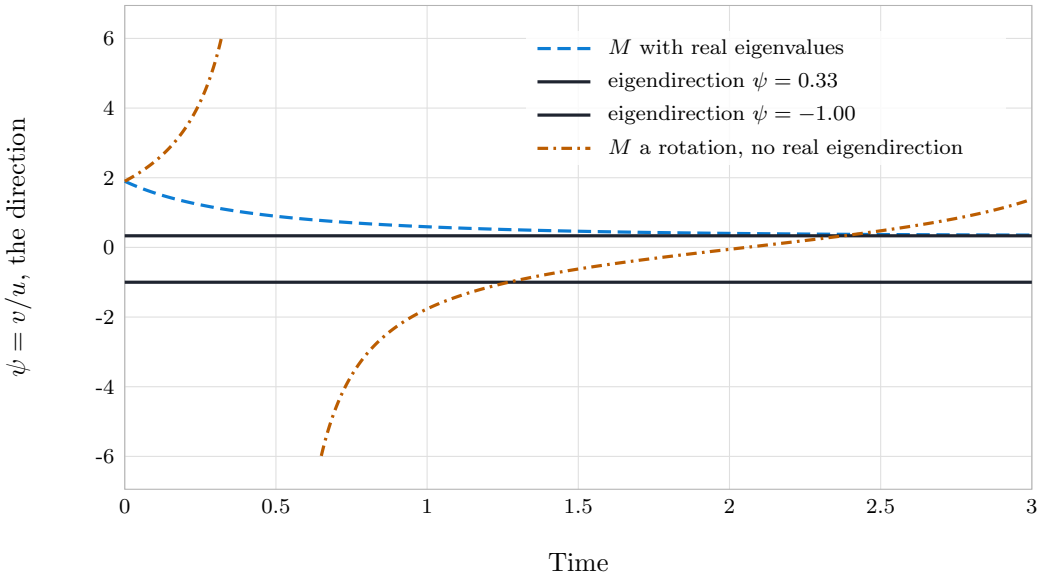

Figure 14.2: The Riccati flow, plotted as the direction $\psi = v/u$ of a linear flow in the plane, for two matrices from the same starting direction. Where M has real eigenvalues the two horizontal lines are the corresponding eigendirections, and the trajectory is trapped between them and settles on one — a fixed point of the projected flow is a direction the linear flow preserves. Where M is a rotation there is no real eigendirection, nothing traps the trajectory, and it runs off the top of the chart and re-enters from the bottom, passing through infinity at $t = \pi/2 - \arctan 1.9 \approx 0.49$: that is the blow-up, and the figure shows what it is, namely the direction passing through $u = 0$ where the coordinate ψ stops covering the line. Nothing happens to the flow there. The curve is computed in the angle $\theta = \arctan \psi$, which covers the whole line and in which the field is bounded, so the integrator passes the blow-up without noticing — one working in ψ could not, and would stop at the edge.

Drawn by `the_ratio_of_a_linear_flow_obeys_a_riccati_equation`.

Remark (The same construction as a blow-up)**.** A reader who has met algebraic geometry will recognise this projection, and the resemblance is exact rather than atmospheric.

The linear flow lives on the plane and vanishes at the origin, so the origin is a singular point of the family of trajectories — every trajectory approaches or leaves it, and no single direction is defined there. *Blowing up* the origin is the standard repair: replace that one point by the projective line $\mathbb{P}^1$ of directions through it. A point of $\mathbb{P}^1$ is a line through the origin, which is to say a value of the ratio $\psi = v/u$, together with the one direction $u = 0$. The Riccati equation is exactly the flow induced on that line.

[9] `measured here`.

Two things follow, and both are visible in what the equation does.

The singularity is resolved into eigendirections. The trajectories' bad point at the origin becomes finitely many bad points on the exceptional line, and they are the fixed points of the Riccati flow, $c + (d-a)\psi - b\psi^2 = 0$. Those are the eigendirections of M: substituting $\psi = (\lambda - a)/b$ turns the quadratic into the characteristic polynomial $(\lambda - a)(\lambda - d) - bc$, so it vanishes exactly when λ is an eigenvalue.[10] One badly behaved point has become two simple ones whose structure can be read off, which is what resolving a singularity buys.

The solution map is a Möbius transformation. A linear map of the plane descends to a fractional linear map of the line of ratios, so with $e^{TM} = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$,

$$\psi(T) = \frac{C + D\,\psi_0}{A + B\,\psi_0}. \tag{14.15}$$

That is a strong statement about a nonlinear differential equation, and it has a testable consequence: Möbius maps are precisely those preserving the cross-ratio, so four solutions started from four different points keep their cross-ratio for all time.[11]

The blow-up in finite time is then a chart artefact rather than an event. $\mathbb{P}^1$ is compact and the flow on it is perfectly well behaved; the coordinate $\psi = v/u$ merely fails to cover the direction $u = 0$, so a trajectory passing through that direction leaves the chart at $+\infty$ and re-enters at $-\infty$.

Remark (Where physics met finance)**.** The same substitution turns the Riccati equation of a stochastic control problem into a Schrödinger equation, and it is the standard route in both directions. The historical route by which affine term structure models were found runs through this correspondence rather than through finance: the equations were recognised as ones already solved elsewhere.

This is the general lesson of the chapter and the reason for the emphasis. New model classes have rarely been invented by writing down dynamics and hoping. They have been found by noticing that a tractable structure from another subject can be mapped onto a financial one, and the mapping is almost always at the level of the generator.

14.5 The Spectral Route

There is a second way to solve $\partial u/\partial t = \mathcal{L}u$, and it is the one a physicist reaches for first: diagonalise the operator.

If there are functions φ_n and numbers λ_n with

$$\mathcal{L}\varphi_n = -\lambda_n \varphi_n, \tag{14.16}$$

then expanding an initial condition in the φ_n reduces the evolution to scalar multiplication:

$$P_t f = \sum_n c_n e^{-\lambda_n t}\varphi_n, \qquad f = \sum_n c_n \varphi_n. \tag{14.17}$$

[10] checked here.

[11] measured, and it does.

Each mode decays at its own rate and nothing mixes. This is an invariant family again — each φ_n spans a one-dimensional one — but arranged so the whole operator is diagonal at once rather than merely block-finite.

The expansion needs the eigenfunctions to form a basis, and that needs $\mathcal{L}$ to be self-adjoint. Not in the plain pairing of chapter 3, though — in a weighted one.

Definition 14.11 ($L^2(\pi)$)**.** Let π be the stationary density. The space $L^2(\pi)$ is the functions square integrable against it, with inner product

$$\langle f, g\rangle_\pi = \int f(x)\, g(x)\, \pi(x)\, dx. \tag{14.18}$$

The weighting says that a function's size is measured where the process actually spends its time. For an Ornstein-Uhlenbeck process with $\pi = N(0, \sigma^2/2\kappa)$, a function that grows wildly at $x = 10$ standard deviations is small in $L^2(\pi)$, because the process is never there. In the flat pairing it would be enormous. The financial reading is direct: what matters about a payoff is its behaviour over the states the model visits.

Definition 14.12 (Reversible)**.** A stationary diffusion is reversible if its law is unchanged by running time backwards — equivalently, if the probability flowing from x to y in time t equals that flowing from y to x, $\pi(x)\, p_t(x, y) = \pi(y)\, p_t(y, x)$.

Concretely: film the process in stationarity, show the film to someone, and ask which way it is running. If they cannot tell, the process is reversible. A mean-reverting rate wandering around its long-run level passes that test; it looks the same either way. The theorem is that $\mathcal{L}$ is self-adjoint in $L^2(\pi)$ exactly when the process passes it.

Calculation 14.13 (Why one dimension is free)**.** A stationary density carries no net probability flux. In one dimension there is nowhere for a flux to circulate to, so it must vanish *pointwise*, which forces $\mu\pi = \frac{1}{2}(\sigma^2\pi)'$ and therefore lets the generator be rewritten as a divergence:

$$\mathcal{L}f = \frac{1}{\pi}\left(\tfrac{1}{2}\sigma^2\pi\, f'\right)'. \tag{14.19}$$

Now pair it. The π in (14.18) cancels the $1/\pi$ in (14.19), and one integration by parts gives

$$\langle \mathcal{L}f, g\rangle_\pi = \int \left(\tfrac{1}{2}\sigma^2\pi f'\right)' g\, dx = -\int \tfrac{1}{2}\sigma^2\pi\, f'\, g'\, dx,$$

which is symmetric in f and g on its face. So $\langle \mathcal{L}f, g\rangle_\pi = \langle f, \mathcal{L}g\rangle_\pi$ for *every* scalar diffusion, whatever its drift and volatility — not merely for the classical ones.[12]

Example 14.2 (A two-dimensional process that fails)**.** Above one dimension probability can circulate, and then it generally does. Take a linear model in the plane,

$$dZ_t = -AZ_t\, dt + dW_t, \qquad A = \begin{pmatrix} 1 & \theta \\ -\theta & 1 \end{pmatrix},$$

[12] checked numerically for an Ornstein-Uhlenbeck process, a double well and a state-dependent volatility.

whose antisymmetric part is a rotation. Its stationary covariance solves $A\Sigma + \Sigma A^\top = I$, and the rotation drops out of that equation entirely: $\Sigma = \frac{1}{2}I$, the same isotropic Gaussian a symmetric A would give. A snapshot of this process is indistinguishable from a reversible one.

The film is not. The process circulates around the origin at rate θ, so running it backwards makes it circulate the other way, and one can tell. Its generator is correspondingly not self-adjoint: on linear functions $\mathcal{L}$ acts as $-A^\top$, so

$$\langle \mathcal{L}f, g\rangle_\pi = -\alpha^\top A\Sigma\beta, \qquad \langle f, \mathcal{L}g\rangle_\pi = -\alpha^\top \Sigma A^\top \beta,$$

and these agree for all α, β only when $A\Sigma$ is symmetric, which the swirl prevents. The spectrum moves off the real line to match — the eigenvalues of $-A$ are $-1 \mp i\theta$ — so there is no real orthogonal expansion to be had.[13]

So the spectral method is a one-dimensional tool in practice, and the reason is structural rather than technical: reversibility is nearly automatic on a line and generically false in a plane. Most multi-factor models of interest have a swirl of exactly this kind, because correlated factors with different mean reversion rates produce one. That is the honest boundary of the technique, and it is the same shape as the boundary of every other method here: a structural condition, cheap in low dimension and expensive above it.

Example 14.3 (The three classical cases)**.** Where the method works completely, the eigenfunctions are the classical orthogonal polynomials, and the correspondence is not a coincidence but the polynomial invariance of the previous section seen through a self-adjoint lens:

Process	**Stationary law**	**Eigenfunctions**	**Eigenvalues**
Ornstein-Uhlenbeck	Gaussian	Hermite	$-n\kappa$
Square root (CIR)	Gamma	Laguerre	$-n\kappa$
Jacobi	Beta	Jacobi	$-n(\cdots)$

These are the Pearson diffusions, and they are the intersection of the two ideas in this chapter: their generators preserve polynomials, and they are reversible, so the polynomial invariant subspace can be diagonalised rather than merely triangularised.

Calculation 14.14 (The Ornstein-Uhlenbeck case, checked)**.** For $dX = -\kappa X\,dt + \sigma\,dW$ with stationary variance $s^2 = \sigma^2/2\kappa$, the eigenfunctions are the Hermite polynomials $He_n(x/s)$ with $\mathcal{L}He_n = -n\kappa\,He_n$. The case $n = 2$ shows the mechanism: with $f = (x/s)^2 - 1$,

$$\mathcal{L}f = -\kappa x \cdot \frac{2x}{s^2} + \frac{1}{2}\sigma^2 \cdot \frac{2}{s^2} = -\frac{2\kappa x^2}{s^2} + 2\kappa = -2\kappa f,$$

using $\sigma^2/s^2 = 2\kappa$. The drift term lowers the degree and the diffusion term supplies exactly the constant needed to close the polynomial back on itself.

`quant/src/generator.rs` verifies this for the first five eigenfunctions on a grid, and then verifies the expansion (14.17) end to end against the exact Gaussian transition law integrated numerically — two routes sharing nothing but the definition of the process.

[13] `measured, including that the stationary covariance is unchanged.`

Remark (What the spectrum is good for)**.** Three things, in increasing order of how much they matter.

The expansion is cheap across maturities. The eigenfunctions and coefficients do not depend on t, so once computed, every maturity costs one exponential per mode. A model that must be re-solved for each expiry becomes one that is solved once.

It gives an honest error. Truncating after N modes leaves an error bounded by the next eigenvalue's decay, which is a statement one can make in advance — unlike a Monte Carlo standard error, which is a statement about the run one happened to do.

And the spectrum has a financial reading. The gap between the first two eigenvalues is the rate at which the process forgets where it started, so for an Ornstein-Uhlenbeck process the spectral gap *is* the mean reversion rate and $\ln 2/\lambda_1$ is the half-life. Chapter 20's calculation Where the bias comes from and what it scales with spends a section on estimating that number badly. In this language it is estimating a spectral gap from a finite sample, and the difficulty is the usual one: the slowest mode is the one a short window sees least of.

Remark (The gap is an asymptotic rate)**.** A deviation from stationarity is a sum over modes,

$$P_t f - \mathbb{E}_\pi[f] \;=\; \sum_{n\geq 1} c_n\, e^{-\lambda_n t}\, \varphi_n,$$

and every term of it is decaying, the higher ones faster. The effective decay rate at time t is $-\frac{d}{dt}\ln$ of the sum, which is

$$\frac{\sum_n \lambda_n\, c_n e^{-\lambda_n t}}{\sum_n c_n e^{-\lambda_n t}},$$

a weighted average of the λ_n. When the weights c_n share a sign, as they do for a deviation started by a positive bump, the average lies above λ_1, so the total falls off *more quickly* than $e^{-\lambda_1 t}$ at first and only settles to that rate once the higher modes have spent themselves. The gap is the rate the decay approaches, not the rate at which the first halving happens; the first halving is quicker. With weights of mixed sign the early rate can sit on either side of λ_1, and the gap is still the asymptotic rate.

That distinction has teeth when a half-life is estimated from data. A sample of a given length contains the fast early relaxation as well as the slow tail, and fitting one exponential to the whole of it splits the difference — returning, when the weights share a sign, a rate above λ_1, which is to say overstating the mean reversion. It is a second and independent route to the bias chapter 20's calculation Where the bias comes from and what it scales with measures, arriving from the spectrum rather than from the estimator.

14.6 What This Chapter Is For

When a model resists, ask which reduction has failed. If the state is infinite, no grid exists and the state must be enlarged or the model abandoned. If the state is finite but no family is preserved, the model is fine and merely needs computing, which is chapter 19. If a family is preserved, find out whether it is a subspace or a manifold, because that determines whether the equations are linear or Riccati.

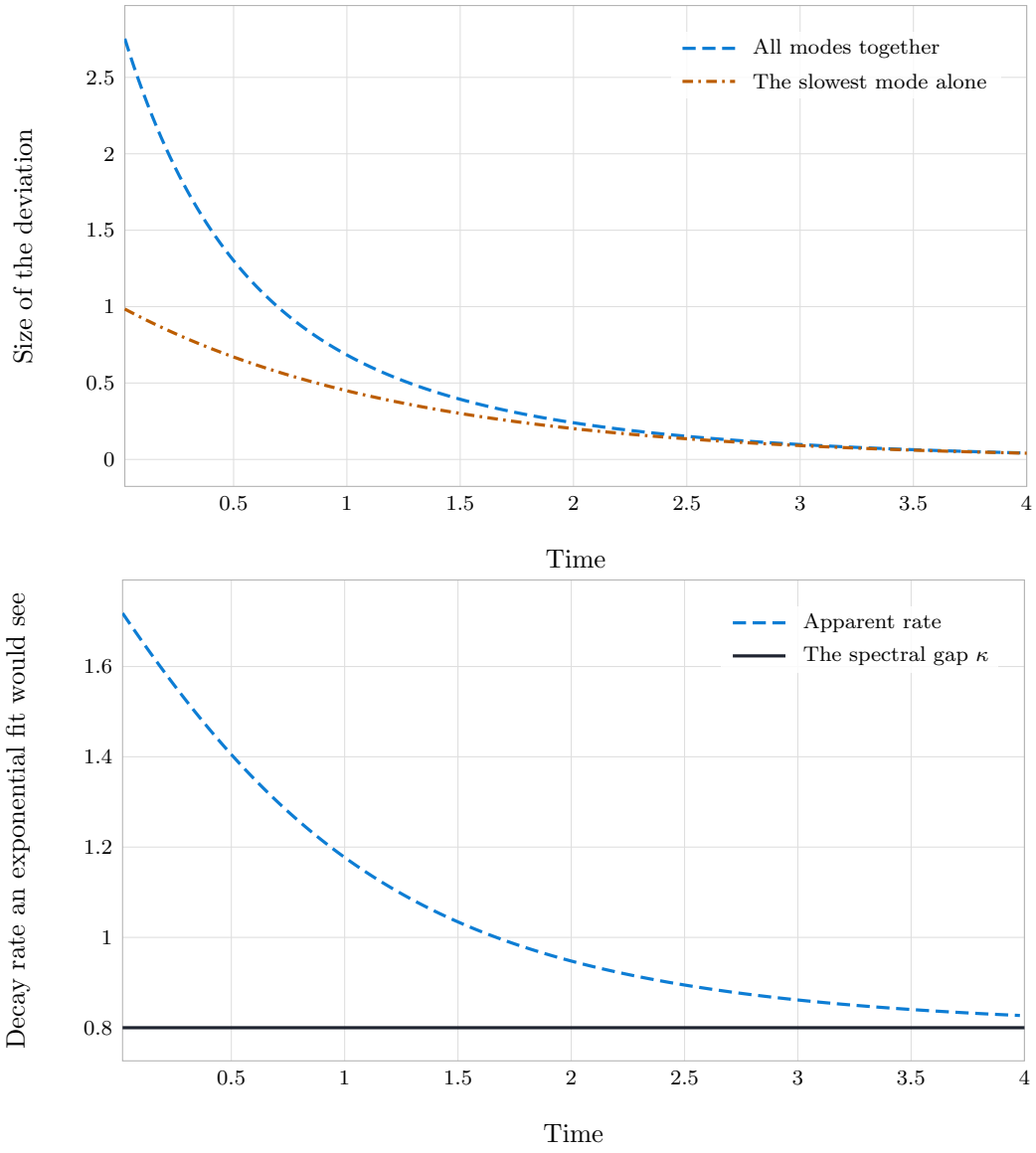

Figure 14.3: Above: a deviation from stationarity with positive weight on four modes, against the slowest mode alone. The total starts above the slowest mode and approaches it from above, because the higher modes are still alive early and decaying faster. Below: the decay rate an exponential fit would report if it began at that time, against the spectral gap it is supposed to be estimating. Early windows see a rate well above κ and the bias is one-sided — which is the same overstatement of mean reversion chapter 20's calculation Where the bias comes from and what it scales with finds from the estimator's side, arriving here from the spectrum's.

Drawn by `ou_spectral`.

References

- Carr, P., & Madan, D. (1999). Option valuation using the fast Fourier transform. *Journal of Computational Finance*, 2(4), 61–73.

- Duffie, D., Filipovic, D., & Schachermayer, W. (2003). Affine processes and applications in finance. *Annals of Applied Probability*, 13(3), 984–1053.

- Cuchiero, C., Keller-Ressel, M., & Teichmann, J. (2012). Polynomial processes and their applications to mathematical finance. *Finance and Stochastics*, 16(4), 711–740.

- Linetsky, V. (2004). The spectral decomposition of the option value. *International Journal of Theoretical and Applied Finance*, 7(3), 337–384.

- Forman, J. L., & Sorensen, M. (2008). The Pearson diffusions: a class of statistically tractable diffusion processes. *Scandinavian Journal of Statistics*, 35(3), 438–465.

- Reid, W. T. (1972). *Riccati Differential Equations*. Academic Press.

Part III

Products and Dependence

Chapter 15

CMS and Static Replication

A swap rate paid on its own annuity needs no model. Paid on any other date it needs a correction, and the obvious way to find that correction — calibrate a term structure model and ask it — is both more work and less honest than the alternative. This chapter derives the alternative: express the payoff as a portfolio of swaptions held to expiry, and read the answer off one expiry's smile. No short rate model, no term structure of volatility, no dynamics. It is chapter 9's replication argument moved into the annuity measure, and it prices constant maturity swaps and cash settled swaptions the way the market actually prices them.

15.1 A Rate Paid on the Wrong Date

Chapter 8 established the one fact this chapter leans on. The annuity is a positive price, so it is a numeraire, and the swap rate is a price divided by it, so

$$S(t) = \mathbb{E}_t^{\mathrm{Ann}}\big[S(T)\big] \tag{15.1}$$

is a martingale in the annuity measure. A swaption pays off on the annuity and (15.1) is why Black's formula applies to it unchanged.

Now change one thing. Instead of paying $S(T) - K$ over the life of a swap, pay it once, in cash, at a single date. The payoff is the same function of the same rate; only the numeraire it settles against has changed. And (15.1) was a statement about a particular numeraire, so it no longer applies: under the measure the new payment belongs to, the swap rate is not a martingale, and its expectation is not today's forward.

Remark (The same mistake, three chapters running)**.** This is chapter 8's Future-Forward Basis again, and chapter 16's quanto adjustment in advance. Each time a quantity that is a martingale under one measure is paid under another, and the gap is a convexity correction. What differs here is the remedy. The future-forward basis was computed inside a model because there was nothing else to compute it from. Here there is: the swaption market quotes the whole distribution of $S(T)$, and a distribution is enough.

15.2 The Annuity Mapping Function

The first step is to say what the wrong measure costs. Part of that can be said exactly and part of it cannot, and the chapter turns on knowing which is which.

Definition 15.1 (CMS Linked Products)**.** A CMS (Constant Maturity Swap) linked product is a contract that references swap rates of swaps of a fixed maturity in their cashflows. For example, a CMS linked strip could pay at every 6 months the swap rate of a swap starting then and ending 10 years from then.

Take a CMS caplet referencing a swap rate S, maturing at time T, with strike K. Cash settled at T, its price is

$$\text{CMSCap}(0) = P(0,T)\mathbb{E}^T[(S(T)-K)^+]$$

A swaption pays the same difference over the life of the underlying swap; the CMS cap settles it in cash at the option's own expiry instead.

One route from here is to calibrate a Hull-White model's parameters to bonds, futures, forwards and swaptions, then price the caplet inside it. The alternative below avoids that.

Allow also for a payment delay: the caplet settles at $T_p \geq T$ rather than at T itself.

$$\begin{aligned}
\text{CMSCap}_{T_p}(0) =& P(0,T_p)\mathbb{E}^{T_p}[(S(T)-K)^+] \\
=& P(0,T_p)\mathbb{E}^{\text{Ann}}[(S(T)-K)^+\frac{P(T,T_p)\text{Ann}(0)}{P(0,T_p)\text{Ann}(T)}] \\
=& \text{Ann}(0)\mathbb{E}^{\text{Ann}}[(S(T)-K)^+\frac{P(T,T_p)}{\text{Ann}(T)}] \\
=& \text{Ann}(0)\mathbb{E}^{\text{Ann}}[(S(T)-K)^+\mathbb{E}^{\text{Ann}}[\frac{P(T,T_p)}{\text{Ann}(T)}|S(T)=s]]
\end{aligned}$$

Everything to this point is exact. The change of numeraire is chapter 6's remark Prices in a numeraire are martingales in its measure and the last line is the tower property, so

$$M(s,T_p) \;=\; \mathbb{E}^{\text{Ann}}\Big[\tfrac{P(T,T_p)}{\text{Ann}(T)}\;\Big|\;S(T)=s\Big],$$

the *annuity mapping function*, is a well-defined object and not an approximation to one. It is also unknown. It is a conditional expectation under the pricing measure, so computing it means knowing the joint law of the ratio $P(T,T_p)/\text{Ann}(T)$ with the swap rate — which is exactly the term structure model this section set out to avoid.

Assumption 15.2 (Linear terminal swap rate)**.** M is affine in the rate,

$$M(s,T_p) \;=\; a(T_p)\,s + b(T_p). \tag{15.2}$$

This is the modelling content of the chapter and everything after it is again exact and the exactness is easy to read backwards onto (15.2).

$$\begin{aligned} M(s,T_p) =& \mathbb{E}^{\mathrm{Ann}}[\frac{P(T,T_p)}{\mathrm{Ann}(T)}|S(T)=s] \\ =& a(T_p)s + b(T_p) \end{aligned}$$

Once the distribution of $S(T)$ under the annuity measure is known — inferred by swaption replication, from market prices at the same expiry and swap end date across strikes — and a and b are fixed, the CMS cap price is the price of a call on a quadratic function of $S(T)$. No term structure model, Hull-White or otherwise, is needed to get there.

Three no-arbitrage identities pin a and b. They are facts the *true* M satisfies whatever its shape, so they constrain the assumption rather than establish it: they say which affine map to use, not that an affine map is right.

Calculation 15.3 (Pinning down the linear map)**.** *Martingale property.* $M(S(T),T_p)$ is itself a price divided by the annuity, so it is a martingale in the annuity measure:

$$\mathbb{E}^{\mathrm{Ann}}\big[M(S(T),T_p)\big] = \frac{P(0,T_p)}{\mathrm{Ann}(0)}.$$

Substituting (15.2) and using $\mathbb{E}^{\mathrm{Ann}}[S(T)] = S(0)$ from (15.1),

$$\begin{aligned} a(T_p)S(0) + b(T_p) =& \frac{P(0,T_p)}{\mathrm{Ann}(0)} \\ b(T_p) =& \frac{P(0,T_p)}{\mathrm{Ann}(0)} - a(T_p)S(0) \\ M(s,T_p) =& a(T_p)(s-S(0)) + \frac{P(0,T_p)}{\mathrm{Ann}(0)}. \end{aligned}$$

The annuity, by definition, in zero-coupon bonds. With the fixed leg the payer leg,

$$\sum_{i=1}^{N_p} \delta_i^p P(T,\tau_i^p) = \mathrm{Ann}(T).$$

Dividing by $\mathrm{Ann}(T)$, taking the conditional expectation used to define M, and substituting the martingale-property result just derived,

$$\begin{aligned} \sum_{i=1}^{N_p} \delta_i^p \mathbb{E}^{\mathrm{Ann}}\Big[\frac{P(T,\tau_i^p)}{\mathrm{Ann}(T)}\,\Big|\,S(T)=s\Big] =& 1 \\ \sum_{i=1}^{N_p} \delta_i^p M(s,\tau_i^p) =& 1 \\ \sum_{i=1}^{N_p} \delta_i^p a(\tau_i^p)(s-S(0)) + \sum_{i=1}^{N_p} \delta_i^p \frac{P(0,\tau_i^p)}{\mathrm{Ann}(0)} =& 1. \end{aligned}$$

The second sum is $\mathrm{Ann}(0)/\mathrm{Ann}(0) = 1$ by the same definition applied at time 0, so the identity holds for every s only if

$$\sum_{i=1}^{N_p} \delta_i^p a(\tau_i^p) = 0.$$

The swap rate, by definition, in bond prices. Writing L_i for the forward rate fixing over $[\tau_{i-1}^p, \tau_i^p]$ as seen at T, the fixed-versus-floating par condition and the single-curve telescoping identity give

$$S(T) = \frac{\sum_{i=1}^{N_p} L_i(T, \tau_{i-1}^p, \tau_i^p)\, \delta_i^p\, P(T, \tau_i^p)}{\mathrm{Ann}(T)} = \frac{1 - P(T, \tau_{N_p}^p)}{\mathrm{Ann}(T)}.$$

Conditioning on $S(T) = s$ and substituting the martingale-property result at $T_p = T$ and at $T_p = \tau_{N_p}^p$,

$$\begin{aligned} s =& M(s,T) - M(s, \tau_{N_p}^p) \\ =& (a(T) - a(\tau_{N_p}^p))(s - S(0)) + \frac{P(0,T) - P(0,\tau_{N_p}^p)}{\mathrm{Ann}(0)} \\ =& (a(T) - a(\tau_{N_p}^p))(s - S(0)) + S(0), \end{aligned}$$

using $\big(P(0,T) - P(0,\tau_{N_p}^p)\big)/\mathrm{Ann}(0) = S(0)$ from the same telescoping identity applied at time 0. This holds for every s only if

$$a(T) - a(\tau_{N_p}^p) = 1.$$

Solving. The two boxed identities constrain a only at the dates $T, \tau_1^p, \ldots, \tau_{N_p}^p$, which is not enough to fix a whole function — a second modelling choice, alongside Assumption 15.2, is needed to close the system:

Assumption 15.4 (The mapping's slope is affine in the payment date)**.** $a(t) = ut + v$ for constants u, v.

With this, the swap-rate identity gives $u(T - \tau_{N_p}^p) = 1$, so $u = 1/(T - \tau_{N_p}^p)$; the annuity identity then gives $u \sum_i \delta_i^p \tau_i^p + v \sum_i \delta_i^p = 0$, so

$$u = \frac{1}{T - \tau_{N_p}^p}, \qquad v = -u\, \frac{\sum_{i=1}^{N_p} \delta_i^p \tau_i^p}{\sum_{i=1}^{N_p} \delta_i^p},$$

which fully specifies the linear annuity mapping and, with it, the CMS caplet price — without a term structure model.

Remark (What the assumption buys, and what it costs)**.** Without it, pricing the caplet needs the joint law of the curve with the swap rate, which is a term structure model — calibrated, simulated, and carrying every choice chapter 12 and chapter 13 spent their length measuring. With it, the payoff becomes a function of $S(T)$ alone, and the price then needs only the *marginal* law of $S(T)$ under the annuity measure. That marginal is quoted: it is the swaption smile at that expiry and tenor, read off by chapter 9's theorem Breeden-Litzenberger. So the assumption is what converts

a model problem into a replication problem, and it is why the method exists at all — it prices a CMS from the swaption smile rather than from a model of the curve.

What it costs is harder to state than it looks, because the assumption is better than a first reading suggests and the exposure is somewhere else.

Note first that the mapping can be *derived* rather than assumed, and the chapter's own opening said how: pick a term structure model. In any one-factor model the whole curve at T is a deterministic function of the single state, so the swap rate and the ratio $P(T,T_p)/\mathrm{Ann}(T)$ are both explicit functions of it, the state can be eliminated between them, and no conditioning is required at all — the conditional expectation collapses to a function. In Hull-White the curve is exponential-affine in that state, so the mapping is an explicit ratio of sums of exponentials, and it is certainly not a straight line.

But it is very nearly one. Over a band of two hundred basis points either side of the forward, the exact Hull-White mapping departs from its own best straight line by about a fifth of one per cent of its level.[1] Assumption 15.2 is therefore a good description of the mapping's *shape*, and the worry it invites — that a curved object is being straightened — is not where the money is.

Where the money is, is which straight line. The slope the model gives depends on the mean reversion: at κ near zero it is 0.63, and at $\kappa = 0.10$ it is 0.70, eleven per cent higher.[2] That is fifty times the shape error, and it is a fact about a parameter chapter 12 identifies as the largest model risk in a callable book and as invisible in the instruments used to calibrate.

So deriving the mapping from a model does not remove the assumption. It renames it. A free coefficient $a(T_p)$, pinned by the arbitrage conditions up to one number, becomes a coefficient determined by κ — and κ is not quoted. What is gained is that the remaining freedom now has a name, a plausible range, and a literature; what is not gained is a number the market has told anyone.

Where the curvature does show up is far from the money, and it grows about as fast as the square of the distance: a twentieth of a per cent of the mapping's level at a hundred basis points either side, a per cent and a half at six hundred.[3] That matters because the replication below weights the wings, so the one region where a straight line is a poor fit is the region the method leans on. It remains a smaller effect than the slope.

It is not, incidentally, made worse by a payment delay. The natural guess is that delaying the payment makes the mapping matter more and bends it further; the opposite happens. A delayed bond falls with the rate as well, which partly offsets the annuity falling underneath it, so the slope drops from 0.64 to 0.35 as the delay runs out to two years and the departure from a straight line falls with it.[4]

So the recommendation to price the trade a second way and reserve the difference, which chapter 20 makes generally, needs sharpening here. The second way *is* a term structure model, and the model is where the mapping came from — so the two prices do not differ by method at all. They differ by κ, and the reserve to carry is a reserve against mean reversion.

[1] `measured`.
[2] `measured`.
[3] `measured`.
[4] `measured`.

15.3 The Payoff as a Portfolio

Everything above is exact and none of it is yet computable, because M multiplies the payoff inside an expectation whose distribution we have not written down.

Theorem 15.5 (Static replication)**.** *Let g be twice differentiable and let $F = S(0)$. Then for every s,*

$$g(s) = g(F) + g'(F)(s - F) + \int_F^\infty g''(K)\,(s - K)^+\,dK + \int_0^F g''(K)\,(K - s)^+\,dK. \tag{15.3}$$

Proof. Taylor's theorem with integral remainder, split at F. For $s > F$, $g(s) = g(F) + g'(F)(s - F) + \int_F^s g''(K)(s-K)dK$, and the integrand vanishes for $K > s$, so the upper limit may be taken to infinity with $(s-K)$ replaced by $(s-K)^+$. The case $s < F$ is the same computation downwards. □

Equation (15.3) is an identity in s and not an approximation, so it can be checked at a point rather than only through the prices it produces — and it is, for payoffs of both signs of curvature, at levels either side of the forward.[5]

Structure (Replication is a change of basis)**.** Theorem 15.5 is easier to hold on to as a statement about coordinates than as an identity to verify.

Payoffs on a single rate form a vector space. Equation (15.3) exhibits a basis for it — a constant, a forward, and the calls and puts at every strike — and gives the coordinates of an arbitrary payoff in that basis, namely $g''(K)$. Pricing is then linear algebra: apply the pricing functional to each basis element, which the market has already done and published as a smile, and take the same combination.

Chapter 9's Breeden-Litzenberger is the dual statement in the same basis. There the coordinates of the *density* were extracted from prices; here the coordinates of the *payoff* are extracted from the payoff. The two are the two halves of chapter 3's pairing, and the reason both are model-free is that neither ever asks how the underlying moves — a basis decomposition is not a dynamical statement.

Remark (What it says)**.** Any European payoff on the swap rate is a portfolio: some cash, a forward, and a continuum of swaptions with weights $g''(K)$. The portfolio is bought once and held — nothing is rebalanced, no hedge ratio is computed, no process is assumed. This is chapter 9's Breeden-Litzenberger read in the other direction: there, option prices gave the density; here, the density is never formed and the option prices are used directly.

Taking expectations in the annuity measure, where $\mathbb{E}^{\mathrm{Ann}}[S(T) - F] = 0$ by (15.1),

$$\mathbb{E}^{\mathrm{Ann}}[g(S(T))] = g(F) + \int_F^\infty g''(K)\,C(K)\,dK + \int_0^F g''(K)\,P(K)\,dK, \tag{15.4}$$

with C and P the payer and receiver swaption prices per unit annuity — which is to say, the market's own quotes with the annuity divided out.

[5] `checked`, together with `the fact that the expansion point is arbitrary`: expanding about the forward is a convenience that makes the linear term vanish in expectation, not a requirement of the algebra.

Calculation 15.6 (The adjustment, from the smile)**.** Put the two halves together. With the linear map, $M(s) = a(s - F) + P(0, T_p)/\text{Ann}(0)$, and the CMS swaplet is worth

$$\begin{aligned}\text{Ann}(0)\,\mathbb{E}^{\text{Ann}}\big[S(T)M(S(T))\big] &= \text{Ann}(0)\Big(a\,\mathbb{E}^{\text{Ann}}[S(S-F)] + \tfrac{P(0,T_p)}{\text{Ann}(0)}\,\mathbb{E}^{\text{Ann}}[S]\Big)\\ &= \text{Ann}(0)\,a\,\text{Var}^{\text{Ann}}(S(T)) + P(0,T_p)\,F,\end{aligned}$$

so that

$$\mathbb{E}^{T_p}[S(T)] - S(0) = \frac{\text{Ann}(0)}{P(0,T_p)}\,a(T_p)\,\text{Var}^{\text{Ann}}\big(S(T)\big). \tag{15.5}$$

And the variance is (15.4) with $g(s) = s^2$, whose second derivative is the constant 2:

$$\text{Var}^{\text{Ann}}(S(T)) = 2\left[\int_F^\infty C(K)\,dK + \int_0^F P(K)\,dK\right]. \tag{15.6}$$

Remark (Read what is on the right hand side)**.** Nothing in (15.5) or (15.6) is a model output. The annuity and the discount factor are chapter 7's curve. The coefficient a came from three no-arbitrage conditions. The variance is an integral of quoted swaption prices across strikes, with equal weight at every strike — which is the same object a variance swap pays, met here in a rates market instead of an equity one.

So the CMS convexity adjustment is a statement about the swaption smile and nothing else. A desk that has calibrated a Hull-White model to price it has done a great deal of work to obtain an answer that its own screen already contained, and has additionally made the answer depend on the model's parameterisation — a one factor Gaussian model has a particular smile, generally not the market's, and the CMS price inherits it.

Remark (Why the whole smile and not the at-the-money)**.** Equation (15.6) weights every strike equally. That is the sentence to remember, because it explains why CMS is a wing product: a payoff that is quadratic in the rate has constant g'', so the far strikes contribute as much per unit of strike as the near ones, and they are precisely the strikes with the least liquidity and the widest bid-offer.

It is also why a CMS book is exposed to the volatility of volatility rather than merely to volatility. Raising ν in the SABR model of chapter 10 lifts the wings while leaving the at-the-money almost unmoved, and (15.6) charges for the wings.

15.4 Cash Settled Swaptions

The same defect, in a product that looks like it should not have one.

A cash settled swaption does not deliver a swap. It pays a cash amount computed from the swap rate at expiry using a *formula* for the annuity rather than the annuity itself — typically

$$\text{Ann}_{\text{cash}}(S(T)) = \sum_{i=1}^{N} \frac{\delta}{(1 + \delta S(T))^i},$$

a function of the single rate, standing in for a sum of discount factors that are not all determined by it.

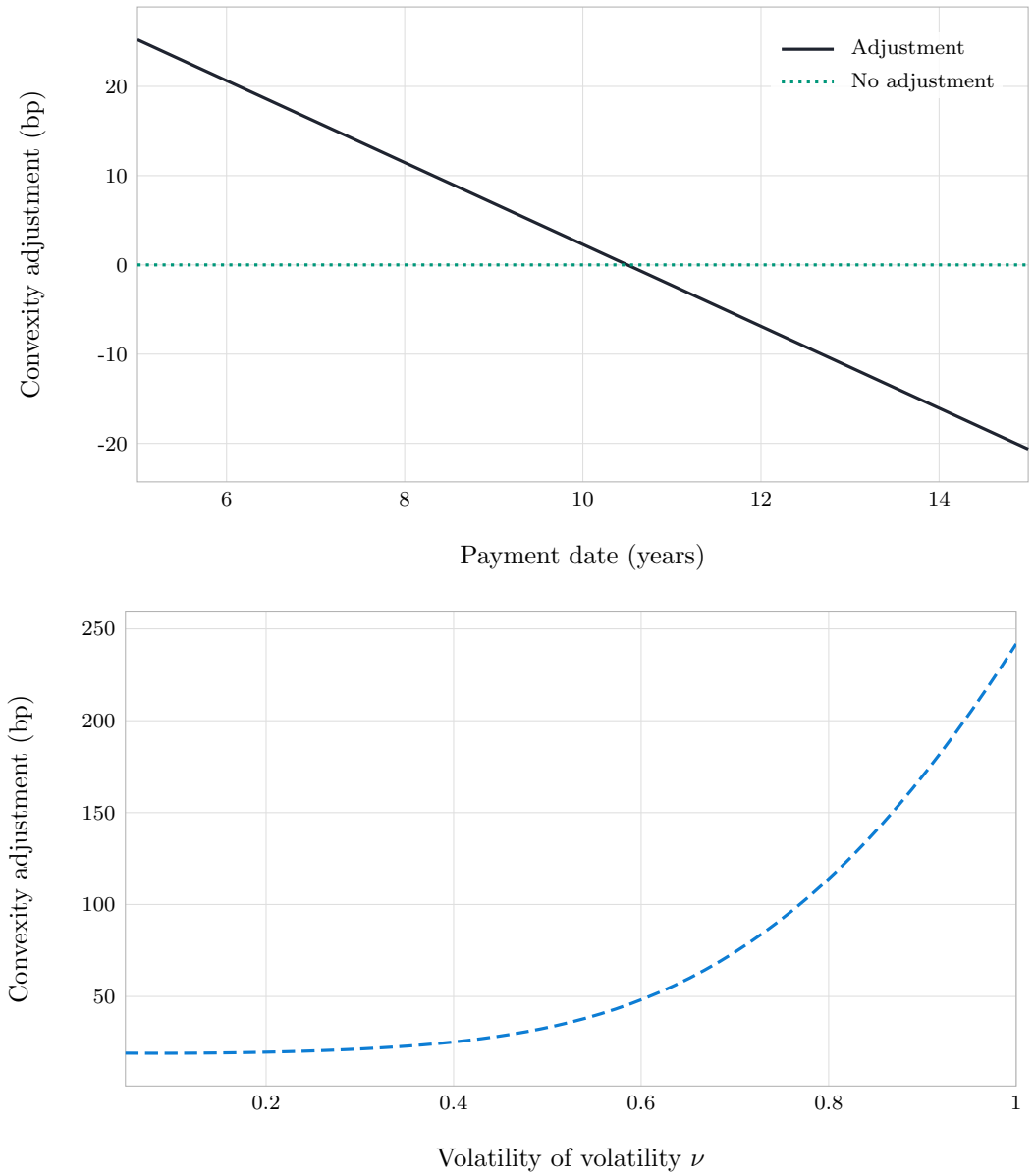

Figure 15.1: The CMS convexity adjustment on a five year expiry into a ten year swap, computed by replication over a SABR smile. The left panel is the adjustment against the payment date. It is largest when the rate is paid at the start of the swap it references, falls as the payment is delayed, and crosses zero — the coefficient $a(t)$ declines across the swap, so far enough out the timing effect more than cancels the convexity. The right panel is the adjustment against the volatility of volatility at a fixed at-the-money level, which is flat only if one believes the wings do not matter. They contribute equally per unit of strike, by (15.6), and a CMS book is long them.
Drawn by `convexity_adjustment` and `annuity_variance`.

Remark (Why this is the same problem)**.** The physically settled swaption pays $\text{Ann}(T)(S(T)-K)^+$ and the annuity in front is the numeraire, so it divides out and Black's formula survives. The cash settled one pays $\text{Ann}_{\text{cash}}(S(T))(S(T)-K)^+$, and Ann_{cash} is *not* the numeraire — it is a different function that happens to agree with it on average. The ratio $\text{Ann}_{\text{cash}}(S)/\text{Ann}(T)$ does not divide out, and what is left is again an expectation of the payoff against something correlated with it.

So the two are not the same trade with a settlement convention between them. They differ by a convexity correction of the same species as the CMS one, and it is computed the same way: the payoff is a function of $S(T)$ alone, therefore Theorem 15.5 applies, therefore it is a portfolio of ordinary swaptions and the smile prices it.

Exercise. A cash settled payer swaption struck at the forward is worth less than the physically settled one at the same strike. Show this from the fact that Ann_{cash} is a decreasing function of S while the payoff $(S-K)^+$ is increasing, so the two are negatively correlated. (Hint: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] + \text{Cov}(X,Y)$, and the covariance of an increasing and a decreasing function of the same variable is negative.) Deduce that the difference grows with volatility, and explain why the two conventions were therefore never interchangeable, even before anyone computed the size of the gap.

15.5 What This Chapter Did Not Need

No short rate was modelled. No volatility was made a function of time. No parameters were calibrated to a surface, and nothing was simulated. The inputs were a discount curve, a swaption smile at one expiry, and three conditions that follow from the definitions of an annuity and a swap rate.

Remark (When a model is unavoidable)**.** The argument above does not generalise as far as it first appears.

Static replication works when the payoff is a function of *one* rate observed at *one* date. Then the smile at that expiry is the whole of the relevant distribution and there is nothing left to assume. A CMS caplet qualifies. A cash settled swaption qualifies.

A Bermudan swaption does not. Its value depends on the joint behaviour of the rate at several exercise dates, and no collection of European quotes determines a joint distribution — which is chapter 18's point about marginals arriving again. A callable range accrual does not qualify either, nor does anything whose payoff depends on the path. For those a model is genuinely required, and which model is the subject of chapter 20.

References

- Hagan, P. S. (2003). Convexity conundrums: pricing CMS swaps, caps, and floors. *Wilmott Magazine*, March, 38–44.

- Carr, P., & Madan, D. (2001). Optimal positioning in derivative securities. *Quantitative Finance*, 1(1), 19–37.

- Mercurio, F., & Pallavicini, A. (2006). Smiling at convexity: bridging swaption skews and CMS adjustments. *Risk*, 19(8), 64–69.

Chapter 16

Credit

We price the instruments that trade on default. Chapter 2 built the default time and observed that its survival probability is formally a discount factor; here we collect on that observation, deriving the credit default swap, the relation between a spread and a hazard rate — which turns out to be exact rather than approximate — and the bootstrap that turns a handful of quotes into a curve. We end with the one thing a single-name model cannot do, which is the subject of chapter 18.

16.1 What We Already Have

Chapter 2 did most of the modelling. Default is the first jump of a Cox process with intensity λ_t, the intensity exists because default arrives at a totally inaccessible time, and the survival probability is

$$Q(t) = \mathbb{P}(\tau > t) = \mathbb{E}\Big[e^{-\int_0^t \lambda_s\, ds}\Big], \tag{16.1}$$

which is chapter 7's bond price with the intensity where the short rate was.

What is missing is everything on the instrument side: what trades, how it is priced, and how λ is recovered from what trades. That is this chapter, and the reader who has been through chapters 7 and 8 will find the shape familiar — an instrument that pays a stream against a contingent payment, a par rate at which the two balance, and a bootstrap. It is the swap market again with a different underlying.

Throughout, R is the recovery rate: the fraction of notional a claim is worth after default. The loss given default is $1 - R$. We take λ deterministic, which makes (16.1) an ordinary integral, and note where stochastic intensity would change things.

16.2 The Credit Default Swap

Definition 16.1 (Credit default swap)**.** A credit default swap is an agreement in which the buyer of protection pays a periodic spread s on a notional until the earlier of maturity and default, and

in exchange receives $1 - R$ at default if it happens before maturity.

It is an insurance contract written as a swap, and like a swap it has two legs to value.

Calculation 16.2 (The protection leg)**.** The buyer receives $1 - R$ at τ if $\tau \le T$. Default happens in $[t, t + dt]$ with probability $Q(t)\lambda_t\, dt$ — survive to t, then default immediately — so

$$\text{Prot} = (1 - R) \int_0^T P(0,t)\, \lambda_t\, Q(t)\, dt, \tag{16.2}$$

with $P(0,t)$ the ordinary discount factor of chapter 7. Note that $-dQ/dt = \lambda_t Q(t)$, so this is the loss integrated against the density of the default time.

Calculation 16.3 (The premium leg)**.** The seller receives s per period while the name survives. On payment dates t_i with accruals δ_i, the expected discounted receipts are

$$s \sum_i \delta_i P(0,t_i) Q(t_i),$$

plus one correction. If default falls partway through a period, the protection buyer still owes the spread accrued up to that moment; conventionally this is taken as half a period, contributing

$$s \sum_i \tfrac{1}{2} \delta_i P(0,t_i) \big(Q(t_{i-1}) - Q(t_i)\big).$$

The sum of the two, per unit of spread, is the *risky annuity* RiskyAnn — chapter 7's annuity with survival probabilities inserted alongside the discount factors.

Definition 16.4 (Par spread)**.** The par spread is the s making the contract worth zero:

$$s = \frac{\text{Prot}}{\text{RiskyAnn}}. \tag{16.3}$$

This is chapter 7's par swap rate again — a contingent leg over an annuity — and everything structural follows as it did there. The par spread is what is quoted. The risky annuity is the sensitivity of the contract's value to the spread, so it is the desk's risk measure, quoted in the same units and for the same reason.

16.3 The Credit Triangle

There is a relation between spread, hazard and recovery that every credit desk uses to convert between them in their head. It is usually presented as an approximation. In the model it is an identity: no expansion is involved, and the interest rate curve cancels. What is approximate is the model's fit to the contract, which pays quarterly, and the size of that error is the calculation *What the payment convention costs* below.

Theorem 16.5 (The credit triangle)**.** *With a flat hazard rate λ and premium paid continuously, the par spread is*

$$s = \lambda\,(1 - R) \tag{16.4}$$

exactly, for any deterministic interest rate curve.

Proof. With λ constant, $Q(t) = e^{-\lambda t}$. Writing the premium leg as a continuous integral, the two legs are

$$\text{Prot} = (1-R)\int_0^T P(0,t)\,\lambda\,e^{-\lambda t}\,dt,$$
$$\text{RiskyAnn} = \int_0^T P(0,t)\,e^{-\lambda t}\,dt.$$

The integrands differ by the constant factor $(1-R)\lambda$ and by nothing else. So the ratio (16.3) is that constant, whatever $P(0,\cdot)$ happens to be. □

The interest rate curve cancelling entirely deserves a pause. A credit default swap is a swap between two streams that stop at the same moment and are discounted by the same factors, so the discounting divides out. That is why a credit spread can be quoted and traded without reference to the rate curve, and why the credit market and the rates market can move independently without either repricing the other.

Example 16.1 (Converting in your head)**.** A name quoted at 150 basis points, at the conventional forty percent recovery, is defaulting at

$$\lambda = \frac{0.0150}{0.6} = 0.025,$$

two and a half percent a year. Over five years that is a survival probability of $e^{-0.125} = 88\%$, so the market is pricing about a one in eight chance of this name failing within five years. Running it the other way, a desk that believes a name defaults with probability one in twenty over five years believes $\lambda \approx 0.01$ and should be paid about 60 basis points to insure it.

Doing this conversion quickly is most of what the triangle is for. It turns a quoted price into a statement about the world.

Calculation 16.6 (What the payment convention costs)**.** Real contracts pay quarterly, not continuously, so (16.4) is not exactly the quoted spread.

A premium paid at the end of each period is on average half a period later than one paid continuously, and that delay is discounted by both the interest rate and the survival probability. But the accrued-interest convention — the half-period term in the premium leg — was put there precisely to compensate the survival part of the delay. What it does not compensate is the interest rate part. So the error should be about half a period of interest,

$$\frac{s_{\text{quarterly}} - \lambda(1-R)}{\lambda(1-R)} \approx \frac{r\,\delta}{2}, \tag{16.5}$$

with δ the accrual period. At a three percent rate and quarterly payment this predicts 0.375%, and the model in `quant/src/credit.rs` gives 0.376%. At annual payment it predicts 1.5% and gives 1.517%. At a zero interest rate it predicts nothing at all, and the computed error is zero to eight decimal places at every spread tried.

Notice what does *not* appear in (16.5): the spread. The error is 0.38% at thirty basis points and 0.36% at twelve hundred. This is the opposite of the usual guess — that the triangle "breaks

down for distressed names" — and the reason is Theorem 16.5: the hazard rate cancels between the legs whatever it is, so nothing that survives the cancellation can depend on it. What survives is the interest rate, which the theorem removed only because continuous payment let it cancel too.

Remark (Recovery cannot be fitted)**.** Equation (16.4) pins down the product $\lambda(1-R)$ and not the two factors. A quoted spread of 125 basis points is equally consistent with a two percent hazard and forty percent recovery, or a three percent hazard and sixty percent recovery, and no amount of staring at that quote will separate them.

This is why recovery is fixed by convention — forty percent for senior unsecured, conventionally — rather than calibrated. It is not laziness. There is nothing in the price to calibrate it against, and pretending otherwise would produce a number that looked estimated and was in fact assumed. Separating the two requires an instrument sensitive to recovery on its own, which is what a recovery swap or a pair of bonds at different seniority provides.

Compare this with the identification problems of chapters 10 to 12. It is the same shape — a product of parameters is observable, the factors individually are not — but with an unusually honest resolution, since the market agreed on a convention rather than each desk picking its own.

16.4 Bootstrapping the Hazard Curve

One quote gives one hazard rate. A curve of quotes — one, three, five, seven and ten years are the liquid points — gives a term structure, and the procedure is chapter 7's.

Take the hazard rate to be piecewise constant between quoted maturities. Then the one year quote involves only the first segment and fixes it. The three year quote involves the first segment, now known, and the second, which it therefore fixes. And so on outward.

Remark. The parallel with chapter 7 is exact enough:

Rates	**Credit**
short rate r_t	hazard rate λ_t
discount factor $P(0,T)$	survival probability $Q(T)$
par swap rate	par CDS spread
annuity	risky annuity
bootstrap from swaps	bootstrap from CDS
forward rate	forward hazard

And the correspondence is not an analogy but an identity of formulas, for the reason chapter 2 gave: (16.1) and the bond price are the same expectation.

Example 16.2 (The curve behind the figure)**.** The five quotes drawn opposite, at forty percent recovery and a three percent rate, bootstrap to[1]

[1] the whole worked example, checked against the bootstrap.

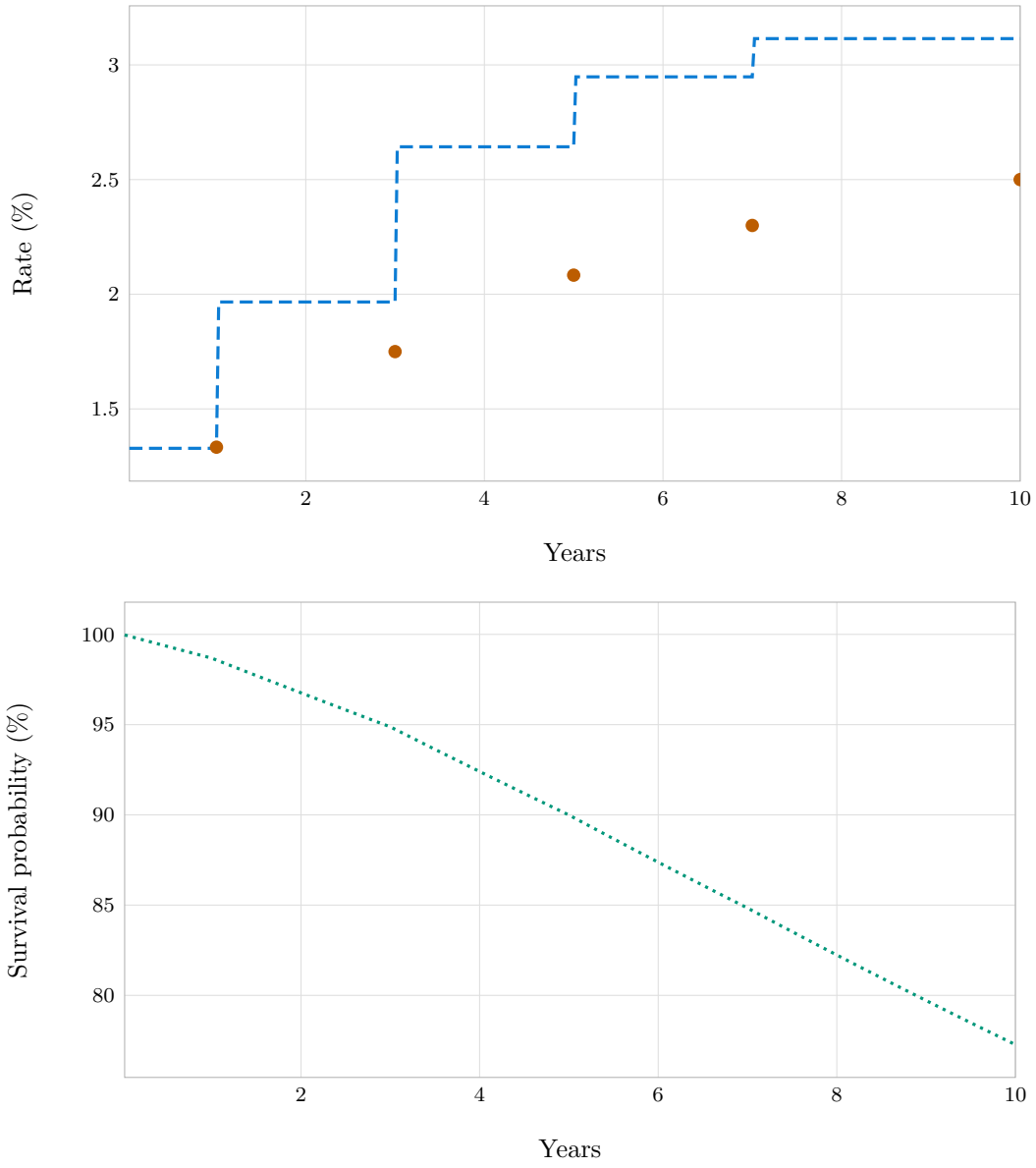

Figure 16.1: A hazard curve bootstrapped from five quoted spreads, with the survival curve it implies. The circles are the triangle applied to each quote individually, which is what a desk computes in its head; the line is the forward hazard the bootstrap actually implies. They agree at the front, where there is nothing before to average, and separate as the curve extends — the forward hazard reaching over three hundred basis points where the ten year triangle reads two hundred and fifty. This is chapter 7's lesson repeating: a quote is an average over its whole life, forwards move further than the averages that contain them, and it is the forwards that anything between the quoted tenors prices off.

Drawn by `bootstrap_hazards` and `par_spread`.

Tenor	Quote	Triangle	Forward hazard
1y	80 bp	133 bp	133 bp
3y	105 bp	175 bp	197 bp
5y	125 bp	208 bp	264 bp
7y	138 bp	230 bp	295 bp
10y	150 bp	250 bp	311 bp

Read the last two columns against each other. At one year they agree, because there is nothing before the first quote for it to be an average of. By ten years the triangle reads 250 basis points and the forward hazard is 311, a quarter higher — the ten year quote is an average over the whole decade, and the market is saying the last three years of it are considerably worse than the first.

A desk pricing a seven-to-ten year forward CDS off the ten year triangle would be using 250 where the curve says 311, and would sell that protection a fifth too cheap. This is chapter 7's warning again in a new market: a quote is an average, a product prices off a forward, and the two are not interchangeable.

16.5 Marking a Position

The par spread prices a new contract. An existing one was struck at some spread K that the market has since left behind.

Calculation 16.7 (The value of an off-market contract)**.** A protection buyer paying K when the market charges s is paying $K - s$ too much, on a stream that survives exactly as the annuity says. So the position is worth

$$V = (s - K) \times \text{RiskyAnn}, \tag{16.6}$$

to the buyer of protection, which is chapter 7's swap valuation with the risky annuity in place of the ordinary one.

Concretely, using the curve above: a five year contract struck at 100 basis points, marked against a market at 125, with a risky annuity of 4.43, is worth[2]

$$0.0025 \times 4.43 = 1.11\%$$

of notional to the buyer of protection. Twenty five basis points of spread became a hundred and eleven basis points of value, because the position runs for about four and a half years' worth of annuity.

Remark (Why contracts trade with an upfront)**.** Equation (16.6) explains a market convention that otherwise looks arbitrary. Standard contracts do not trade at their par spread at all; they pay a fixed coupon, 100 basis points for investment grade names and 500 for high yield, and the difference between that coupon and the par spread is settled as a cash payment at the start.

That payment is exactly V. Fixing the coupon means every contract on a name has identical cash flows regardless of when it was struck, so two positions can be netted against each other and torn up, instead of accumulating as an ever-growing book of offsetting trades with different

[2] the risky annuity and the mark to market, checked.

coupons. The convention exists to make positions fungible, and (16.6) is the formula that converts between the running spread a trader thinks in and the upfront the contract actually settles.

Remark (The annuity is shorter than it looks)**.** The riskless five year quarterly annuity at a three percent rate is 4.63.[3] The risky one above is 4.43, four percent shorter. The missing four percent is the premium that will not be paid because the name may not survive to pay it — which is to say the risky annuity is doing real work, and substituting the ordinary annuity into (16.6) would overstate the position by that much. On a distressed name the gap is far larger.

16.6 The Two Risk Numbers

A credit position carries two risks that cannot be netted, and chapter 2 explained why: they are the diffusive part and the jump part of the same position, and they require different instruments.

Definition 16.8 (CS01)**.** The change in value for a one basis point widening of the spread curve. By (16.3) it is the risky annuity times a basis point, exactly as a swap's DV01 is its annuity.

Definition 16.9 (Jump to default)**.** The change in value if the name defaults immediately. For a protection buyer it is a gain of $(1 - R)$ less the value already marked; for a bond holder a loss of $(1 - R)$ times notional.

CS01 is the risk of the spread drifting. Jump to default is the risk of the event. A position can be flat on one and enormously exposed on the other — selling protection on a name trading at fifty basis points has a small CS01 and a jump to default of sixty percent of notional — which is precisely why they sit on separate lines of a risk report rather than being combined into one number.

16.7 Making the Intensity Random

Everything so far took λ deterministic.

A deterministic intensity makes the whole future spread curve known today. So an option on a spread is worth its intrinsic value, a position in spread volatility does not exist, and the only uncertainty left in the model is *when* the default arrives rather than *what the market will think* of the name in a year. That is the same complaint chapter 9 made about a deterministic volatility, arriving in a different market, and it has the same remedy.

Let λ be a process. The survival probability is then

$$Q(0,T) \;=\; \mathbb{E}\Big[\exp\Big(-\int_0^T \lambda_s\, ds\Big)\Big], \tag{16.7}$$

and the reader should recognise that expectation, because it is the one chapter 8 spent a chapter on with r in place of λ. A credit curve is a discount curve on a different rate, so every model in this book's term structure chapters transfers to it unchanged — and so does every one of their limitations.

[3] `the riskless annuity, checked`.

Remark (Which model, and the one constraint credit adds)**.** The choice is not quite free, because an intensity is an arrival rate and a negative arrival rate is meaningless. Hull-White, which allows r to go negative and is none the worse for it in rates, is not admissible here.

The usual answer is a square-root process, $d\lambda = \kappa(\theta - \lambda)dt + \eta\sqrt{\lambda}\,dZ$, whose diffusion switches off as the intensity approaches zero. That is chapter 10's backbone question with a hard constraint attached: the exponent is not being fitted to data here but forced to a half by the requirement of positivity.

The reward is that the model is affine, so (16.7) is exponential-affine in λ_0 with coefficients solving a Riccati pair, by the theorem of chapter 14 — and for this process the pair has a closed form. The credit curve is then computed by evaluating a formula rather than by simulating, which is what makes calibration to a strip of CDS quotes practical.[4]

Calculation 16.10 (What randomness alone does to the spread)**.** Randomness in the intensity changes the curve even when it changes nothing about the intensity's average, and the direction is fixed rather than a matter of parameters.

Take $\lambda_0 = \theta = 2\%$, so the intensity's mean is flat at two percent for every horizon, and $\kappa = 0.5$, $R = 40\%$. With $\eta = 0$ the credit triangle gives a flat five year spread of 120 basis points. Turning on the vol of intensity lowers it monotonically, and at $\eta = 0.20$ it has fallen by 4.2 basis points.[5]

The reason is Jensen: (16.7) is a convex function of the integrated intensity, so $\mathbb{E}[e^{-X}] > e^{-\mathbb{E}[X]}$ and a random intensity survives better than its own average. A desk bootstrapping a deterministic curve from quotes is therefore not recovering the mean intensity; it is recovering something below it, by an amount that is a pure convexity and grows with the volatility of the spread.

16.8 Options on Credit

With the intensity random there are credit derivatives to price, and the machinery for them is already in this chapter.

Definition 16.11 (Payer and receiver swaptions on a CDS)**.** A payer gives the right at T to buy protection to T_b at a strike spread K; a receiver, the right to sell it.

The valuation follows chapter 8's swaption argument with one substitution. Equation (16.6) says a position struck at K is worth $(s-K)$RiskyAnn, so the payer's payoff at T is $\text{RiskyAnn}(T)\,(s(T)-K)^+$ — an option on the forward spread, with the *risky* annuity multiplying it. The risky annuity is a positive traded portfolio, so it is an admissible numeraire; the forward spread is a traded value divided by it, so it is a martingale in the corresponding measure; and the option prices by Black in that measure, exactly as a swaption does.

So a credit desk quotes spread volatility, runs a smile, and inherits every question of chapter 10 — with lognormal the usual convention, since a spread cannot go negative and is far from zero.

Remark (What is different: the numeraire can vanish)**.** The parallel with swaptions is exact but for one thing. The risky annuity is not merely positive: it can jump to zero, because the name can

[4] `credit::CirIntensity`, checked against a simulation sharing none of its algebra.

[5] `measured`.

default before the option expires. A numeraire that can vanish is not a numeraire on that event, and the argument above says nothing about what happens there.

The market resolves this by contract rather than by mathematics, and in two different ways. A single name option that *knocks out* on default is worthless if the name goes, and for it the Black argument holds on the surviving paths, which are the only paths that matter. An index option does not knock out — it carries *front-end protection*, paying the losses incurred between today and expiry — and its value is then the Black term plus the value of that protection, which is not an option at all but a forward-starting exposure to the jump risk of §16.6.

The two are quoted in the same units and are not the same instrument. It is also the cleanest illustration in the book of a numeraire argument being contract-dependent: what makes the pricing work is a clause, not a theorem.

Constant maturity CDS. A contract paying a periodically reset spread of fixed tenor, rather than a fixed one. It is chapter 15's problem in credit: the rate is observed under one measure and paid under another, so it carries a convexity adjustment, and the same replication over the spread smile prices it under the same kind of mapping assumption.

Recovery locks and digital default swaps. This chapter's own identification problem was that λ and R enter the par spread only through their product, so no CDS quote separates them. A recovery lock trades R directly, and a digital swap pays a fixed amount on default rather than $1-R$. Where they trade, the identification failure stops being an assumption and becomes a market price — which is the honest resolution of a degeneracy, and rarer than one would like.

16.9 What One Name Cannot Tell You

Everything above concerns a single name, and for a single name the model is in good order: the intensity is identified up to the recovery convention, the curve bootstraps, and the two risks are separately measurable and separately hedgeable.

Portfolios are a different matter, and the reason is not a defect in the machinery so far but a question it never asked. A collateralised loan obligation, an index tranche, or the counterparty exposure of chapter 24 all depend on *how many* names default together, and nothing in a collection of single-name curves says anything about that. Each curve fixes its own marginal default distribution and the joint distribution is left entirely open — which is chapter 9's Gyongi problem in another costume: the marginals are pinned by the market and everything about the dependence is unconstrained.

Assuming independence is not neutral, it is a strong and badly wrong assumption. Defaults cluster: they share exposure to the economy, to funding conditions, and to each other through direct claims. A portfolio model that assumes independence will price the senior tranche of anything at nearly zero, because it assigns essentially no probability to many names failing at once, and that specific error has a history.

How dependence is modelled, what the industry standard gets wrong, and what to use instead is the subject of chapter 18.

References

- Duffie, D., & Singleton, K. J. (1999). Modeling term structures of defaultable bonds. *Review of Financial Studies*, 12(4), 687–720.

- Lando, D. (1998). On Cox processes and credit risky securities. *Review of Derivatives Research*, 2(2), 99–120.

- O'Kane, D. (2008). *Modelling Single-name and Multi-name Credit Derivatives*. Wiley.

Chapter 17

Foreign Exchange, Quantos and Inflation

Three markets that look unrelated on a trading floor are one piece of mathematics. An exchange rate is the price of one numeraire in units of another; a quanto is a payoff settled in the wrong numeraire; and inflation is foreign exchange with the real economy as the foreign country, the consumer price index as the exchange rate and index-linked bonds as the foreign bonds. One adjustment, $-\rho\sigma_1\sigma_2$, does the work in all three, and we derive it once and then spend the chapter recognising it. It ends by explaining a correction on year-on-year inflation swaps that is worth a dozen basis points at the long end and is invisible to anyone who has not made the identification.

17.1 An Exchange Rate Is a Ratio of Numeraires

Chapter 6 defined a numeraire as a strictly positive tradable that other prices are quoted against, and showed that each numeraire carries its own measure in which prices divided by it are martingales. Foreign exchange is what happens when there are two of them and they belong to different countries.

Let X_t be the spot exchange rate, quoted as units of domestic currency per unit of foreign. Let B^d_t and B^f_t be the two savings accounts, growing at the domestic and foreign short rates r_d and r_f.

Calculation 17.1 (The forward exchange rate)**.** A domestic investor can hold the foreign savings account, but only by first converting into it. What they hold, valued in their own currency, is $X_t B^f_t$ — and this is a domestic tradable, so it must have drift r_d under the domestic risk neutral measure. Since B^f grows deterministically at r_f,

$$\frac{d(X_t B^f_t)}{X_t B^f_t} = \frac{dX_t}{X_t} + r_f\,dt \stackrel{!}{=} r_d\,dt + \sigma_X\,dW^d_t, \tag{17.1}$$

which forces

$$\frac{dX_t}{X_t} = (r_d - r_f)\,dt + \sigma_X\,dW_t^d. \tag{17.2}$$

The exchange rate drifts at the interest rate differential. The forward rate does not need the dynamics at all. A foreign zero coupon bond, converted at spot, costs $X_0 P_f(0,T)$ domestic units today and is worth X_T domestic units at T: it is a domestic tradable that delivers the exchange rate. The forward strike is the price of receiving X_T, expressed in domestic units at T, so dividing by the price $P_d(0,T)$ of one domestic unit at T gives

$$F_X(0,T) = X_0 \frac{P_f(0,T)}{P_d(0,T)}. \tag{17.3}$$

In the language of chapter 6, the forward is the expectation of X_T in the domestic T-forward measure, the one whose numeraire is $P_d(t,T)$. The domestic tradable $X_t P_f(t,T)$ divided by that numeraire is a martingale in it, and at $t = T$ it equals X_T, so $\mathbb{E}^T[X_T] = X_0 P_f(0,T)/P_d(0,T)$. The volatility does not appear, because no distribution for X_T was used: the forward is fixed by two prices, and it holds whatever the rates do.

Remark (What this says in words)**.** If foreign interest rates are higher than domestic, the forward exchange rate is below the spot: the currency you earn more interest in is expected — under the pricing measure — to weaken by exactly the extra interest. It has to be, or borrowing domestically, converting, lending abroad and selling the proceeds forward would be a riskless profit. This is covered interest parity, and it is not an economic prediction. It is a no-arbitrage statement, enforced by anyone with access to both money markets.

That this is not an economic prediction deserves emphasis, because the historical record disagrees with it loudly: high-yielding currencies have not on average weakened by their interest differential, which is the entire basis of the carry trade. Chapter 6's distinction applies exactly here. (17.3) is a statement about $\mathbb{Q}$; the carry trade is a bet about $\mathbb{P}$; and the gap between them is a risk premium, not an arbitrage.

Remark (The symmetry, and what breaks it)**.** Nothing above singled out the domestic currency. The foreign investor sees the exchange rate $1/X_t$ and, by the same argument in reverse, finds it drifts at $r_f - r_d$ under *their* risk neutral measure. Both are right. The two measures differ, and $1/X$ under the foreign measure is not the reciprocal of X under the domestic one in any naive sense — by Itô's lemma the reciprocal of a lognormal picks up a $+\sigma_X^2$ in its drift, and that term is exactly the difference between the two measures.

This is Siegel's paradox, and it is the first sign that being careless about which currency one is standing in is expensive. The rest of the chapter is about the cases where it is expensive by an amount someone quotes.

Calculation 17.2 (Options on the exchange rate)**.** Since (17.2) is a geometric Brownian motion with drift $r_d - r_f$, everything in chapter 5 applies with r_f playing the part of a dividend yield: the foreign currency pays interest to whoever holds it, exactly as a stock pays dividends. The price of a call on the exchange rate is

$$C = P_d(0,T)\big[F_X N(d_1) - K N(d_2)\big], \qquad d_{1,2} = \frac{\ln(F_X/K) \pm \frac{1}{2}\sigma_X^2 T}{\sigma_X\sqrt{T}}, \tag{17.4}$$

which is Black-76 on the forward (17.3). Under its own name this is the Garman-Kohlhagen formula; it is not a new formula, so that everything already established about the Black world — the smile of chapter 9, the dynamics of chapter 10 — transfers without rederivation.

17.2 The Quanto Adjustment

Now the case where the two currencies genuinely interfere.

Definition 17.3 (Quanto)**.** A quanto is a derivative whose payoff is computed from a foreign asset but paid in domestic currency at a fixed exchange rate.

A typical example: a note paying the return on a foreign equity index, in dollars, with no currency exposure. The buyer wants the index and not the currency, and the quanto gives them exactly that. Someone has to price the separation.

Theorem 17.4 (The quanto drift adjustment)**.** *Let S_t be a foreign asset with volatility σ_S, let X_t be the exchange rate (domestic per foreign) with volatility σ_X, and let ρ be their correlation. Then under the* domestic *risk neutral measure,*

$$\frac{dS_t}{S_t} = \big(r_f - \rho\,\sigma_S\sigma_X\big)dt + \sigma_S\,dW_t, \tag{17.5}$$

and the quanto forward of S to time T is the ordinary forward multiplied by $e^{-\rho\sigma_S\sigma_X T}$.

Proof. S is a foreign tradable, so under the foreign risk neutral measure $\mathbb{Q}_f$ it drifts at r_f. To value a payoff settled at home we need its drift under $\mathbb{Q}_d$, so we need the change of measure between them, and the Radon-Nikodym derivative is fixed by the requirement that the converted asset X_tS_t be a domestic tradable. The quanto contract itself, paying S_T domestic units at T, is also a domestic tradable and drifts at r_d, but that gives nothing here: its price $P_d(t,T)\,\mathbb{E}^d_t[S_T]$ is defined through the law of S under $\mathbb{Q}_d$, so it holds for any drift S might have. The constraint comes from XS, which a domestic investor can buy today at an observed price, and which is a tradable whatever the drift of S turns out to be.

Apply Itô's product rule to XS and impose drift r_d under $\mathbb{Q}_d$. Writing μ_S for the unknown drift of S under $\mathbb{Q}_d$,

$$\frac{d(XS)}{XS} = \frac{dX}{X} + \frac{dS}{S} + \frac{dX}{X}\frac{dS}{S} = (r_d - r_f)dt + \mu_S\,dt + \rho\,\sigma_S\sigma_X\,dt + (\cdots)dW,$$

using (17.2) for the first term. Setting the total drift to r_d gives $\mu_S = r_f - \rho\sigma_S\sigma_X$, which is (17.5). Integrating a drift over $[0,T]$ multiplies the forward by its exponential. □

Remark (Why the cross term is the whole story)**.** The proof is three lines and every one of them is the Itô cross term $\frac{dX}{X}\frac{dS}{S}$.

Converting a foreign asset into domestic currency means multiplying two random quantities. If they tend to move together, the product moves more than either — you get more foreign currency exactly when each unit of it is worth more — and that extra covariance is worth something. A domestic investor holding the unhedged foreign asset receives it. A quanto buyer has been given the asset with the currency stripped out, so they do not receive it, and the price has to be lower by exactly that amount. Equation (17.5) is that statement with the accounting done.

Example 17.1 (The size of it)**.** A foreign index with 25% volatility, a currency with 10% volatility, correlation 0.5. The adjustment is

$$-\rho\sigma_S\sigma_X = -0.5 \times 0.25 \times 0.10 = -1.25\% \text{ per year},$$

so the one year quanto forward is 98.76 where the plain forward is 100.

A hundred and twenty five basis points a year is easy to dismiss on a short trade and impossible to dismiss on a long one, because it is a drift and compounds. At five years the forward is six percent below the plain one. This is the reason the adjustment is quoted as a drift rather than as a price: the number that matters is the rate, and the maturity does the rest.

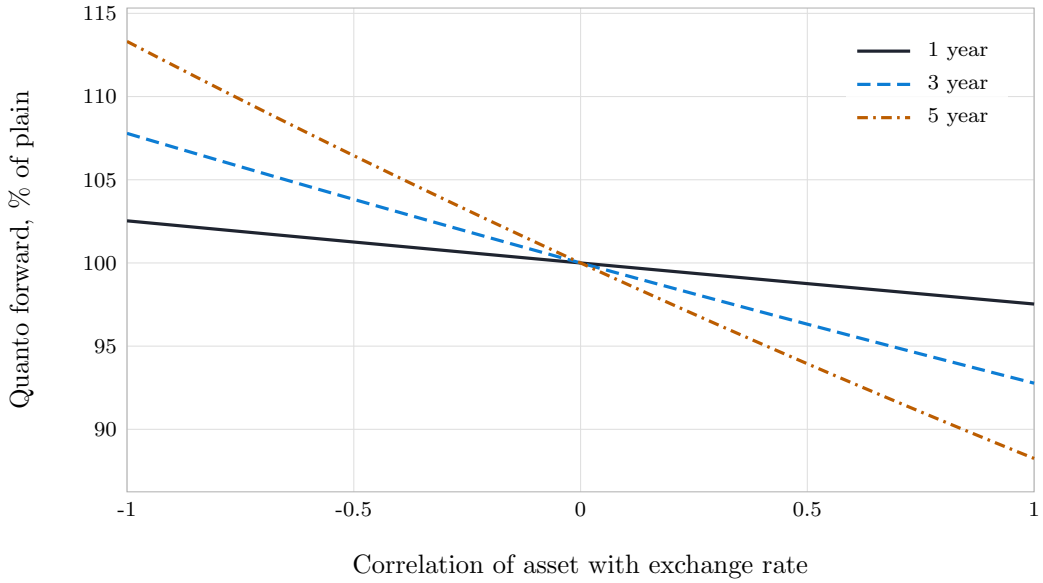

Figure 17.1: The quanto forward as a fraction of the plain forward, at a 25% asset volatility and a 10% currency volatility. It is monotone through zero and almost linear in the correlation, since the adjustment enters as an exponent and the exponent is small; and it fans out with maturity, because the adjustment is a drift. Nothing is wrong at $\rho = 0$: a payoff settled in a currency that does not move with it costs nothing to redenominate.
Drawn by `quanto_forward`.

Exercise. A quanto *option*, rather than a forward, on the same asset. Show that its price is the ordinary Black formula with the forward replaced by the quanto forward, and that the volatility is unchanged. (Hint: (17.5) changed the drift and not the diffusion coefficient.) Then explain why this means a quanto option's implied volatility is the same as the plain option's, even though its price is different — and why that makes the adjustment easy to hide in a trading system that only stores implied volatilities.

17.3 Inflation Is Foreign Exchange

An economy has two ways of measuring value. Nominal terms count currency. Real terms count goods — a basket, whose currency price is the consumer price index I_t. Then:

Foreign exchange	**Inflation**
domestic currency	nominal economy
foreign currency	real economy
exchange rate X_t	price index I_t
foreign short rate r_f	real short rate r
foreign bond $P_f(t,T)$	index-linked bond $P_r(t,T)$
domestic short rate r_d	nominal short rate n

The index I_t is the price of one unit of the real good in nominal currency, which is precisely what an exchange rate is. And an index-linked bond, which pays a principal scaled by realised inflation, is a bond denominated in goods: it pays one basket at maturity, converted at whatever the index then is. It is the foreign bond.

This identification is the Jarrow-Yildirim model. Everything in the first two sections now transfers with no new work.

Calculation 17.5 (The zero coupon inflation swap)**.** A zero coupon inflation swap pays $I_T/I_0 - (1+b)^T$ at T, and b is quoted as the breakeven inflation rate. Valuing it is (17.3) under a different name: $I_t P_r(t,T)$ is a nominal tradable, so the forward index is

$$\mathbb{E}^T[I_T] = I_0 \frac{P_r(0,T)}{P_n(0,T)}, \tag{17.6}$$

and the breakeven follows,

$$(1+b)^T = \frac{P_r(0,T)}{P_n(0,T)}. \tag{17.7}$$

Remark (There is no model in this)**.** Equation (17.7) contains no volatility, no correlation and no model. The breakeven is a ratio of two observable discount curves, and the trade that enforces it is static: buy the index-linked bond, sell the nominal bond, and hold both to maturity. It is the Fisher equation.

This matters because the breakeven is routinely described as "the market's inflation expectation", and it is not one. It is the expectation under the nominal T-forward measure, which differs from the real-world expectation by the covariance of inflation with the pricing kernel of chapter 6,

$$\mathbb{E}^T[I_T] \;=\; \mathbb{E}^{\mathbb{P}}[I_T] + \frac{\mathrm{Cov}^{\mathbb{P}}(M_T, I_T)}{\mathbb{E}^{\mathbb{P}}[M_T]}.$$

The kernel is the discount factor times the density that reweights the paths, so this covariance contains both how inflation lines up with the states investors want compensated and how it lines up with nominal rates. It has no fixed sign. If inflation tends to arrive in bad times, investors pay for protection and the breakeven sits above what they expect; if it arrives in good times, it sits

below. Which one holds is an empirical question and not a consequence of (17.7). The mechanism is the one behind the quanto adjustment, an expectation moved between measures by a covariance with the density that moves it, and the difference is what the covariance is made of: there it is $\rho\sigma_S\sigma_X$, which traded options pin down, and here it is not tied to any traded price. Chapter 6's $\mathbb{P}$ against $\mathbb{Q}$ distinction, for the third time.

Example 17.2 (A breakeven)**.** Ten year nominal rates at 3.0% and ten year real rates at 0.5%. The breakeven is

$$\frac{P_r(0,10)}{P_n(0,10)} = e^{(0.030-0.005)\times 10} = e^{0.25},$$

so $(1+b)^{10} = e^{0.25}$ and $b = e^{0.025} - 1 = 2.53\%$. The difference of the two continuously compounded rates is 2.5%; expressed as an annually compounded rate it is 2.53%. The three basis point gap is a compounding convention.

17.4 Where the Model Comes Back: Year-on-Year Swaps

The zero coupon swap needed no model. The year-on-year swap, which pays $I_{T_i}/I_{T_{i-1}} - 1$ on each of a series of dates, needs one.

Calculation 17.6 (Why the naive answer is wrong)**.** The obvious guess is that a year-on-year payment is the ratio of two zero coupon forwards. It is not, and conditioning shows where it breaks. Writing $S = T_{i-1}$ and $T = T_i$, and taking the expectation under the nominal T-forward measure in which the payment is valued,

$$\begin{aligned}
\mathbb{E}^T\Big[\frac{I_T}{I_S}\Big] &= \mathbb{E}^T\Big[\frac{1}{I_S}\mathbb{E}^T[I_T \mid \mathcal{F}_S]\Big] \\
&= \mathbb{E}^T\Big[\frac{1}{I_S}\cdot I_S\frac{P_r(S,T)}{P_n(S,T)}\Big] \\
&= \mathbb{E}^T\Big[\frac{P_r(S,T)}{P_n(S,T)}\Big], \qquad (17.8)
\end{aligned}$$

where the second line is (17.6) applied at time S instead of time 0, which is legitimate because $I_t P_r(t,T)/P_n(t,T)$ is a T-forward martingale.

The index has cancelled out entirely. What is left is the expectation of a *future* bond price ratio — and unlike a forward, that is not pinned by any static portfolio. Bond prices at S depend on where rates are at S, and (17.8) therefore depends on how they get there. A model is unavoidable.

Remark (The general shape of this)**.** The reader who has been through chapters 7 and 8 has met this twice already. A futures contract differs from a forward because it is settled under the risk neutral rather than the forward measure. A CMS payment differs from a swap rate because it is paid on the wrong date. Each time, an instrument that looks like a forward turns out to reference a rate under a measure in which that rate is not a martingale, and the gap is a convexity correction.

(17.8) is the same phenomenon and will produce the same kind of answer. What is new here is that part of the correction will turn out to be a quanto adjustment, which the other two cases have no analogue of, because they involve only one currency.

We now do the calculation. Take both short rates to be Hull-White as in chapter 8's Hull-White Model section, with mean reversions a_n, a_r and volatilities σ_n, σ_r, the index lognormal with volatility σ_I, and correlations ρ_{nr} between the two rates and ρ_{rI} between the real rate and the index.

Calculation 17.7 (The real rate under the nominal measure)**.** First, the piece that has no counterpart in an ordinary rates model. The real short rate is a foreign short rate, so its drift under the nominal measure is not the drift its own model was written with.

The argument is Theorem 17.4 again. $I_t P_r(t,T)$ must be a nominal tradable, and expanding by Itô's product rule as before,

$$\frac{d(IP_r)}{IP_r} = \underbrace{(n-r)dt}_{\text{index}} + \underbrace{\mu_{P_r}dt + (\cdots)dW}_{\text{real bond}} + \underbrace{\rho_{rI}\,\sigma_I\big(-B_r(t,T)\sigma_r\big)dt}_{\text{cross term}} \stackrel{!}{=} n\,dt + (\cdots)dW,$$

where $-B_r(t,T)\sigma_r$ is the real bond's volatility in the Hull-White model, $B_r(t,T) = (1-e^{-a_r(T-t)})/a_r$. Solving gives $\mu_{P_r} = r + \rho_{rI}\sigma_I\sigma_r B_r$, which is the real bond drifting faster than r under the nominal measure.

To translate that back to the short rate, let the real short rate's drift under $\mathbb{Q}_n$ be its own drift plus an unknown shift δ, so that $dr = (\theta_r - a_r r + \delta)\,dt + \sigma_r\,dW^{\mathbb{Q}_n}$. Then $\sigma_r\,dW^r = \sigma_r\,dW^{\mathbb{Q}_n} + \delta\,dt$, where W^r is the Brownian motion of the real economy's own measure, under which the real bond has $dP_r/P_r = r\,dt - B_r\sigma_r\,dW^r$. Substituting,

$$\frac{dP_r}{P_r} = (r - B_r\,\delta)\,dt - B_r\sigma_r\,dW^{\mathbb{Q}_n},$$

and matching the drift to μ_{P_r} gives $\delta = -\rho_{rI}\sigma_r\sigma_I$. The maturity has dropped out, since $B_r(t,T)$ multiplies both sides, so a single shift of the short rate accounts for the bonds of every maturity at once, which is what allows the adjustment to be attributed to the short rate and not to any bond. The short rate is therefore

$$dr_t = \big(\theta_r(t) - a_r r_t - \rho_{rI}\,\sigma_r\sigma_I\big)dt + \sigma_r\,dW_t^{\mathbb{Q}_n}. \tag{17.9}$$

The extra term is exactly (17.5), with the real rate in the place of the foreign asset and the price index in the place of the exchange rate.

Calculation 17.8 (The convexity)**.** With both rates Gaussian, (17.8) is an expectation of an exponential of a normal and can be done in closed form. Writing y^n, y^r for each rate's deviation from its own initial forward curve and B_n, B_r for the two Hull-White factors over $[S,T]$, the bond reconstitution formula of chapter 8's Hull-White Model section gives

$$\frac{P_r(S,T)}{P_n(S,T)} = \underbrace{\frac{P_r(0,T)\,P_n(0,S)}{P_r(0,S)\,P_n(0,T)}}_{\text{the naive answer}} \cdot e^{D_r - D_n} \cdot e^{-B_r y_S^r + B_n y_S^n},$$

with D_i the deterministic terms. Taking the expectation of the lognormal and using that $D_i = -\frac{1}{2}B_i^2 V_i$ exactly cancels half of each variance, the whole correction collapses to

$$C = \exp\Big(-B_r m_r + B_n m_n + B_n^2 V_n - B_r B_n \mathrm{Cov}(y_S^n, y_S^r)\Big), \tag{17.10}$$

where m_r, m_n are the two rates' means at S under the nominal T-forward measure and V_n is the nominal variance. The year-on-year payment is the naive ratio of forwards multiplied by C.

Remark (Reading (17.10))**.** The formula is opaque and the content is not. Every term is a mean or a covariance under the nominal T-forward measure, and the drifts that produce those means are three, listed in `Inflation::real_mean`:

- the Hull-White term each model carries to stay fitted to its own initial curve, which would be there in a single-currency world and contributes nothing interesting;
- $-\rho_{rI}\sigma_r\sigma_I$, the quanto adjustment (17.9), present only because the real economy is a foreign one;
- $-\rho_{nr}\sigma_n\sigma_r B_n(u,T)$, from changing to the nominal forward measure, which is the ordinary convexity of chapters 7 and 8.

So the correction has two parents. Setting $\rho_{rI} = 0$ — or $\sigma_I = 0$ — deletes the quanto term and leaves the rest untouched, and that is a sharp enough statement to check numerically, which `quant/src/crosscurrency.rs` does.

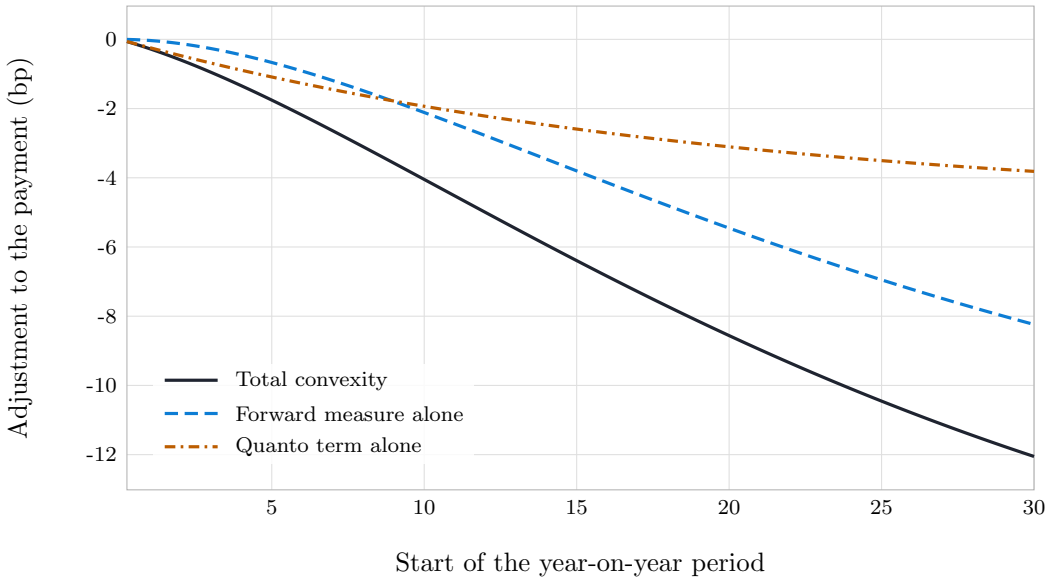

Figure 17.2: The correction a one year inflation payment needs, against the zero coupon curve, as a function of when the year starts. It grows with the start date because every term carries the accumulated effect of the measure changes up to that point, and the two components grow differently: the forward measure term is quadratic-ish in the start while the quanto term is closer to linear, so the quanto share falls from most of the correction at the front end to about a third of it at thirty years. Drag the correlation to zero and only one curve moves.
Drawn by `Inflation::yoy_convexity` and `Inflation::real_mean`.

Example 17.3 (Is it worth anything)**.** At the calibration in the figure — a 1.0% nominal volatility, 0.7% real, 1.2% index volatility, $\rho_{nr} = 0.6$, $\rho_{rI} = -0.3$ — and against a flat 2.53% breakeven:

Period starts	Total correction	Quanto part
1y	−0.3 bp	−0.3 bp
5y	−1.8 bp	−1.1 bp
10y	−4.2 bp	−2.0 bp
20y	−8.8 bp	−3.2 bp
30y	−12.4 bp	−3.9 bp

A basis point is nothing at the front end and a dozen of them at thirty years is not: on a long dated year-on-year book it is the difference between a position that makes money and one that does not.

Remark (Trusting this)**.** Equation (17.10) is exactly the kind of result that is easy to get wrong and hard to notice being wrong: a sign error inside it changes a number by a few basis points and nothing visibly breaks. So the implementation carries a second, independent route to the same answer — a Monte Carlo that steps both rates forward under (17.9) and averages the bond price ratio it finds, sharing none of the algebra above. The two agree across a range of correlations and volatilities to well inside the simulation error.

17.5 What Carries Over

Two things carry over from this chapter beyond the formulas.

The first is that a change of numeraire is a change of country, and that the cost of moving between them is always the same object: the covariance between the thing being valued and the exchange rate connecting the two numeraires. Written as $-\rho\sigma_S\sigma_X$ it looks like a quanto convention. It is not; it is the Itô cross term, and it appears wherever a quantity is valued in a measure other than the one it was defined in.

The second is that recognising two markets as the same market is worth more than any individual result derived here. The inflation desk and the foreign exchange desk sit on different floors and quote in different units, and (17.9) is invisible from inside either one. It becomes obvious — almost forced — the moment the index is seen as an exchange rate. Chapter 18 does the same thing again, for correlation.

References

- Garman, M. B., & Kohlhagen, S. W. (1983). Foreign currency option values. *Journal of International Money and Finance*, 2(3), 231–237.

- Jarrow, R., & Yildirim, Y. (2003). Pricing treasury inflation protected securities and related derivatives using an HJM model. *Journal of Financial and Quantitative Analysis*, 38(2), 337–358.

- Brigo, D., & Mercurio, F. (2006). *Interest Rate Models — Theory and Practice*, 2nd ed. Springer.

Chapter 18

Dependence

Chapter 16 left a gap: single-name curves pin every marginal default distribution and say nothing at all about the joint one. This chapter is about what to put in that gap. It builds the tools that describe dependence — Sklar's theorem, rank correlation, and the tail dependence coefficient that turns out to be the number most portfolio instruments are actually a bet on — and then gives four things to use, in increasing order of how much structure they demand. The Gaussian copula appears last, as a diagnosis rather than a target: it is a coherent model whose one defining property is that it sets to zero the exact quantity a senior tranche is priced off, and we finish by measuring what that costs.

18.1 The Gap, Stated Precisely

A basket of n names, each with a default time τ_i whose distribution F_i chapter 16 bootstrapped from that name's own CDS quotes. Anything depending on more than one name — a first-to-default swap, a tranche, the counterparty exposure of chapter 24 — depends on the joint distribution $F(t_1, \dots, t_n)$, and the marginals do not determine it.

The following theorem says exactly how much freedom that leaves.

Theorem 18.1 (Sklar)**.** *For any joint distribution F with marginals $F_1, \dots, F_n$ there is a function $C : [0,1]^n \to [0,1]$ with*

$$F(t_1, \dots, t_n) = C\big(F_1(t_1), \dots, F_n(t_n)\big), \tag{18.1}$$

and C is itself a distribution function on the unit cube with uniform marginals. It is unique wherever the marginals are continuous. Conversely, any such C combined with any marginals gives a valid joint distribution.

Proof. With continuous marginals the copula is not merely shown to exist, it is written down.

The probability integral transform is the whole idea: if F_i is continuous then $U_i = F_i(\tau_i)$ is uniform on $[0,1]$, since $\mathbb{P}(F_i(\tau_i) \le u) = \mathbb{P}(\tau_i \le F_i^{-1}(u)) = u$. So define

$$C(u_1, \dots, u_n) = F\big(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n)\big), \tag{18.2}$$

which is the joint distribution of $(U_1, \dots, U_n)$. It is a distribution function because F is; its marginals are uniform by the transform; and substituting $u_i = F_i(t_i)$ into (18.2) recovers (18.1). Uniqueness is immediate from the same substitution: (18.1) determines C at every point of the form $(F_1(t_1), \dots, F_n(t_n))$, and continuity of the marginals makes those points the whole cube.

The converse needs no work at all. Given any copula C and any marginals, the composition $C(F_1(t_1), \dots, F_n(t_n))$ is a composition of a distribution function with non-decreasing right-continuous maps, so it is again a distribution function, and setting all but one argument to their upper limits returns F_i. This half is what does the damage in this chapter: it says the construction never fails, so no choice of C can ever be ruled out by the marginals.

When a marginal has an atom the argument breaks at exactly one point — F_i is no longer invertible, and (18.1) constrains C only on the range of F_i, which now has gaps. Existence survives, by interpolating across the gaps, and uniqueness does not. That is the whole content of the continuity hypothesis, and it is why a default-time distribution with a lump of probability at a coupon date is a case where "the" copula is not well defined. □

Remark (What it means)**.** A joint distribution factors cleanly into two independent pieces: the marginals, which say how each name behaves on its own, and the copula C, which says how they are coupled. The two can be chosen separately, and every choice is legitimate.

The intuition is a change of coordinates. $F_i(\tau_i)$ is uniform on $[0,1]$ whatever τ_i was — this is the probability integral transform — so applying F_i to each name strips out everything specific to that name and leaves only its rank. The copula is the joint distribution of the ranks. It is what remains of the dependence after every marginal has been standardised away.

Sklar's theorem is usually presented as a technical result. For our purposes it is the precise statement of chapter 16's complaint: the market gives us the F_i and says nothing whatever about C, and by the converse half of the theorem, *every* C is consistent with the quotes. Choosing one is not calibration. It is an assumption, and it should be argued for rather than defaulted into.

Example 18.1 (The two extremes)**.** Two names, each defaulting within five years with probability 10%. What is the probability both do?

If they are independent, $C(u,v) = uv$ and the answer is 1%. If they are comonotone — one defaults exactly when the other does — then $C(u,v) = \min(u,v)$ and the answer is 10%. Both are consistent with identical CDS quotes on both names. A tenfold range in the joint probability, and nothing in the single-name market narrows it by a basis point.

18.2 Correlation Is the Wrong Word

The industry names this problem "correlation". That is a poor choice.

Remark (Correlation is not invariant)**.** Linear correlation is a property of the variables, not of their copula. Apply a strictly increasing function to one of them — take a log, or a square — and the copula is unchanged by construction, since ranks are unchanged, while the correlation moves. So correlation mixes up the dependence with the marginals, which is exactly the separation Theorem 18.1 was useful for making.

Rank statistics do not have this problem. Kendall's τ and Spearman's ρ depend only on the copula, and are the right things to quote if a single number must be quoted.

Theorem 18.2 (Attainable correlations)**.** *Let $X = e^{\sigma_1 Z_1}$ and $Y = e^{\sigma_2 Z_2}$, with Z_1 and Z_2 standard normal. Then whatever the joint law of (Z_1, Z_2),*

$$\frac{e^{-\sigma_1\sigma_2} - 1}{\sqrt{(e^{\sigma_1^2} - 1)(e^{\sigma_2^2} - 1)}} \;\le\; \mathit{Corr}(X, Y) \;\le\; \frac{e^{\sigma_1\sigma_2} - 1}{\sqrt{(e^{\sigma_1^2} - 1)(e^{\sigma_2^2} - 1)}}. \tag{18.3}$$

Proof. The marginals fix the means and variances, so only $\mathbb{E}[XY]$ varies with the coupling, and the correlation is increasing in it. The Hoeffding-Frechet bounds say the extremes of $\mathbb{E}[XY]$ over joint laws with given marginals are attained by the comonotone coupling, $C = \min(u, v)$, and the countermonotone one, $C = \max(u + v - 1, 0)$. For standard normals these are $Z_2 = Z_1$ and $Z_2 = -Z_1$: the first sends both to the same quantile, and $\Phi(-z) = 1 - \Phi(z)$ makes the second send them to opposite quantiles.

Now compute. For a standard normal Z, $\mathbb{E}[e^{sZ}] = e^{s^2/2}$, so

$$\mathbb{E}[X] = e^{\sigma_1^2/2}, \qquad \mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = e^{2\sigma_1^2} - e^{\sigma_1^2} = e^{\sigma_1^2}\big(e^{\sigma_1^2} - 1\big),$$

and likewise for Y. With $Z_2 = \varepsilon Z_1$ for $\varepsilon = \pm 1$, the product is $XY = e^{(\sigma_1 + \varepsilon\sigma_2)Z_1}$, so

$$\mathbb{E}[XY] = e^{(\sigma_1 + \varepsilon\sigma_2)^2/2} = e^{(\sigma_1^2 + \sigma_2^2)/2}\, e^{\varepsilon\sigma_1\sigma_2} \quad\Longrightarrow\quad \mathrm{Cov}(X, Y) = e^{(\sigma_1^2 + \sigma_2^2)/2}\big(e^{\varepsilon\sigma_1\sigma_2} - 1\big),$$

using $\mathbb{E}[X]\,\mathbb{E}[Y] = e^{(\sigma_1^2 + \sigma_2^2)/2}$. Dividing by $\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)} = e^{(\sigma_1^2 + \sigma_2^2)/2}\sqrt{(e^{\sigma_1^2} - 1)(e^{\sigma_2^2} - 1)}$, the exponential prefactor cancels and

$$\mathrm{Corr}(X, Y) = \frac{e^{\varepsilon\sigma_1\sigma_2} - 1}{\sqrt{(e^{\sigma_1^2} - 1)(e^{\sigma_2^2} - 1)}},$$

which is the upper end of (18.3) for $\varepsilon = +1$ and the lower end for $\varepsilon = -1$. □

Remark (The Gaussian copula's correlation is not the Pearson correlation)**.** The same computation gives the whole range and not just its ends. If $Z_2 = \rho Z_1 + \sqrt{1 - \rho^2}\, W$ with W an independent standard normal, which is the Gaussian copula with parameter ρ, then $\mathbb{E}[XY] = e^{(\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2)/2}$, and

$$\mathrm{Corr}(X, Y) \;=\; \frac{e^{\rho\sigma_1\sigma_2} - 1}{\sqrt{(e^{\sigma_1^2} - 1)(e^{\sigma_2^2} - 1)}}. \tag{18.4}$$

This is increasing in ρ, and $\rho = \pm 1$ are the two ends of (18.3). The parameter ρ of the Gaussian copula is always attainable and the joint law it defines is always valid. The Pearson correlation of the two lognormals it produces is (18.4), which is not ρ, and for volatile positions is far closer to

zero. Reading one number as the other is the mistake, and no joint law is impossible until that is done.

Inverting (18.4) turns a target Pearson correlation into a copula parameter, $\rho = \ln\big(1 + c\sqrt{(e^{\sigma_1^2}-1)(e^{\sigma_2^2}-1)}\big)/(\sigma_1\sigma_2)$, and this fails exactly when the target c is outside (18.3): the logarithm has no argument below the lower bound, and ρ leaves $[-1,1]$ above the upper one. It is the same bounds theorem, now stated for the copula parameter.[1]

Example 18.2 (The bounds are not decorative)**.** Evaluating (18.3) numerically:[2]

Volatilities	Lowest attainable	Highest attainable
25% and 25%	−0.94	+1.00
100% and 100%	−0.37	+1.00
200% and 200%	−0.02	+1.00
50% and 200%	−0.16	+0.44

Read the last two rows. Two lognormals at 200% volatility cannot be more than two percent negatively correlated no matter how they are coupled, because the closest they can come to opposed is e^{2Z} against e^{-2Z}, whose product is identically one. They are opposed in rank, and yet their correlation is $-e^{-\sigma^2} \approx -0.02$. Each has a mean of about 7.4 and a standard deviation of about 54, and that variance is carried by rare huge values, which occur on opposite sides: one is enormous only when the other is tiny, so the covariance $1 - e^4$ is small next to the variance $e^4(e^4-1)$ that the tails create.[3] And a 50% name and a 200% name cannot exceed 0.44 however tightly they are bound, because the fatter-tailed one has a variance the thinner one cannot track.

A system asked for a Pearson correlation of −0.5 between two 200% volatility positions has been asked for something no joint distribution can produce, and a careful one fails to calibrate. A system that takes the same −0.5 as the parameter of a Gaussian copula builds a valid joint law without complaint, and (18.4) says the Pearson correlation of the positions it implies is −0.016.[4]

Remark (And zero correlation is not independence)**.** The familiar one, included because it is the reason the next section exists. Correlation measures one particular linear summary of the joint law, and there are many ways to be strongly dependent with none of it. In particular, two variables can be uncorrelated and still almost always crash together, which is the case a portfolio of credits actually lives in.

18.3 The Number That Actually Matters

The failures above are diagnostic. This is the one that decides prices.

Definition 18.3 (Tail dependence)**.** The coefficient of upper tail dependence of a copula is

$$\lambda = \lim_{u \to 1^-} \mathbb{P}\big(U > u \mid V > u\big), \tag{18.5}$$

[1] simulating the coupling and measuring the Pearson correlation; the two ends and the monotonicity; the inversion, and where it fails.

[2] quant/src/dependence.rs.

[3] the closed form, and the mean and standard deviation quoted.

[4] the number, pinned.

where U, V are the two uniform coordinates. It is the probability that one variable is extreme, given that the other is, in the limit of extreme.

Remark (Why this and not correlation)**.** Consider what a senior tranche is. It absorbs nothing until the portfolio has already lost, say, fifteen percent, which requires a large fraction of the names to have failed together. Its entire value is the probability of a joint extreme. It is a bet on λ and on almost nothing else.

Correlation, by contrast, is an average over the whole distribution, dominated by the middle where most of the probability sits. Two copulas can agree on correlation to three decimal places, agree on every marginal exactly, and disagree about λ completely — and a senior tranche will price off the disagreement rather than the agreement.

Theorem 18.4 (Tail dependence of the two standard copulas)**.** *The Gaussian copula with correlation $\rho < 1$ has $\lambda = 0$. The Student-t copula with correlation ρ and ν degrees of freedom has*

$$\lambda = 2\, t_{\nu+1}\left(-\sqrt{\frac{(\nu+1)(1-\rho)}{1+\rho}} \right) > 0 \tag{18.6}$$

for every $\rho > -1$ and every finite ν.

Proof. Both halves come from the same question — conditional on one variable being far out in the tail, how far out is the other — and the two answers differ for one structural reason, which the calculation makes visible.

The reduction. Both copulas are unchanged by reflecting both coordinates, so the upper tail coefficient equals the lower one, and in terms of the standardised variables X, Y (normal or t, both symmetric),

$$\lambda = \lim_{x \to -\infty} \mathbb{P}(Y < x \mid X < x) = \lim_{x \to -\infty} \frac{\mathbb{P}(X < x,\, Y < x)}{\mathbb{P}(X < x)}.$$

Numerator and denominator both vanish, so by L'Hôpital the limit is the ratio of their derivatives in x. The denominator's is the density $f(x)$. The numerator's corner (x, x) moves in both coordinates,

$$\frac{d}{dx}\,\mathbb{P}(X < x,\, Y < x) = f(x)\,\mathbb{P}(Y < x \mid X = x) + f(x)\,\mathbb{P}(X < x \mid Y = x) = 2 f(x)\,\mathbb{P}(Y < x \mid X = x),$$

the two terms being equal because the pair is exchangeable. So $\lambda = 2 \lim_{x \to -\infty} \mathbb{P}(Y < x \mid X = x)$, a statement about the conditional law of Y given $X = x$.[5]

The Gaussian case. Write (X, Y) as standard bivariate normal with correlation ρ and condition on $X = x$. Then

$$Y = \rho x + \sqrt{1-\rho^2}\,\epsilon, \qquad \epsilon \sim N(0,1) \text{ independent of } X,$$

so λ is twice the limit of $\mathbb{P}(Y < x \mid X = x)$ as $x \to -\infty$, which is

$$\mathbb{P}\big(\epsilon < x\,\tfrac{1-\rho}{\sqrt{1-\rho^2}}\big) = \Phi\left(-|x| \sqrt{\frac{1-\rho}{1+\rho}} \right) \;\longrightarrow\; 0$$

[5] the joint tail by quadrature at $x = -1000$, against twice the conditional probability.

for any $\rho < 1$, since the argument goes to $-\infty$. The conditioning moved Y's mean out by ρx, but the residual ϵ has a fixed scale, and a fixed-scale Gaussian cannot keep up with a barrier receding at rate x. So the second name is dragged part of the way into the tail and left there. Only at $\rho = 1$ does the drag become complete.

The Student case. The t copula is the Gaussian construction with one change: the pair is divided by an independent random scale,

$$(X, Y) = \sqrt{\frac{\nu}{W}}\,(Z_1, Z_2), \qquad W \sim \chi^2_\nu,$$

with (Z_1, Z_2) Gaussian of correlation ρ. Now conditioning on X being extreme is informative in a way it was not before: a large $|X|$ is evidence not only about Z_1 but about W being small, and a small W inflates Y as well. The tail event is explained by the common factor rather than by the individual one, and the common factor moves both names together.

Making that quantitative needs the conditional law of Y given $X = x$, which for the bivariate t is again a t. Three steps show it.

What $X = x$ says about W. Given W, $X = \sqrt{\nu/W}\,Z_1$ is $N(0, \nu/W)$, so the joint density of (X, W) is proportional to

$$w^{\nu/2-1}e^{-w/2} \cdot \sqrt{w}\, e^{-wx^2/(2\nu)} \;=\; w^{(\nu+1)/2-1} \exp\Big(-\tfrac{w}{2}\big(1 + \tfrac{x^2}{\nu}\big)\Big).$$

As a function of w this is a gamma density, and it says that given $X = x$ the variable $W' = W(\nu + x^2)/\nu$ is $\chi^2_{\nu+1}$. So W has one more degree of freedom than before, and it is scaled by $\nu/(\nu + x^2)$: a large $|x|$ makes a small W likely, $\mathbb{E}[W \mid X = x] = \nu(\nu + 1)/(\nu + x^2)$.

What is left of Y. Given W and $X = x$, $Z_1 = x\sqrt{W/\nu}$ is fixed, so

$$Y = \sqrt{\tfrac{\nu}{W}}\big(\rho Z_1 + \sqrt{1 - \rho^2}\,\epsilon\big) = \rho x + \sqrt{1 - \rho^2}\,\sqrt{\tfrac{\nu}{W}}\,\epsilon,$$

where ϵ is standard normal and independent of (X, W).

Recognising a t. Substituting $\sqrt{\nu/W} = \sqrt{(\nu + x^2)/W'}$,

$$Y = \rho x + \sqrt{1 - \rho^2}\,\sqrt{\tfrac{\nu+x^2}{\nu+1}} \cdot \epsilon\sqrt{\tfrac{\nu+1}{W'}},$$

and $\epsilon\sqrt{(\nu + 1)/W'}$ is by definition a t variable with $\nu+1$ degrees of freedom, ϵ being independent of $W' \sim \chi^2_{\nu+1}$. So given $X = x$, Y is a $t_{\nu+1}$ variable with mean ρx and scale $\sqrt{1 - \rho^2}$ times the factor $\sqrt{(\nu + x^2)/(\nu + 1)}$. That factor is the information about W: it equals one at $x^2 = 1$ and grows like $|x|$, and as $\nu \to \infty$ it tends to one, which returns the Gaussian case's fixed scale $\sqrt{1 - \rho^2}$.[6]

Then

$$\mathbb{P}(Y < x \mid X = x) = t_{\nu+1}\left(\frac{(x - \rho x)\sqrt{\nu + 1}}{\sqrt{(1 - \rho^2)(\nu + x^2)}}\right),$$

[6] a simulation of the construction, keeping the draws whose X lies near -2.

and as $x \to -\infty$ the $\sqrt{\nu + x^2}$ in the denominator grows like $|x|$ and cancels the x in the numerator, leaving the finite limit

$$t_{\nu+1}\left(-\frac{(1-\rho)\sqrt{\nu+1}}{\sqrt{1-\rho^2}}\right) = t_{\nu+1}\left(-\sqrt{\frac{(\nu+1)(1-\rho)}{1+\rho}}\right).$$

Twice this limit is (18.6), the factor of two being the one from the reduction above. Note where the cancellation came from: the scale of the conditional distribution grew with $|x|$, and it grew because the mixing variable W is shared. That is the entire difference between the two copulas. □

Structure (The cancellation is the point)**.** The two calculations are the same calculation with one term changed, and the term is the scale of the conditional law. In the Gaussian case it is constant, so a barrier receding at rate $|x|$ eventually outruns it and the limit is zero. In the Student case it grows like $|x|$, the two rates match, and a finite limit survives.

The mechanism carries beyond copulas, and it is general: a fixed-scale residual cannot produce tail dependence, and a shared random scale always can. Any model in which the extreme behaviour of several quantities is driven by a common multiplicative factor — a stochastic volatility shared across names, a funding cost that hits every position, a liquidity parameter — has tail dependence for this reason, and any model whose only coupling is through the mean of a fixed-variance residual does not, whatever correlation is fed into it. Chapter 24's stress work is the practical form of the same observation: the scenario that matters is the one that moves the shared scale.

Remark (Read that again)**.** The Gaussian statement is not that tail dependence is small, or that it has been approximated away. It is exactly zero, for every correlation short of perfect. Condition on one name being in a one-in-a-million event and the probability that the other is too converges to zero. Joint extremes still occur; they become negligible next to single extremes.

The reason is regression to the mean. Given $X = x$, the most likely value of Y is ρx, closer to the centre, and reaching x takes a deviation of $(1-\rho)x$ from a residual whose scale $\sqrt{1-\rho^2}$ is fixed. The cost of that deviation is a probability of order $e^{-x^2(1-\rho)/(2(1+\rho))}$, which vanishes as x grows for every $\rho < 1$. The correlation controls how far along Y is expected to come, and never how far it can go in the tail.

The Student-t fixes it with one extra parameter and no extra conceptual machinery. A t is a normal divided by an independent random scale, and it is that shared scale that produces joint extremes: occasionally the whole system is drawn from a wide distribution, and then everything moves at once. This is not a mathematical trick. It is a common volatility factor, and it is what the world does.

18.4 Four Things To Use Instead

Now the constructive half. These are ordered by how much structure they ask for, and the right choice depends on what is being priced and on whether it has to be hedged.

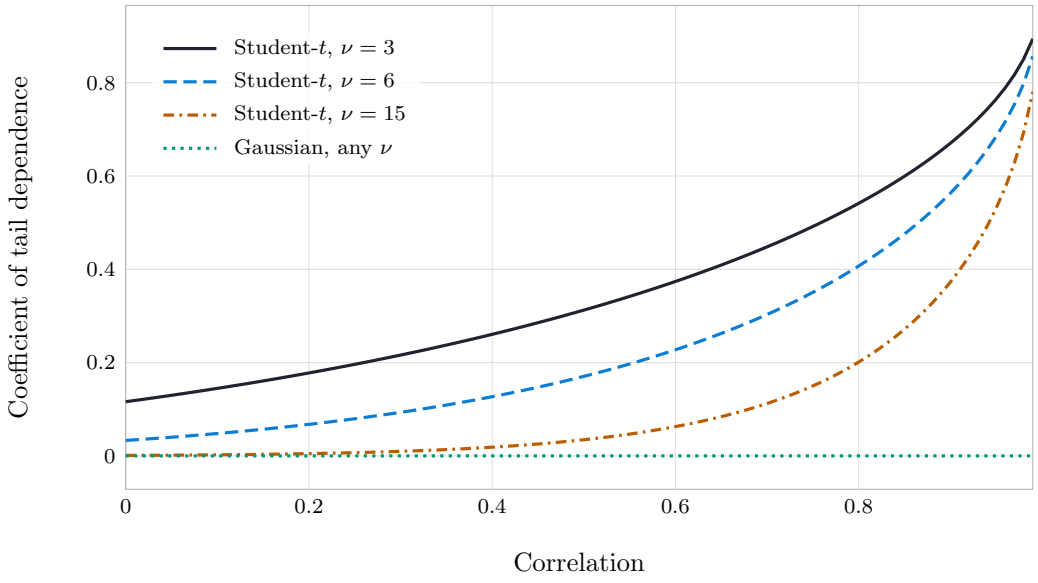

Figure 18.1: The tail dependence coefficient against correlation. The Gaussian curve is the horizontal axis — not close to it, on it, for every correlation up to one, where it jumps discontinuously to one. The Student-t curves sit well above and rise as the degrees of freedom fall. At a correlation of 0.3 a t with four degrees of freedom gives $\lambda = 0.16$: condition on one name being in an extreme event and there is a one in six chance the other is as well. The Gaussian model, calibrated to exactly the same correlation, says the chance is zero.

Drawn by `t_tail_dependence`.

18.4.1 Give the copula tails

The cheapest fix, and often enough. Replace the Gaussian copula by a Student-t with the same correlation matrix and one degrees-of-freedom parameter. Everything about the implementation is unchanged — the same factor structure, the same simulation, the same calibration to marginals — and λ goes from zero to (18.6).

The ν parameter can be calibrated to the tranche market, or set from history, and history gives two different answers depending on what is fitted. A t fitted by maximum likelihood to a single series measures the tail thickness of that series: EM at each ν for the location and scale, then a search over ν on the profile likelihood. That is not the parameter here. The copula's ν governs how two names move together in the tail, and it is fitted to the pair after replacing each series by its ranks, so that only the joint behaviour is scored — maximum pseudo-likelihood, with the copula density in place of the t density. On 11,241 daily changes in Treasury yields since 1981, the marginal fits give $\nu = 2.5$ for the two year and 4.1 for the ten year, and the copula of the two has $\nu = 5.1$ at a correlation of 0.82.[7] The marginals and the copula disagree, which is the reason to keep them apart. For equity indices, fits of a t are commonly reported with ν between three and six, and that range is the usual starting point when a t copula is set by convention; it is a rule of thumb and not something measured here, and a credit book's own data, or the tranche market, is the better source for its own ν.

Remark (Its limitation, stated honestly)**.** The t copula is symmetric: it produces joint booms as readily as joint crashes. Credit is not symmetric — names default together far more often than they are jointly upgraded — so a t that is fitted to the downside will overstate the upside. Whether that matters depends on whether anything in the book pays off on the upside. For a tranche, largely not.

18.4.2 Give it asymmetric tails

If the asymmetry matters, the Archimedean family provides it. A Clayton copula has lower tail dependence and none in the upper tail; a Gumbel copula the reverse. Each is a one-parameter family generated by a single convex function, which makes them easy to simulate and easy to reason about.

The cost is that Archimedean copulas in dimension n are exchangeable: they have one parameter for the whole basket and cannot express that two names in the same sector are more tightly coupled than two names in different ones. That is a serious limitation for a real portfolio, and the usual repair — nesting them hierarchically by sector — works but restores much of the complexity that made them attractive.

18.4.3 Model the cause, not the coupling

The first two are still copulas, and every copula shares a defect that the previous section did not mention because it is not about tails.

[7] `the three fits on the committed panel`; `recovery of a known` ν; `recovery of the copula's` ν`, with the marginals distorted by a monotone map`.

Remark (A copula has no dynamics)**.** A copula on default times is a static object. It gives the joint distribution of the τ_i as seen from today, and it says nothing about how single-name spreads move. The only thing it can update on is which names have defaulted, so between defaults its own prices are deterministic: there is jump-to-default risk in it, and no spread risk. There is no diffusion for a replication argument to cancel, and the dynamic hedging of chapter 5 has nothing to act on. The static content of chapter 4 still applies, since a copula can be arbitrage-free at a point in time, but the hedge ratio of a tranche against spread moves has no source in the model.

This is a deeper problem than the tails. A trader running a tranche book must hedge it with single-name CDS, and the hedge ratio has to come from somewhere. In a copula model it comes from bumping a static parameter and re-running, which is a finite difference of a quantity that was never claimed to be a price process.

The alternative is to model dependence where it comes from. Chapter 2 built default as the first jump of a Cox process with intensity λ_t; make the intensities share a factor,

$$\lambda_t^i = a_i\, Y_t + Z_t^i, \tag{18.7}$$

with Y a common process and Z^i idiosyncratic ones. Now everything is a process. Names default together because their intensities rise together; the joint distribution is an output rather than an assumption; and the model has a filtration, so a tranche has a genuine delta with respect to the single-name spreads that hedge it.

Remark (What this buys, and what it costs)**.** It buys dynamic consistency, which is the thing a copula cannot provide at any price. Spread moves and defaults are the same mechanism seen at different scales, which matches what a credit book actually experiences: correlated spread widening first, defaults later.

It costs calibration difficulty. Getting enough tail dependence out of (18.7) requires Y to be able to jump — a diffusive common intensity produces correlated spreads and puts far less loss in the far tail than a jump factor of the same mean and variance, which is the Gaussian problem in a new costume. So Y is usually taken with jumps, and the model is harder to fit than a copula. That is the trade, and for a book that has to be hedged rather than merely priced, it is usually worth making. The calculation below shows what fitting it involves.

Calculation 18.5 (Fitting the common-factor model)**.** Take $dY = -\kappa Y\, dt + dJ$ with J a compound Poisson process of rate ℓ and exponential jumps of mean μ, and write $B(t) = (1 - e^{-\kappa t})/\kappa$.

Only one integral matters. Take the idiosyncratic intensities Z^i deterministic. Given Y, name i survives to T with probability $\exp(-a_i\Lambda_T - \int_0^T Z^i)$, where $\Lambda_T = \int_0^T Y\, dt$, and the names are independent given Λ_T. So every joint quantity is an integral over the law of the single variable Λ_T. Integrating $Y_t = y_0 e^{-\kappa t} + \int_0^t e^{-\kappa(t-s)} dJ_s$ and exchanging the order of integration,

$$\Lambda_T = y_0 B(T) + \int_0^T B(T-s)\, dJ_s,$$

a sum over the jumps of $B(T-s)$ times the jump size.

Its transform is in closed form. For a Poisson process the exponential formula gives $\mathbb{E}\big[e^{-a\Lambda_T}\big] = \exp\big(-ay_0B(T) + \ell\int_0^T (\mathbb{E}[e^{-aB(s)X}] - 1)\, ds\big)$, and for an exponential jump X of mean μ, $\mathbb{E}[e^{-uX}] =$

$1/(1+\mu u)$. With $c = \mu a/\kappa$ the integrand is $1/\big((1+c) - c\,e^{-\kappa s}\big) - 1$, so

$$\mathbb{E}\big[e^{-a\Lambda_T}\big] = \exp\Big(- a y_0 B(T) + \ell\Big[\frac{\ln\big((1+c)e^{\kappa T} - c\big)}{\kappa(1+c)} - T\Big]\Big). \tag{18.8}$$

This is the affine structure of chapter 14 in the simplest case, and it is checked against an exact simulation of Λ_T, which is a Poisson number of jumps at uniform times and needs no time discretisation.[8]

What the jump is. It is in the common factor, so every intensity rises at once: a system-wide shock and not a default. Given the path of Y the names default independently, so a default in one name does not by itself move the others, and there is no direct contagion; all dependence is through common shocks. That is what reduces the model to the single variable Λ_T. Contagion in the direct sense, where a default makes the other names' intensities jump, needs a self-exciting term and breaks that reduction, and a default can also move the other spreads without any intensity jumping if the factor is hidden and the market learns about it from the default. This model has neither.

The first layer of the fit is the single-name curves. Because Y and Z^i are independent, $Q_i(T) = \mathbb{E}[e^{-a_i\Lambda_T}] \cdot e^{-\int_0^T Z^i}$. The first factor is (18.8) at a_i, so a deterministic function $Z^i(t)$, bootstrapped maturity by maturity exactly as the hazard rate is bootstrapped in chapter 16, reproduces each name's CDS curve for any values of κ, ℓ, μ and a_i. The single-name market is fitted exactly, as a Hull-White shift fits the rates curve, and it constrains nothing about the dependence.

The second layer is the tranches. Given Λ_T, a large homogeneous portfolio loses the deterministic fraction $(1-R)(1-e^{-a\Lambda_T - zT})$, so the expected loss of a tranche is $\mathbb{E}[f(\Lambda_T)]$ for a call spread f, an integral over the law of Λ_T. There is no closed form for it, which is the first way the model is harder than a copula: the law comes from inverting (18.8) or from sampling Λ_T as above, and the calibration is a least squares over that inner computation, with common random numbers so that the objective is smooth in the parameters. The parameters κ, ℓ, μ and a are chosen to minimise the squared error against the quoted tranches.

What identifies the parameters. The n-th cumulant of Λ_T is $\ell\, n!\, \mu^n \int_0^T B(s)^n\, ds$. The mean and variance fix $\ell\mu$ and $\ell\mu^2$ and so, given κ, both ℓ and μ, and every higher cumulant is then predicted. The equity and mezzanine tranches, which price the middle of the law, therefore pin the parameters, and the senior tranche, which prices the far tail, is a prediction of the model and not a free fit. A miss there is information about the model. A copula reaches each tranche with its own correlation and so fits every quote exactly, which is why it is easier to fit and why the fit says less.

Take $\kappa = 0.5$, $\ell = 0.4$ per year, $\mu = 0.05$, $y_0 = 0.01$ and $T = 5$, and compare with a diffusive factor, whose integral is Gaussian, of the same mean and variance of Λ_T. The mezzanine tranches, 3–10% and 10–20%, lose less under the jump factor, and the 20–60% tranche loses more than twice as much: jumps take loss out of the middle of the law and put it in the tail, with the first two moments unchanged.[9]

Remark (Contagion, modelled directly)**.** The common-factor model has no contagion: given the path of Y the names default independently. To put it in, add to each intensity a term that jumps

[8] `the transform at three loadings, and the mean and variance`.

[9] `the tranche loss ratios`.

when another name defaults and then fades, $\lambda_t^i = a_i Y_t + Z_t^i + \sum_{j \neq i} b_{ij}\, e^{-\delta(t-\tau_j)} \mathbf{1}_{\{\tau_j \le t\}}$. This is the infectious-defaults model of Davis and Lo, and the primary-secondary counterparty structure of Jarrow and Yu. A default now raises the other intensities at once, so defaults cluster and spreads jump when a default is observed, which the common factor can only imitate by jumping on its own.

The price is conditional independence: the intensities depend on each other's default histories, and the reduction to Λ_T is lost. What survives is the affine structure. For an exchangeable portfolio with an exponential kernel, the pair of the default count and the intensity is Markov, and the generator applied to $e^{uN+v\lambda}$ has a multiplier affine in λ, so the invariance test of chapter 14 passes and the transform of the loss solves a Riccati pair. Errais, Giesecke and Goldberg develop this for affine point processes. The two mechanisms are complements, shared shocks and feedback between defaults, and a book that wants both has to fit both from the same few tranche quotes.

18.4.4 Do not model dependence at all

The fourth option is the most interesting, and it is available whenever a liquid tranche market exists.

Notice what a tranche actually is. It absorbs portfolio loss L between a and d, so its payoff is

$$\frac{(L-a)^+ - (L-d)^+}{d-a}, \tag{18.9}$$

which is a call spread on a single scalar variable. The whole capital structure is a strip of call spreads on L — and by chapter 9's theorem Breeden-Litzenberger, a strip of call spreads on a variable determines that variable's distribution.

Remark (The same move as chapter 9)**.** This is the local volatility argument again. There we stopped trying to guess the dynamics of the underlying and read the risk-neutral density straight out of the option prices. Here we stop trying to guess the copula and read the loss distribution straight out of the tranche prices. In both cases the object that was being modelled turns out to be observable, and modelling it was never necessary.

The advantages transfer too, and so do the limitations. What is recovered is the loss distribution at each horizon and nothing more — the same partial information chapter 11 had to work around with local-stochastic volatility, and for the same reason. A bespoke tranche with different attachment points can be priced by interpolating the recovered distribution. A product depending on which names defaulted, or on the order they defaulted in, cannot: that is information about the copula and not about L, and it is not in the tranche prices.

Exercise. Show from (18.9) that the notional-weighted expected losses of any partition of $[0, 1-R]$ into tranches sum to the expected loss of the portfolio itself, whatever the dependence. (Hint: the call spreads telescope.) This is the credit analogue of put-call parity: a model-free constraint, which makes it the first thing to check in any implementation. `quant/src/dependence.rs` tests exactly this, for both copulas and three correlations.

18.5 The Gaussian Copula, and What It Cost

We can now say what went wrong precisely.

The Gaussian copula is not a broken model. It is a coherent, tractable, easily simulated dependence structure that reproduces any correlation matrix exactly, and for a great many purposes it is perfectly adequate. Its one structural property is Theorem 18.4: it sets tail dependence to zero.

It was then adopted as the market standard for pricing tranches, whose senior end is a pure bet on tail dependence. The single quantity those instruments were sensitive to was the single quantity the model asserted to be zero.

Example 18.3 (What the assumption is worth)**.** A large homogeneous portfolio, five percent of names defaulting by the horizon, forty percent recovery, thirty percent correlation. The same marginals, the same correlation, under a Gaussian copula and a Student-t with four degrees of freedom:[10]

Tranche	**Gaussian**	**Student-t**	**Ratio**
$0-3\%$ (equity)	54.11%	39.40%	0.73
$3-7\%$	19.58%	18.25%	0.93
$7-15\%$	5.83%	8.35%	1.43
$15-30\%$	0.81%	2.41%	2.98
$30-60\%$ (super senior)	0.019%	0.195%	10.16

Every number a risk report records is identical between the two columns: the same default probabilities, the same recovery, the same correlation. The super senior tranche is worth ten times more under one than the other.

And note the direction. The fat-tailed copula makes the equity tranche *safer* — it puts more probability at low losses as well as at high ones. So the effect is not a uniform increase in risk that a conservative haircut would have caught. It is a transfer of risk up the capital structure, and it is invisible at the bottom, where the market's attention was, and largest at the top, where the notional was.

Remark (Base correlation was the tell)**.** The market did notice, in the way markets do. Fitting one Gaussian correlation to all tranches simultaneously turned out to be impossible: each tranche implied a different one, and the market began quoting a "base correlation" per attachment point, rising steeply with seniority.

This is precisely chapter 9's volatility smile, and it means precisely the same thing. A parameter that the model says is a single number, quoted instead as a curve, is the market saying the model is wrong in a structured way and telling you the shape of the error. In the volatility case, the profession read the message and built chapters 10 to 12 in response. In the correlation case, the curve was largely treated as a quoting convention.

The lesson generalises past credit. When a model requires a different parameter value for each instrument, the parameter is not a parameter. It is a record of the model's error.

Remark (The fair verdict)**.** It is fashionable to blame the formula, and that is too easy — and it lets the actual mistake escape unexamined. The Gaussian copula did what it says. The failure was in the layer above it: an instrument was priced with a model that had been chosen for tractability

[10] every tranche in the table.

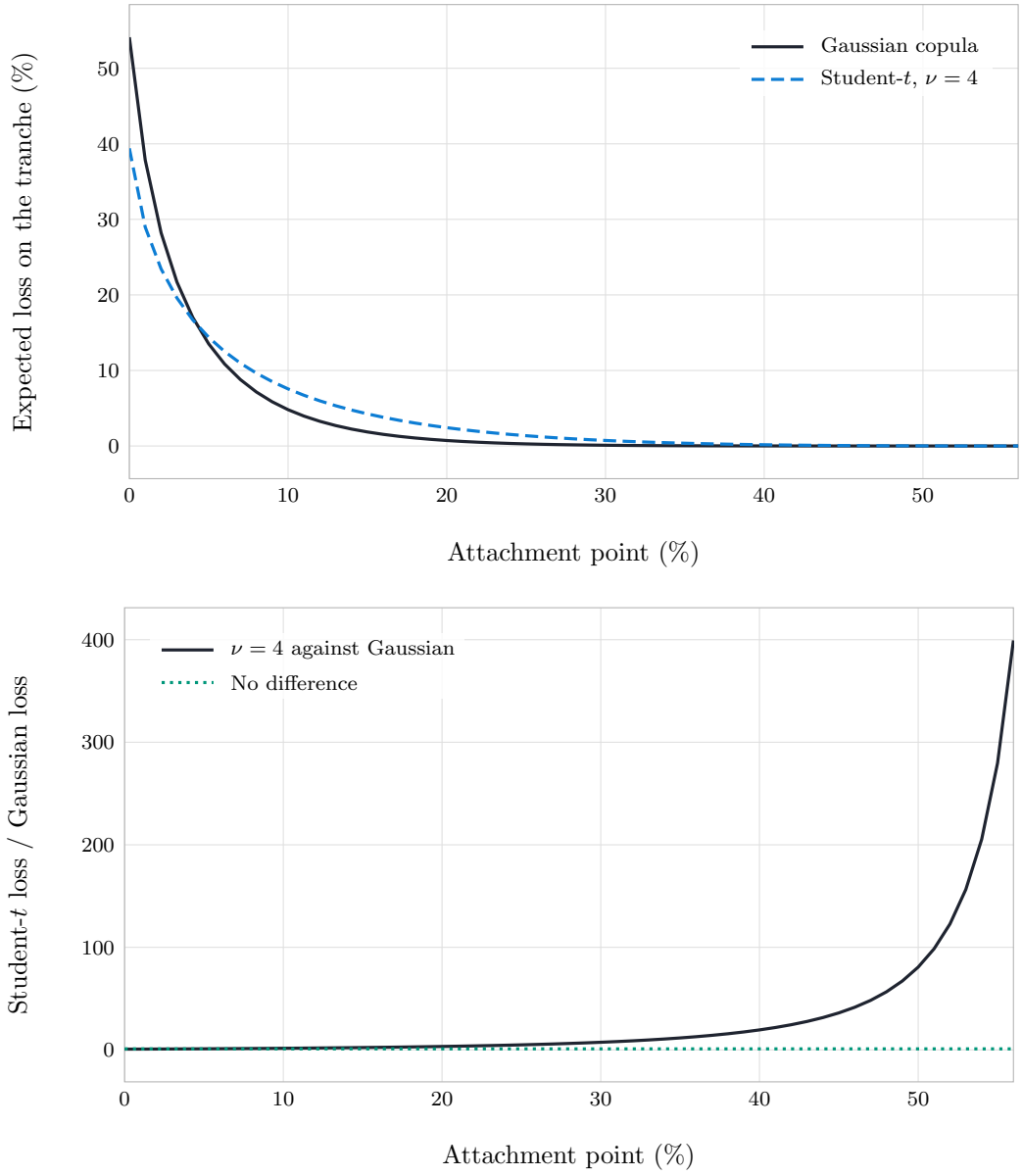

Figure 18.2: The same portfolio sliced into three point wide tranches at every attachment point. The upper panel shows the levels, which are dominated by the equity end and reveal nothing about the top. The lower panel shows the ratio of the two models, which reveals everything: below one at the bottom, crossing at around six percent attachment, and running away above it. A model chosen for its behaviour on the tranches that trade most was applied to the tranches that carried the most notional.

Drawn by `Portfolio::tranche_expected_loss` and `Portfolio::exceedance`.

and for its fit to the part of the capital structure that traded, and nobody asked what the senior end was sensitive to and whether the model had anything to say about it.

That question — what is this instrument's payoff a bet on, and does my model have a view on that quantity or has it assumed one — is not specific to credit, and it is the one thing to take away from this chapter.

18.6 The Same Gap in Rates

A CMS spread option pays on the difference between two swap rates of different tenors,

$$\big(S_1(T) - S_2(T) - K\big)^+, \tag{18.10}$$

and the callable range accruals and steepener notes that make up much of the structured rates market are strips of them. Chapter 12 met this payoff already, as the instrument that exposes what a one-factor model cannot do.

Now notice the shape of the pricing problem. Each rate's marginal distribution is *known*: its swaption smile gives the law of the rate under its own annuity measure by Breeden-Litzenberger, without any model of the curve, and chapter 15 shows how to carry that law to the measure of a payment date at the price of one assumption about the annuity mapping, not a model of the curve. What is not known is the joint distribution, and (18.10) depends on nothing else. That is exactly the position §18.1 opened this chapter with, in a different market: marginals supplied by the market, dependence supplied by the modeller.

So the market does exactly what credit did. It takes the two marginals from the smiles, joins them with a Gaussian copula carrying a single correlation, and integrates (18.10) against the result. And it has arrived at the same symptom: no single correlation fits the quoted spread options across strikes, so a correlation is quoted per strike — the base correlation of the previous section, in rates, for the same reason.

Calculation 18.6 (What the copula is worth on a spread option)**.** Set up as the tranche comparison was, so that nothing but the dependence differs. Two rates at 4% and 2.5%, both with an absolute volatility of 100 basis points, five years, correlation 0.8 — and both keeping exactly the same normal marginal under either copula.[11]

Against a Student-t with four degrees of freedom, the option struck at a spread of 1.5% is worth 0.95 times its Gaussian value. Struck at 4.5% it is worth 2.6 times.

Tail dependence makes extremes arrive together, and one expects two rates that move together to produce a narrower spread. What actually happens is that a Student-t copula is a bivariate normal divided by a common random scale, and a small draw of that scale makes both rates extreme *without making them equal*. The spread inherits the heavy tail. So the Gaussian copula understates a far strike spread option for precisely the reason it understated a senior tranche, and the mass it is missing there is the mass it has put in the middle.

[11] `measured`. The Gaussian column is checked against Bachelier, which it must reproduce, since normal marginals joined by a Gaussian copula give a normal spread.

Remark (Two assumptions, not one)**.** Rates adds a layer that credit does not have.

The marginals in credit come from CDS quotes with no modelling in between. The marginals here come from chapter 15, which reaches them only through the linear annuity mapping — an assumption in its own right, and one calibrated to different instruments than the copula is. A CMS spread option therefore rests on two dependence-free approximations stacked on each other: a mapping to get each rate's law under the payment measure, and a copula to join them. Neither is implied by the other and a reserve against one says nothing about the other.

A quanto CMS spread adds a third, since chapter 17's adjustment needs each rate's correlation with the exchange rate. The object required is then a three-dimensional joint law of which the market quotes only pairwise pieces, and the standard practice — assemble it from those pieces and hope the assembly is consistent — is not guaranteed to produce a valid joint distribution at all.

Calculation 18.7 (What "valid" means, and how it fails)**.** Standardise the three variables and read each as a unit vector in the $L^2(\Omega)$ of chapter 1. A correlation is the cosine of the angle between two of them, $\rho(X,Y) = \langle X, Y\rangle/(\|X\|\|Y\|)$, and any three vectors have a positive semi-definite Gram matrix of their pairwise inner products — $v^\top G v = \|\sum_i v_i u_i\|^2 \geq 0$ for every v. So three numbers can be the pairwise correlations of three real random variables only when

$$\begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}$$

is positive semi-definite, which for three variables is one inequality beyond $|\rho_{ij}| \leq 1$,

$$1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23} \;\geq\; 0. \tag{18.11}$$

Nothing in three separate calibrations enforces it. Take $\rho_{12} = 0.8$ for the two rates, as calibrated to the spread option, and $\rho_{13} = 0.7$, $\rho_{23} = -0.1$ for each rate's correlation with the exchange rate, as calibrated to a quanto adjustment on that rate alone — both plausible on their own, since the two rates are not the same variable and can disagree about their exchange rate exposure. The left side of (18.11) is -0.252. No real $(\text{rate}_1, \text{rate}_2, \text{FX})$ has this triple of correlations: the matrix has no square root, so a Gaussian copula built from it cannot be sampled, and nothing downstream of "assemble the pieces" was ever a joint distribution.[12]

Which leaves the choice this book keeps returning to. A term structure model of chapter 12 or chapter 13 supplies the joint law by construction and gives up the exact fit to the smiles; a copula keeps the exact fit to each marginal and supplies the joint law by assumption. Chapter 12's calculation How much a factor costs measured what the first costs on this very product. Calculation 18.6 measures what the second costs.

References

- Sklar, A. (1959). Fonctions de repartition a n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Universite de Paris*, 8, 229–231.

[12] this triple fails a Cholesky factorisation and a consistent one succeeds.

- Embrechts, P., McNeil, A., & Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. In *Risk Management: Value at Risk and Beyond*, Cambridge University Press, 176–223.

- Li, D. X. (2000). On default correlation: a copula function approach. *Journal of Fixed Income*, 9(4), 43–54.

- Duffie, D., & Garleanu, N. (2001). Risk and valuation of collateralized debt obligations. *Financial Analysts Journal*, 57(1), 41–59.

- Davis, M., & Lo, V. (2001). Infectious defaults. *Quantitative Finance*, 1(4), 382–387.

- Jarrow, R. A., & Yu, F. (2001). Counterparty risk and the pricing of defaultable securities. *Journal of Finance*, 56(5), 1765–1799.

- Errais, E., Giesecke, K., & Goldberg, L. R. (2010). Affine point processes and portfolio credit risk. *SIAM Journal on Financial Mathematics*, 1, 642–665.

- Frechet, M. (1951). Sur les tableaux de correlation dont les marges sont donnees. *Annales de l'Universite de Lyon, Section A*, 14, 53–77.

- Hoeffding, W. (1940). Massstabinvariante Korrelationstheorie. *Schriften des Mathematischen Instituts der Universitat Berlin*, 5, 179–233.

- Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd ed. Springer.

- McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative Risk Management*, revised ed. Princeton University Press.

Part IV

Numerical and Statistical Technique

Chapter 19

Numerical Methods

Almost every model in these notes was chosen for the same reason, and it was not realism. Affine dynamics give Riccati equations — closed-form bonds for Gaussian short rates, a characteristic function for Heston — SABR gives an expansion, and each of those buys a price in microseconds, which is what made calibration possible on the hardware the models were designed for. That constraint shaped the canon, and it has largely lifted. This chapter is about what replaced it: what a lattice and a simulation actually cost, why early exercise is where the difficulty concentrates, and what becomes available once a model is no longer required to have a formula. We measure the costs rather than assert them, including two biases in the standard method for early exercise that pull in opposite directions and are usually discussed as one.

19.1 The Constraint That Chose the Models

Model	**What it was built to give in closed form**
Affine models: Vasicek, Hull-White, Heston (ch. 14)	Bonds, or a characteristic function, via Riccati equations — bonds become swaptions by chapter 8's Jamshidian decomposition
SABR (ch. 10)	An asymptotic expansion for implied volatility
Quasi-Gaussian (ch. 12)	A finite Markov state, so a low-dimensional grid

None of these was selected because the world behaves that way. Each was selected because a calibration needs thousands of model prices, and a model that takes a second to price cannot be calibrated at all if a second is what one has for the whole exercise.

Remark (What the constraint cost)**.** The earlier chapters have already itemised the bill without presenting it as one. Chapter 10's local volatility model fits every European price and predicts the wrong dynamics. Chapter 10's SABR fits one expiry and is not consistent across expiries. Chapter 12's quasi-Gaussian models buy a finite state by insisting the volatility structure decay

exponentially, which nothing in the market requires. Chapter 18's Gaussian copula is the extreme case: chosen because it is easy to simulate and calibrate, and structurally unable to represent the one quantity it was used for.

In each case a functional form was adopted because it was solvable, and the market then declined to agree. That is not an argument against the models — they are still what one reaches for, and most of this chapter's alternatives are slower and less understood. It is an argument for knowing which of a model's properties are markets and which are convenience.

Structure (Which constraint is binding now)**.** The models of chapter 5 through chapter 12 were selected under a constraint: a price had to be computable by hand, or by a machine far weaker than the one on any desk today. That constraint chose affine dynamics, exponential volatility structures and one-factor curves, and it chose them for tractability rather than for realism — which is what chapter 14 spends a chapter making precise.

That constraint has lifted. What has not lifted, and what this chapter argues is now the binding one, is *identification*. A model can be solved numerically at almost any complexity; it cannot be told apart from its neighbours by the instruments the market quotes. These notes have measured that three times over — SABR's β against its correlation, local-stochastic volatility's mixing weight, and quasi-Gaussian's mean reversion — and in each case the flatness was a property of the quoted instruments rather than of the optimiser. Compute stopped binding and data started.

So the question is not whether a solvable model is still *necessary*. It is not. The question is what solvability was providing besides speed, and the answer turns out to be two things: prices that are arbitrage-free for structural rather than statistical reasons, and errors one can bound in advance. The next section is about the first and §19.7 about the second.

19.2 Two Ways to Compute

Everything that is not a closed form is one of two things, and their weaknesses are complementary in a way that decides most practical choices.

Lattices and grids

Backward induction on a tree, or a finite difference solution of the pricing equation of chapter 5. Chapter 9's Dupire forward equation can be run the other way for a whole strike surface at once.

The virtues are exactness and early exercise. There is no sampling error — refine the grid and the answer converges, at a rate that is second order in the spacing for a decent scheme — and early exercise is free, because backward induction already knows the continuation value when it needs to compare against intrinsic. That is why chapter 20's model selection table sends callables to short rate models: not because those models are more realistic, but because their small state fits on a grid.

The vice is dimension. A tensor grid with n points per axis in d dimensions has n^d points:

d	**points**	d	**points**
1	10^2	4	10^8
2	10^4	5	10^{10}
3	10^6	6	10^{12}

at a hundred points an axis. Three dimensions is comfortable, four is a project, and five is not happening. A basket over ten underlyings, or a market model with twenty forward rates, is not a large grid — it is no grid at all.

Simulation

Monte Carlo has exactly the opposite profile. Its error depends on the number of paths and not on the dimension, and it handles path dependence without any special effort because it has the path.

Calculation 19.1 (The rate, measured)**.** The standard error of a Monte Carlo estimate falls as $1/\sqrt{N}$. Pricing a one year European put in `quant/src/numerics.rs`:

Paths	**Standard error**	**Error** $\times\sqrt{N}$
10^3	0.349	11.02
10^4	0.107	10.67
10^5	0.0343	10.85
10^6	0.0109	10.89

The last column is flat across three decades, which is the rate confirmed rather than assumed.

The second half of that claim decides which method a problem gets. Doing it honestly needs an integral whose *dimension* can be varied while everything else is held still, so build one: drive a single lognormal with d independent normals combined into one,

$$Z = \frac{z_1 + \cdots + z_d}{\sqrt{d}}, \qquad F_T = F \exp\left(-\tfrac{1}{2}\sigma^2 T + \sigma\sqrt{T}\, Z\right). \tag{19.1}$$

Z is standard normal for every d, so a call on F_T has the same Black-76 price at every d, and its payoff has the same distribution — and therefore the same variance. Only the dimension of the integral changes. Anything that varies with d is then a property of the method and not of the problem.

Calculation 19.2 (The dimension, measured)**.** A one year at-the-money call, $F = K = 100$, $\sigma = 20\%$, worth 7.9656 at every dimension. Both methods are given the same budget of 10^5 payoff evaluations: the grid spends it as n^d nodes with n as large as that allows, the simulation as 10^5 paths.[1]

[1] `numerics::DimensionTest`.

d	**nodes an axis**	**grid error**	**simulation error**
1	100000	0.00000	0.025
2	316	0.00001	0.033
3	46	0.026	0.039
4	17	0.007	0.039
5	10	0.294	0.012
6	6	6.134	0.016
7	5	5.942	0.028

The grid is not merely better in one and two dimensions, it is in a different league — a deterministic rule converges in the mesh, not in the square root of a sample. By six dimensions it is wrong by three quarters of the price. The simulation column has no trend in it at all.

The middle rows wobble because n moves in integer steps and the payoff's kink lands between nodes differently each time; the collapse is the signal, not the ordering of $d = 3$ against $d = 4$.

The cliff has a location and a reason. The grid gets $n = \lfloor B^{1/d} \rfloor$ nodes an axis out of a budget B, so a budget spread over more axes leaves too few points on each to resolve anything: a hundred thousand evaluations is a hundred thousand nodes on one axis and five nodes on each of seven. Equivalently, to hold the mesh fixed the cost is n^d, which is the table at the top of this section read as an error rather than as a point count.

The simulation's flatness is exactly true here and only approximately true in practice. The standard error is $\sigma_{\text{payoff}}/\sqrt{N}$, and the dimension appears nowhere in it. Measured across $d = 1, 2, 4, 8, 16$ it is 0.0416 every time, to three figures. But (19.1) was built so that the payoff's own standard deviation is the same at every dimension, and that is the part a real problem does not give you: a basket over ten assets, or a market model with twenty rates, generally has a payoff whose variance grows with the number of factors. Dimension independence is a statement about the *exponent*. The constant is still the payoff's standard deviation, and that is a modelling quantity, not a numerical one.

Remark (What the rate means for a desk)**.** Four times the work to halve the error. A price good to a cent needs a hundred times the paths of one good to ten cents, and a Greek computed by bumping needs the difference of two such numbers, which is why Greeks are the expensive part of a simulation and why pathwise and adjoint methods exist.

Variance reduction — antithetics, control variates, importance sampling — improves the constant, sometimes by a large factor, and leaves the exponent alone. Quasi-Monte Carlo is the one technique that changes the rate, approaching $1/N$ when the effective dimension is low, which is a genuinely different proposition and worth reaching for when it applies.

Structure (Both methods are applying the same semigroup)**.** The two techniques look unrelated and are two ways of computing $P_t f = \mathbb{E}[f(X_t)]$, the semigroup of chapter 3.

A lattice or a grid discretises the *operator*: replace $\mathcal{L}$ by a matrix and $P_t = e^{t\mathcal{L}}$ by repeated multiplication. The error is a discretisation error, it is deterministic, and it shrinks with the mesh — and the matrix has one row per grid point, which is where the dimensional cost comes from.

A simulation samples the *measure* instead: draw from the law of X_t and average f. Where that law is exactly samplable — a lognormal terminal price, a noncentral chi-square draw for a CIR-type

variance — nothing is discretised at all, and the error is a sampling error that shrinks with the square root of the count. Most models and most payoffs need a path rather than one terminal draw, and then time is stepped, Euler or otherwise, which adds a second and deterministic error from that discretisation — the one the third discipline below, *correct the discretisation*, is about. Either way nothing costs n^d: a time grid is refined along one axis, in time, whatever d is, which is the asymmetry with the lattice that survives.

Written that way the trade is not a matter of taste. One method pays for accuracy in dimensions and the other in variance, and early exercise is hard for simulation for a structural reason: the exercise decision needs $P_t f$ at an intermediate state, which is exactly the object a lattice computes and a simulation does not.

19.3 Arbitrage-Free Becomes a Property of the Estimator

A closed form has a property that is easy to overlook because it is never stated: its prices are mutually consistent by construction. Black-Scholes cannot produce a negative butterfly, because it is an expectation under a measure, and an expectation of a convex function of a random variable is convex. There is nothing to check.

Once the expectation is estimated rather than evaluated, that guarantee has to be earned, and it is not automatic.

Calculation 19.3 (A butterfly priced two ways)**.** Take three strikes with K_2 midway between K_1 and K_3, and price each call by simulation. The payoff is convex in the strike, so for any single path

$$\big(S-K_1\big)^+ - 2\big(S-K_2\big)^+ + \big(S-K_3\big)^+ \;\geq\; 0. \tag{19.2}$$

Price the three on a *common* set of paths and (19.2) is inherited term by term: the estimated butterfly is an average of non-negative numbers and cannot be negative. Ten thousand paths — far too few for an accurate price — give exactly zero violations across every width tried.[2]

Price them on *independent* paths, which is what pricing each option separately amounts to, and nothing preserves it. Each price carries its own sampling error, the butterfly is a difference of nearly equal numbers, and at the same ten thousand paths over a quarter of narrow triples come out negative.[3] The price set admits a static arbitrage: one could buy the butterfly for a negative amount.

Remark (Why more paths is the wrong answer)**.** The instinct is to spend more. It does not work: sampling error falls as $N^{-1/2}$ while a butterfly's true value falls as the square of the strike spacing, so for any path budget there is a spacing at which the violations return.[4] Sixteen times the paths buys four times the resolution in strike, and a desk quoting a surface always wants more resolution than that.

Note also which direction the failure hides in. Widening the butterfly makes its value large against the noise and the violations disappear, so a check on widely spaced strikes passes while the

[2] `measured`.

[3] `measured`.

[4] `measured at two budgets and two widths`.

surface a desk actually shows — adjacent strikes, tightly spaced — fails. A validation that samples coarsely will not find this.

Structure (Three disciplines in place of one guarantee)**.** Calculation 19.3 generalises.

Arbitrage-freeness in a solvable model is a *structural* property: it follows from the price being an expectation under a measure. In a numerical model it is an *estimator* property, and it survives only if three things are done deliberately.

- *Price from a model, not from prices.* A surface interpolated or learned directly from quotes has no measure behind it and no reason to be convex in strike or monotone in maturity. A simulation of an arbitrage-free process does, whatever else is wrong with it.
- *Share the randomness.* Common random numbers across strikes, maturities and scenarios is what carries pathwise relations like (19.2) through to the estimates. The same discipline is what makes a bumped Greek meaningful, since a delta from two independent simulations is mostly noise.
- *Correct the discretisation.* A discrete scheme does not preserve the martingale property exactly, so numeraire-relative assets drift. Chapter 8 measures the consequence: a Gaussian forward curve model simulated with the wrong drift stops repricing its own initial curve, by five basis points at a year and half a per cent at ten. The repair is to force the known quantities to reprice, which is what a control variate or a martingale correction does.

None of the three is expensive. All three are easy to omit, and each omission produces prices that look like prices.

19.4 Early Exercise, Where the Difficulty Concentrates

The two methods divide the world neatly except at one point, where most of the practical difficulty lives.

Early exercise needs the continuation value: at each exercise date, the holder compares what the option is worth if kept against what it pays if stopped. A lattice knows the continuation value exactly, because it has already solved the layer after. A simulation does not, because it runs forwards and the future of a path is not known when the decision is made — and using it anyway is not merely inaccurate, it is a different and much more valuable contract, one exercised with foresight.

The standard resolution is Longstaff and Schwartz's: estimate the continuation value by regressing realised discounted cashflows on functions of the current state, across paths, and exercise where intrinsic beats the estimate. It works, it is what essentially every desk uses for Bermudans, and it is biased.

Remark (Two biases, not one)**.** They are usually discussed as though there were one, and they have different causes, different signs, and different remedies.

Foresight bias, upwards. If the same paths are used to fit the regression and to value the option, the exercise rule has seen the outcomes it is deciding about. It is fitted to the noise as well as to the

signal, and a rule fitted to noise flatters itself on the data it was fitted to. This is a finite sample effect and it decays as paths are added.

Policy bias, downwards. The estimated rule is not the optimal stopping rule — a quadratic in the spot cannot be — and any suboptimal rule undervalues the option. This has nothing to do with sample size. It is the basis's fault, and adding paths does not touch it.

The usual implementation contains both. Which one dominates depends on how many paths there are, and that is exactly what makes the situation confusing.

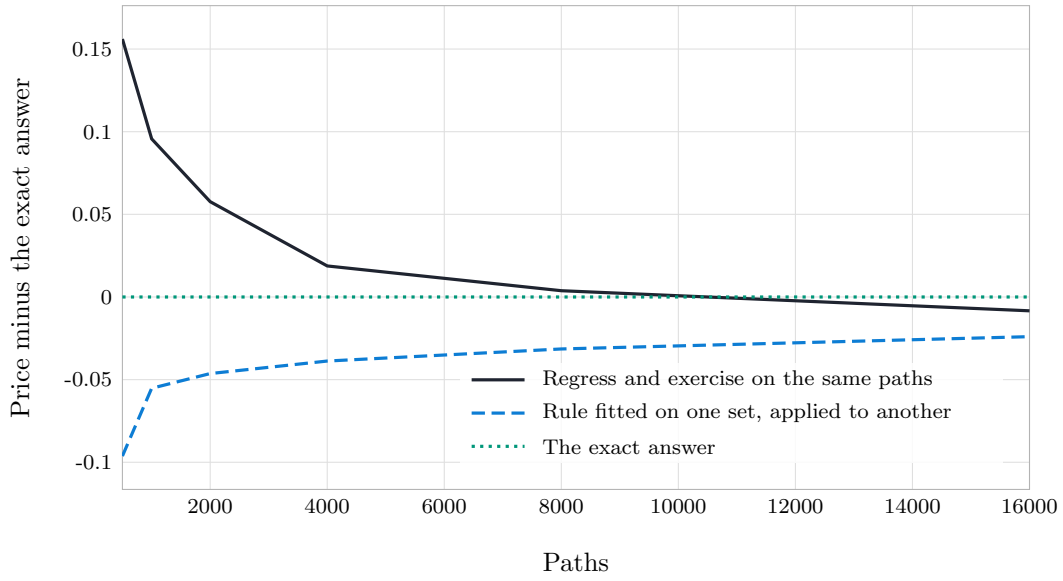

Figure 19.1: Longstaff-Schwartz against an exact lattice answer, averaged over two hundred and forty independent runs at each path count. The upper curve reuses its paths and starts fifteen cents high, falling through the exact answer as the foresight it enjoyed is priced out. The lower curve fits the rule on one set of paths and applies it to another, which removes the foresight and leaves only the loss from an imperfect exercise rule — so it stays below, and flattens rather than converging, because more paths cannot repair a basis that is too poor. The distance between the curves at any path count is the foresight; the distance from the lower curve to zero is the policy.
Drawn by `longstaff_schwartz` and `binomial_bermudan`.

Example 19.1 (The numbers)**.** A one year Bermudan put, ten exercise dates, struck at the money, at a 25% volatility. The lattice answer is 7.915 against a European value of 7.459, so early exercise is worth about forty-six cents. Averaged over two hundred and forty runs:

Paths	Same paths	Fresh paths
500	+0.156	−0.096
2,000	+0.058	−0.046
8,000	+0.004	−0.031
16,000	−0.008	−0.024

Read the two columns against each other. The left one falls by a factor of twenty and changes sign; the right one falls by a factor of four and is levelling off. That is the two biases behaving as their causes predict.

Remark (One run cannot see its own bias)**.** The practical warning, and the reason the table above needed two hundred and forty runs per entry rather than one.

At a thousand paths the bias is about ten cents and the standard deviation of a single estimate is several times that. So a single Longstaff-Schwartz run compared against a lattice tells you nothing about which way the method leans — the answer will be above or below according to the seed, and both happen. A desk that validates its Bermudan pricer by running it once and eyeballing the difference has not validated anything, and `the test that pins this` exists because it is easy to believe otherwise.

Remark (The harder problem is choosing what to regress on)**.** The biases above are measurable, which makes them the easy part. The difficulty that actually decides whether a Bermudan price is any good is upstream of them: what should the continuation value be regressed *on*?

The method needs variables in which the continuation value is a function of the current state — Markovian in them — and it needs a basis in which that function is close to linear. For a Bermudan put on one asset both are easy: the state is the spot and a quadratic is adequate, which is why the example above behaves. For anything real neither is obvious.

A Bermudan swaption in the market model of chapter 13 has twenty forward rates as its state. One cannot regress on twenty variables with any hope of conditioning, so the modeller picks summary variables — the underlying swap rate, perhaps a slope, perhaps the first two principal components — and hopes the continuation value depends on the rest only weakly. A callable range accrual depends on the whole path of accruals so far, and the natural summary, the accrual accumulated to date, is an extra state variable that must be carried and regressed on. A callable with a barrier depends on whether the barrier has been touched, which is not a function of the current level at all.

And the error this causes is invisible from inside the run. If a relevant state variable has been left out, the regression still fits, the fit statistics still look reasonable, and the price comes out low by an amount nothing in the calculation reveals. It is the policy bias of the previous section with a cause that adding paths cannot touch and adding basis functions cannot touch either, because the missing information is not in the variables being regressed on.

The only real defences are a dual upper bound, which brackets whatever the policy lost, and pricing the same trade with a genuinely different state — for instance on a lattice in a low-factor model — and treating the gap as the model risk it is. Which is chapter 20's discipline again, arriving from a numerical direction rather than a statistical one.

Remark (Neither is an upper bound)**.** A tempting conclusion from the figure is that the two curves bracket the answer. They do here, and it is not a bracket one can rely on: the upper curve's position depends on the path count, and at sixteen thousand paths it has already crossed below.

A genuine upper bound needs a different idea.

Fix the notation for what follows: exercise dates $t_1 < \cdots < t_K$, state X_{t_k}, discount factor D_k to t_k, intrinsic payoff h, and

$$V_0 \;=\; \sup_{\tau} \mathbb{E}\big[D_\tau\, h(X_\tau)\big], \tag{19.3}$$

the supremum over stopping times τ valued in $\{t_1, \dots, t_K\}$. Longstaff-Schwartz attacks (19.3) from below: any implementable rule is one particular τ, so its price is one term of the supremum and a lower bound. The dual formulation attacks it from the other side.

Theorem 19.4 (Weak duality: any martingale bounds it)**.** *For any martingale M adapted to the same filtration with $M_0 = 0$,*

$$V_0 \;\le\; \mathbb{E}\Big[\max_{1\le k\le K}\big(D_k\,h(X_{t_k}) - M_k\big)\Big]. \tag{19.4}$$

Proof. Optional stopping gives $\mathbb{E}[M_\tau] = 0$ for every stopping time τ valued in the finite set $\{t_1, \dots, t_K\}$, so

$$\mathbb{E}\big[D_\tau h(X_\tau)\big] \;=\; \mathbb{E}\big[D_\tau h(X_\tau) - M_\tau\big] \;\le\; \mathbb{E}\Big[\max_k\big(D_k h(X_{t_k}) - M_k\big)\Big],$$

the inequality because $D_\tau h(X_\tau) - M_\tau$ is one of the terms maximised over, whichever date τ happens to land on. The right side does not depend on τ, so taking the supremum over τ on the left proves (19.4) for that M — and the argument used nothing about M beyond $M_0 = 0$, so it holds for every martingale at once. □

Theorem 19.5 (Strong duality: the right martingale is exact)**.** *Let $Y_k = D_k V_k$ be the discounted value process, V_k the true continuation-or-exercise value at t_k — the Snell envelope of $D.h(X.)$, and V_0 its value at the start. Then Y is a supermartingale, $Y_{k-1} \ge \mathbb{E}[Y_k \mid \mathcal{F}_{k-1}]$, because continuing is by definition worth at least what any one choice, in particular waiting one more step and then following the optimal rule, would give. Its Doob decomposition $Y_k = Y_0 + M^*_k - A_k$, with M^* a martingale and A increasing, $A_0 = 0$, $M^*_0 = 0$, is*

$$M^*_k - M^*_{k-1} \;=\; Y_k - \mathbb{E}[Y_k \mid \mathcal{F}_{k-1}], \qquad A_k - A_{k-1} \;=\; Y_{k-1} - \mathbb{E}[Y_k \mid \mathcal{F}_{k-1}] \;\ge\; 0. \tag{19.5}$$

Equality holds in (19.4) *for $M = M^*$.*

Proof. V_k dominates the intrinsic payoff pointwise, $Y_k \ge D_k h(X_{t_k})$, since exercising is one of the choices continuation is worth at least as much as. So for every k,

$$D_k h(X_{t_k}) - M^*_k \;\le\; Y_k - M^*_k \;=\; Y_0 - A_k \;\le\; Y_0 \;=\; V_0,$$

using (19.5) for the middle equality and $A_k \ge 0$ for the last inequality. This holds pathwise, so the maximum over k is at most V_0, and taking expectations gives $\mathbb{E}[\max_k(\cdots)] \le V_0$. Theorem 19.4 gives the reverse inequality for this M^* in particular, so the two meet. □

Remark (What the two theorems divide between them)**.** Theorem 19.4 is cheap and universal: write down any martingale and get a valid bound, no matter how bad. Theorem 19.5 says the bound can be made exact, but the martingale that does it is built from V, the very function the whole chapter is about not having. Practice therefore substitutes an approximation for V in (19.5) and inherits both theorems separately: the bound stays valid by theorem 19.4, whatever the approximation, and stops being exact by theorem 19.5, to a degree that depends on how good the approximation is. That is the sense in which the passage's "rough" is doing real work: rough buys validity for free and tightness not at all, and the two have to be assessed separately.

Calculation 19.6 (Andersen and Broadie's construction)**.** Take $\hat{V}_k(x) = \max\big(h(x), \hat{C}_k(x)\big)$ for $k < K$ and $\hat{V}_K(x) = h(x)$, with $\hat{C}_k$ the Longstaff-Schwartz continuation-value fit at t_k — the same quadratic already sitting in the regression, nothing further done to it. Equation (19.5) applied to $\hat{Y}_k = D_k \hat{V}_k(X_{t_k})$ in place of Y_k defines a martingale $\hat{M}$ by the same formula, and theorem 19.4 makes it valid whether or not $\hat{V}$ is any good.

The one thing (19.5) needs that is not already in hand is the conditional expectation $\mathbb{E}[\hat{Y}_k \mid \mathcal{F}_{k-1}]$, and $\hat{V}$ has no closed form to take it in. Andersen and Broadie's move is to estimate it by simulation: from each outer path's state at t_{k-1}, draw a handful of *inner* one-step continuations to t_k, average $\hat{Y}_k$ over them, and use that average in place of the conditional mean. The inner draws are independent of the outer path's own step to t_k, which is what keeps $\mathbb{E}[\Delta \hat{M}_k] = 0$ exactly, whatever the inner sample size — more inner paths reduce the *variance* of the estimated bound, not a bias in it.[5]

Example 19.2 (The bracket, on the same put)**.** The Bermudan put of this section's earlier table, priced at the money for one year on ten exercise dates: the lattice value is 7.915. The Longstaff-Schwartz rule, fitted on one set of paths and priced out of sample on sixteen thousand fresh ones, gives a lower bound of 7.886. The martingale with no information in it at all, $M \equiv 0$, gives an upper bound of 12.370 — valid by theorem 19.4, and useless, since it is simply the value of stopping with perfect foresight of the whole path. The same quadratic fit that produces the lower bound, run through calculation 19.6, gives an upper bound of about 8.2: an order of magnitude closer than the martingale with nothing in it, from a martingale built out of a continuation-value estimate nobody trusted enough to call exact.[6]

7.886 and 8.2 are both computable without ever knowing the answer, and 7.915 sits inside them. That is what the passage means by converting "our Bermudan number is probably about right" into an interval: not a tighter point estimate, but a certificate that the true value cannot be outside a range, built from the same regression the point estimate already required.

19.5 What Lifting the Constraint Permits

Now the constructive half. If a model no longer needs a formula, what can it be?

Models with more parameters than quotes

Neural stochastic differential equations, market generators, and non-parametric local-stochastic volatility all abandon the idea that a model is a handful of interpretable numbers. The dynamics are specified by something with thousands or millions of parameters, fitted to whatever data is

[5] the construction, including the one detail that is not in the two-paragraph sketch above: $\hat{C}_k$ was fitted only on paths in the money, exactly as longstaff_schwartz restricts its own regression, and evaluating it out of the money — where the inner sampling and every date after the first go wandering regardless — extrapolates a quadratic fitted on one side of the strike to the other, which came out wrong by tens of dollars before it was floored at the payoff itself. The primal never notices this, since its exercise test starts from "intrinsic > 0" before it ever consults the continuation estimate; the dual evaluates $\hat{V}$ everywhere, so it notices immediately.

[6] the naive bound, and that it is loose; the fitted one, and the comparison; lower and upper measured together.

available, and priced by simulation because nothing else is possible. The first and third keep the form of a diffusion and learn its coefficients; a market generator gives that up too. It is typically a network trained — adversarially, or by some other distributional-matching objective — to output entire simulated paths directly from noise, fitted to reproduce a historical sample rather than to solve any equation at all.

Where the parameters are put matters more than how many there are, and whether a learned model is arbitrage-free depends entirely on it. A network fitted directly to a price surface, or a market generator fitted to reproduce simulated paths, has no structural guarantee: chapter 4 established that a price system without a martingale measure admits a money pump, and nothing in a training loss enforces one — a constraint can be added as a penalty, but a penalty is a preference and not a guarantee.

A network placed inside the coefficients of a stochastic differential equation is in a completely different position. Chapter 2 established that a continuous arbitrage-free price has no choice but to be an Itô diffusion, and chapter 4 that its discounted value must be a local martingale under the pricing measure — which fixes the drift and leaves only the diffusion coefficient free. So writing

$$\frac{dS_t}{S_t} = r\,dt + \sigma_\theta(t, S_t, v_t)\,dW_t \tag{19.6}$$

with σ_θ a neural network produces an arbitrage-free model for every value of θ, the constraint carried by the form of the equation rather than by the loss. Nothing has been given up in flexibility either: by chapter 2 the equation is not a restriction on the model class, it is the model class. The remaining requirements are mild — σ_θ positive and regular enough for a solution to exist — and the remaining problem is not this one: chapter 21 works out why a finite set of quotes still does not determine a function, however it is parametrised. It also explains why these methods are simulation methods necessarily: a learned coefficient admits no closed form by construction, so the model is defined by the equation and priced by solving it.

The calibration problem then changes shape rather than disappearing. What is needed is not a closed form but a *fast approximation of the map from parameters to prices*, which can be learned once, offline, and then evaluated in microseconds — so the model is free to be slow, provided its price map is smooth enough to be learned. Chapter 21 discusses that technique under deep calibration; here the point is what it unlocks, which is the freedom to choose dynamics on their merits.

Learning the exercise rule instead of the continuation value

The regression above estimates a continuation value and derives a policy from it. Reinforcement learning inverts that: the policy is the object being optimised, and the value is whatever the policy achieves.

This is a genuine improvement in one respect — a policy can be represented by something far richer than a quadratic, so the policy bias above can be made small — and it changes nothing structural. A learned policy is still a policy, so the estimate is still biased low, and it still needs a dual bound to be turned into an interval. The bias moves; it does not go away.

Optimising the hedge instead of the price

The methods above still price first and hedge afterwards. Deep hedging drops that order, and it is the one item on this list that changes what is being computed rather than how.

Fix a risk measure ρ — expected shortfall, from chapter 24, is the usual choice and the one its axioms recommend. Parametrise the hedging strategy directly as a function of whatever is observable, hand it a simulated market, and choose its parameters to minimise ρ of the terminal profit and loss. The price falls out as the cash that makes the optimised risk acceptable, which is an indifference price rather than a replication cost.

What that buys is everything the replication argument had to assume away. Transaction costs, position limits, discrete rebalancing and an incomplete market are not obstructions to be handled afterwards but terms in the objective, so the strategy is optimised against the market it will actually trade in. Chapter 5 measured what discrete rebalancing alone leaves behind — a residual whose spread falls only as the square root of the rebalancing count — and the delta it was rebalancing was optimal for a market with none of these features.

Two things it does not do, and both matter more than the machinery.

It does not produce a no-arbitrage price, and cannot. Replication is what pinned the Black-Scholes price down uniquely. Take away frictionless, continuous trading — with transaction costs, position limits or only finitely many rebalancing dates — and some payoffs are no longer attainable at zero residual risk. Chapter 6's structure The middle row and its theorem work out what that buys: no unique martingale measure, and so no single arbitrage-free price, only an interval bounded by the cost of super-replicating the payoff above and sub-replicating it below. The answer depends on the risk measure and on how much risk is being tolerated, so two desks with different objectives get different numbers and neither is wrong. That is not a defect of the method; it is what an incomplete market means, and that theorem said as much long before anyone trained a network. What the method supplies is a principled way to pick a point in the interval the theorem leaves open.

And it inherits the market it was trained on. A strategy optimised against simulated paths learns the measure that generated them, so every defect of the generator becomes a defect of the hedge — which is why the market generators of the first subsection are not a separate topic from this one. The calibration problem has been moved rather than removed, and it has been moved somewhere harder to inspect: a mis-specified parameter is visible, and a mis-specified simulator is a distribution nobody has plotted.

Remark (What a market generator is free to learn, and what that has to do with statistical arbitrage)**.** A market generator is fitted to history under the physical measure, with no no-arbitrage constraint built into its form the way a neural SDE's drift has one. So it is free to reproduce whatever predictable structure the historical sample contains — and real markets do contain some, which is not the contradiction it looks like. Chapter 4's theorem requires only that an equivalent martingale measure exist; it says nothing about the drift under the physical measure, which is free to carry a risk premium, or — chapter 22's term — a statistical arbitrage: a position with a genuine expected return that is not an arbitrage because it still holds a risk the model does not span. A generator that reproduces history faithfully reproduces that residual along with everything else.

What it cannot supply is chapter 22's other requirement: a structural reason. Finding that a generated path's conditional mean is not flat is a statistical fact; identifying why — a regulatory

flow, a financing constraint, a named mechanism — is not something a distributional match can do for you. That is the sense in which a market generator could point at a candidate statistical arbitrage without ever validating one, and it is a narrower and riskier use than the deep hedging above, which asks only that the generator's paths look like the market a hedge will actually trade in, not that any pattern found in them survives beyond the sample that produced it.

Solving high-dimensional equations directly

The dimensional wall is where the newer numerical methods have most clearly earned their place, and the leading one says something about pricing that the partial differential equation hides.

Definition 19.7 (The pricing problem as a backward equation)**.** A European claim's value Y_t and its hedge satisfy the backward stochastic differential equation

$$Y_t = g(X_T) + \int_t^T f\big(s, X_s, Y_s, Z_s\big)\, ds \;-\; \int_t^T Z_s \cdot dW_s, \tag{19.7}$$

where g is the payoff, f carries the discounting and any funding or default term, and Z_t is whatever integrand makes the equation hold.

Two features of (19.7) deserve emphasis. It runs *backwards* — the terminal condition is given and the initial value is the unknown — which is why it is the natural form of a pricing problem. And Z is not an auxiliary variable: by comparison with chapter 4's replication argument, $Z_t = \sigma^\top \nabla_x Y_t$ is the hedge. So (19.7) states the price and the hedging strategy as one object, which the pricing equation does only implicitly.

Remark (What a deep solver does with it)**.** The construction of E, Han and Jentzen turns (19.7) into a fitting problem, and does it by reading the equation forwards.

Parametrise the unknown initial value Y_0 as a number and the integrand $Z_t = \zeta_\theta(t, X_t)$ as a network. Then simulate X forward and propagate Y alongside it using (19.7) as a recursion,

$$Y_{t+\Delta} = Y_t - f\big(t, X_t, Y_t, Z_t\big)\Delta + Z_t \cdot \Delta W_t,$$

which requires no knowledge of the answer. At the horizon the equation demands $Y_T = g(X_T)$, and in general the propagated Y_T will not equal the payoff. So minimise

$$\mathbb{E}\Big[\big(Y_T - g(X_T)\big)^2\Big] \tag{19.8}$$

over Y_0 and θ. The loss is the terminal condition's failure, the fitted Y_0 is the price, and the fitted network is the hedge.

What has been bought is that nothing is discretised in the state. The cost is $O(\text{paths} \times \text{steps})$ regardless of dimension, which is why this works at fifty factors and a grid does not. What has been given up is the subject of the next section.

The honest test of such a solution is the same as for anything else in these notes: report the residual of the equation at points the fit never saw. A network that satisfies its equation only where it was trained has learned the training set.

States that are not vectors

Some problems have a state that is a graph rather than a list of numbers — a portfolio of counterparties with exposures between them, a limit order book, a network of related instruments. Graph neural networks are the natural representation, and this is the least mature of the strands here: the applications are real but the track record is short, and it is worth treating claims in this area with more suspicion than the rest.

Of the three, the order book is the furthest along. A market maker's quote depends on the whole depth of the book, not on the best bid and offer alone, and the book's natural representation is a graph: a node per price level, an edge to its neighbours, and a message-passing step that lets a shock several levels away — a large resting order cancelled, a burst of activity on one side — reach the level actually being quoted before it arrives there through price alone. What makes this the more advanced case is not the architecture, which is standard, but the feedback loop: a quote is filled or it is not, within milliseconds, so the model is trained against an outcome the market supplies immediately and constantly, unlike a counterparty network or a relative-value graph, where the analogous signal — a default, a reversion — arrives rarely and late. Chapter 23 is where a quote actually gets set from a state like this.

19.6 Why This Matters for Trading and Not Only for Pricing

A natural objection to everything above is that a model which cannot be calibrated in milliseconds cannot be used. That is true for a market making desk quoting a screen, and it is not true generally.

Chapter 22's trades do not need a model that prices in microseconds. They need a model whose view of the future is better than the market's, because the entire proposition is that the market's price implies something implausible and the position collects when it stops doing so. For that, being right matters and being fast does not.

Remark (The division of labour)**.** This suggests a use of models that the earlier chapters have not needed. A desk can reasonably run two: a tractable one for marking, quoting and hedging, where consistency with the market and speed are everything, and a slower and more honest one for deciding what to hold, where realism is everything.

They will disagree, and the disagreement is not a bug to be reconciled. It is the trade. Chapter 20's discipline applies directly — price it in two models that disagree about the thing the payoff is sensitive to — with the difference that here the disagreement is being sought rather than reserved against.

19.7 What Computation Does Not Fix

Two limits, both of which the earlier chapters have already established and neither of which more compute touches.

The first is identifiability. Chapter 20 measured a SABR calibration in which the quotes fixed the level of the smile to one percent and left the correlation free to move across most of its range — with the model exactly right and the data noiseless. That flatness is a property of the instruments

quoted, not of the fitting procedure, and a model with a million parameters fitted to forty quotes is that problem in its extreme form. A richer model does not extract information the market did not supply; it distributes the same information over more unknowns.

That is the complaint against fitting a single day's quotes specifically, and chapter 21's Calibrating to Prices and to History at Once is the escape a richer model can take that a small one cannot use as well: fit to the smile and to the price history at once, and by Girsanov the diffusion coefficient is shared between the two measures, so the history is a second, genuinely independent source of information about exactly the parameters a cross-section leaves flat. More parameters make room to use it — there is more diffusion structure to be pinned down and more history to pin it with. It is not a free upgrade: what the history identifies well is the diffusion coefficient and not the drift, so a slow parameter like a mean reversion still needs a long history rather than a rich model, and where the two sources disagree — implied against realised volatility routinely do — that disagreement is information a joint fit has to model rather than average away.

The second is that every estimate of an exercise policy is biased, by the argument of this chapter, and no amount of computation makes a policy optimal without knowing the optimal policy. Compute changes the size of the bias and the tightness of the bound. It does not remove either.

The third is the one this chapter's argument turns on, and it is not about accuracy but about *knowing* the accuracy.

Structure (Bounded and unbounded errors)**.** A grid's error can be bounded before it is run. A finite difference scheme of order p has error $O(h^p)$ with a constant depending on derivatives of the solution, so halving the mesh and watching the answer move by a factor of 2^p is a check that the bound is operative — and Richardson extrapolation turns two runs into an estimate of the error itself. The same is true of a Monte Carlo: the central limit theorem supplies a standard error from the sample, and it is a genuine confidence statement about the number produced.

A network fitted to (19.8) offers neither. The loss is a residual, and a small residual does not bound the error in the solution: the map from residual to error involves a stability constant for the equation that is not available in the high-dimensional problems the method exists to solve. Refining does not help, because there is no mesh to refine — one trains longer, or with more capacity, and the answer moves by an amount with no theory attached.

So the asymmetry is not that one method is more accurate. Over four dimensions the network is more accurate, sometimes by a wide margin. The asymmetry is that one method's error is a quantity and the other's is a hope, and that difference lands squarely on chapter 24: a valuation uncertainty has to be reserved against, and a reserve requires a number. An unbounded error cannot be reserved for; it can only be provisioned against by judgement, which is the same as saying it is carried as model risk.

That is the residual value of solvability. A solvable model is not more realistic. It is not faster in any way that matters now. What it has is that its answer comes with an error one can write down, and a desk that has to reserve against being wrong finds that this is worth something. The trade is realism for auditability, and stating it that way makes it a decision rather than a habit.

Remark (The rule that survives)**.** The methods in this chapter will date faster than anything else in these notes. What will not is the habit of asking, of any number produced by a numerical

method, three questions: what is its standard error, which way does it lean, and what would it look like if the method were wrong.

References

- Longstaff, F. A., & Schwartz, E. S. (2001). Valuing American options by simulation: a simple least-squares approach. *Review of Financial Studies*, 14(1), 113–147.

- Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Springer.

- Andersen, L., & Broadie, M. (2004). Primal-dual simulation algorithm for pricing multidimensional American options. *Management Science*, 50(9), 1222–1234.

- Rogers, L. C. G. (2002). Monte Carlo valuation of American options. *Mathematical Finance*, 12(3), 271–286.

- Tavella, D., & Randall, C. (2000). *Pricing Financial Instruments: The Finite Difference Method*. Wiley.

- Han, J., Jentzen, A., & E, W. (2018). Solving high-dimensional partial differential equations using deep learning. *PNAS*, 115(34), 8505–8510.

Chapter 20

Fitting and Testing

The chapters so far have built models and taken their parameters as given. This one asks where the numbers come from and how anyone would know a model is wrong. The two questions turn out to be different problems: fitting to today's prices is an inverse problem with no statistics in it, while fitting to a history is a statistical problem whose bias can be several times the truth. Both have the same failure, which is a fit that looks excellent while determining almost nothing, and both have the same remedy, which is to measure how far the answer can move before anyone would notice. We finish with the one general test of whether a model fits at all; chapter 21 continues into the harder case where the parameter itself is never observed.

20.1 Choosing the Model in the First Place

Before a parameter can be fitted, something has to be chosen to fit, and these notes have now built enough models that the choice is real. This section is the one a reader is entitled to ask for and most texts omit: given a trade, which model, and why.

There is a single question underneath it, and it is the one chapter 15 ended on.

Remark (The question that decides it)**.** *What is this payoff a bet on, and does the candidate model have a view on that quantity or has it assumed one?*

Every entry in the table below is an application of that question. A model is adequate when the things the payoff depends on are things the model was calibrated to; it is dangerous exactly when the payoff is sensitive to something the model fixed by assumption — which is chapter 18's Gaussian copula, chapter 9's local volatility, and chapter 10's sticky strike, all over again.

Does it need a model at all?

The first question is whether to reach for one, and surprisingly often the answer is no.

If the payoff is a function of one rate or price observed on one date, its distribution on that date is the whole of the relevant information, and the option market quotes that distribution. Static replication then prices it exactly, as chapter 15 did for constant maturity swaps and cash settled swaptions, and chapter 9's Breeden-Litzenberger did for anything written on a single equity fixing. No model is better than a model here, because a model can only add an assumption to information that was already complete.

The market quotes that distribution at finitely many strikes, not a continuum, so reading it off still needs an interpolation between them, and chapter 7's interpolation is a modelling choice finding applies here exactly as it did to the discount curve: schemes that reprice every quoted strike exactly can still disagree sharply on $g''(K)$ between them. What keeps this from being the model the sentence above excluded is chapter 9's remark that a parameterisation like SVI is not a model at all — it is a claim about today's shape and none whatever about how that shape moves, which is what a model always claims and a smile-fit never does. No-arbitrage narrows the interpolation without pinning it down, so the claim above is exactly true for a payoff whose replication weight sits inside the quoted range, and needs chapter 15's warning about the wings the moment it does not.

A model becomes necessary when the payoff depends on more than one date, on the path, on more than one underlying, or on when the holder chooses to act. Those four are the whole of it.

Which model, and what each one is really for

Model	**Reach for it when**	**It assumes away**
Black-Scholes / Black-76	Quoting and hedging a vanilla	Everything about the smile
Local volatility (ch. 9)	The payoff depends only on terminal distributions	Volatility dynamics: the smile is predicted to slide, wrongly
SABR (ch. 10)	Interpolating one expiry's smile	Any consistency between expiries
SVI	Interpolating one expiry's smile so that it can be differentiated	Dynamics entirely — it is a curve, not a process
Heston (ch. 10)	A consistent surface across expiries; forward volatility	Fit to any single expiry's wings
LSV (ch. 11)	Barriers and cliquets: exact vanilla fit *and* live dynamics	Nothing much, at the cost of being slow and fiddly to build
Hull-White (ch. 8)	Bermudans and callables where the smile is second order	The smile entirely, and any decorrelation along the curve
Quasi-Gaussian / Cheyette (ch. 12)	Callables that are also smile sensitive	Realism in exchange for a finite Markov state
Market models (ch. 13)	Products written on the actual quoted rates; flexible correlation	Tractability: it is a simulation, so callables need regression
Markov functional	Callables needing an exact vanilla fit and speed	Any control over the dynamics beyond the one factor

Remark (The row that is not a model)**.** One entry in that table is a different kind of thing and the difference is the point of including it. Every other row is a process: it says what the underlying

does, and a price follows. SVI — the stochastic volatility inspired parameterisation, which contains no stochastic volatility — says only what one expiry's smile *looks like*. It writes the total implied variance $w(k) = \sigma^2_{\text{imp}}(k)\,T$ at log-moneyness $k = \ln(K/F)$ as

$$w(k) = a + b\Big(\rho\,(k-m) + \sqrt{(k-m)^2 + s^2}\Big), \tag{20.1}$$

and stops. There is no underlying and no dynamics, so it cannot price a barrier and has nothing to say about tomorrow's smile.

The five parameters are readable. Vertical shift a is the level, horizontal shift m is where the vertex sits, s is how rounded the vertex is, b sets the size of the wings, and $\rho \in (-1, 1)$ tilts them: the wing slopes are $b(1-\rho)$ to the left and $b(1+\rho)$ to the right,[1] so a negative ρ is the equity skew.

Structure (Why a hyperbola)**.** Expression (20.1) looks like an arbitrary functional form and is less arbitrary than it looks. It is a hyperbola, written in coordinates that name its degrees of freedom, and the constraints on a smile point at that shape without quite forcing it.

A smile must do three things. It must be smooth and convex in the middle, since chapter 9 makes the second derivative in strike a probability density. It must grow at most linearly in each wing, which is Lee's moment formula: total variance rising faster than $2|k|$ would price deep options as though the underlying had moments it cannot have. And the two wings are free to grow at different rates, because skew exists.

Those constraints narrow the field sharply. They rule out anything with quadratic wings, anything obliged to be symmetric, anything not convex. What they do not do is determine a single curve: adding to any admissible smile a smooth convex bump that dies away at both ends leaves the asymptotes and the convexity exactly as they were, so there are infinitely many curves meeting all three conditions, and (20.1) is not the general solution of them.

What can be said is smaller and still stands. Among *conics*, the condition is decisive: an ellipse has no real asymptote and a parabola has one, so a conic with two distinct real asymptotes is a hyperbola and can be nothing else. So (20.1) is the lowest-degree algebraic curve that does what a smile has to do, and its five parameters are exactly the count those conditions suggest — two asymptotic slopes, two shifts, one radius of curvature at the vertex. That the count matches is an observation rather than a derivation, and the form remains a choice: it is the minimal one, made once and kept because it fits, not a consequence with a proof behind it.

Which also says exactly where it can go wrong. The whole of Lee's condition is that neither asymptote is steeper than two, or $b(1 + |\rho|) \le 2$, and it is a constraint on the fit rather than a property of the form. Violating it produces a smile whose implied density goes negative, and the density is what Dupire divides by — a fit outside the bound does not degrade the local volatility surface, it changes the sign of its denominator.[2]

Remark (The three families, and what actually separates them)**.** The table is long and the underlying taxonomy is short. Interest rate models differ in what they take as the state variable, and that choice, not the label, determines what each is good for.

Short rate models — Hull-White, and the quasi-Gaussian models of chapter 12 — carry a small Markov state. That makes them solvable on a lattice or a partial differential equation grid, which

[1] `svi::Svi::wing_slopes`, checked against the curve far out in each wing.

[2] `measured`, with the density obtained by differencing prices rather than from the parameters.

makes early exercise cheap, which is why callables are priced with them. The price of the small state is that the whole curve moves as a function of one or two numbers, so decorrelation between distant parts of the curve is not available.

Market models take the traded forward rates themselves as state variables. That is the honest choice — they model what is quoted, with as much correlation structure as one cares to specify — and it costs tractability, since the state is high-dimensional and simulation is the only route. Early exercise then needs a regression method, and every regression method is an approximation nobody can bound.

Markov functional models take neither, and instead impose that the state is one dimensional and choose the functional form to reproduce vanilla prices exactly. They are fast and they fit, and what one gives up is any independent control of the dynamics: having spent the freedom on the fit, there is none left for how the curve moves.

Choosing between them is choosing which of exact fit, rich dynamics, and cheap early exercise to give up, because no model in use gives all three.

Structure (How the functional form is chosen)**.** The third family was characterised and never built. Building it is one composition of quantile functions.

Work under the terminal measure, the one whose numeraire is the last discount bond in the structure. Take the single driver x_t driftless under it — a Brownian motion is the simplest choice, so $x_t \sim N(0,t)$ with a known, closed-form law. At each fixing date T_i the market supplies the marginal law of the rate L_i maturing there, via exactly the replication chapter 15 and chapter 9's Breeden-Litzenberger already built: call its quantile function F_i^{-1}. The model defines $L_i = \phi_i(x_{T_i})$, and asking ϕ_i to be increasing and to reproduce that one marginal exactly leaves nothing to choose: x_{T_i}'s own quantile is $\Phi(x/\sqrt{T_i})$, and the only increasing map sending it to F_i^{-1}'s quantile is the composition

$$\phi_i(x) \;=\; F_i^{-1}\Big(\Phi\big(x/\sqrt{T_i}\big)\Big). \tag{20.2}$$

Send the driver's quantile to the rate's, and there is exactly one monotone map that does it — the same construction chapter 18's copulas make from a uniform to a target marginal, run here on a market smile instead of a correlation assumption.

(20.2) says where the freedom went. Fitted at every date, ϕ_i leaves nothing over to also match a correlation between L_i and L_j: that joint law is already decided, entirely, by the law of (x_{T_i}, x_{T_j}), which was fixed the moment x was chosen to be a single driftless Brownian motion and before a single quote was matched. A market model buys correlation by adding factors; this model spent its only factor on the marginals and has none left to spend on how they move together.

This is a choice and not a formality. Φ is not picked independently; it is x's own CDF, forced by whichever law x was given, and (20.2) reproduces the target marginal whatever that law is — so x's dynamics are the one thing the fit never touches, and so the only place left to shape the joint law across dates. Giving x a time-dependent volatility $\eta(t)$, still no function of x itself, moves the autocorrelation of x_{T_i} and x_{T_j} and so the decorrelation along the curve, at no cost to the fit at any single date. Letting η depend on x buys nothing the fit needed and costs the one thing the construction is for: x stops being Gaussian, Φ stops being closed form, and finding it means solving x's own forward equation, which is the numerical cost a lattice model exists to avoid.

Example 20.1 (The flat-smile case is the model everyone already knows)**.** Take the target marginal itself lognormal, forward F and volatility σ: $F_i^{-1}(p) = F\exp\left(\sigma\sqrt{T_i}\,\Phi^{-1}(p) - \frac{1}{2}\sigma^2 T_i\right)$. Substituting into (20.2), the quantile functions cancel against each other's inverse and leave

$$\phi_i(x) \;=\; F\exp\left(\sigma x - \tfrac{1}{2}\sigma^2 T_i\right),$$

the ordinary driftless lognormal transform of the same Brownian motion. A flat smile is the special case in which the general construction reduces to Black-76 driven by one factor, which is the check that (20.2) is the right composition and not merely a plausible one.[3]

Because every quantity is a function of (t, x_t) and x is one-dimensional and Markov, early exercise is exactly the lattice calculation chapter 8's Hull-White and chapter 12's models already do — the same mechanism, on a driver whose only job is to carry the fitted marginals rather than to mean-revert like a rate. There is no rate to interpolate between fixing dates, because ϕ_i is only ever asked for at T_i; what is needed at an intermediate t is some other quantity's value there, and chapter 6's remark that a rebased price is a martingale in x_t answers it directly, by the same backward equation and the same known transition density that supply Φ at each date in the first place.

Example 20.2 (Four trades, four answers)**.**

- *A CMS cap.* One rate, one date per caplet. No model — replicate over the swaption smile, chapter 15. Reaching for Hull-White here obtains a worse answer more slowly, because the model's own smile replaces the market's.

- *A Bermudan swaption.* Several exercise dates, so the joint distribution matters and no set of European quotes supplies it. A short rate model, because early exercise is the binding constraint and a lattice is what makes it cheap.

- *An equity barrier.* The payoff depends on the path, and on the smile at every level the barrier might be crossed at. Local-stochastic volatility, chapter 11, because a pure local volatility model prices the barrier off dynamics it got wrong and a pure stochastic volatility model does not fit the vanillas it will be hedged with.

- *A callable range accrual.* Path dependent, smile sensitive and callable at once, which is the intersection that has no comfortable answer. Quasi-Gaussian if the callability dominates, a market model with regression if the smile does, and in either case a reserve against the difference between the two prices — which is the honest measure of the model risk, and larger than most desks admit.

Remark (The rule that survives when the table does not)**.** Models come and go and the table above will date. What does not is the discipline: price the trade in two models that disagree about the thing the payoff is most sensitive to, and reserve the difference. If the two agree, the choice did not matter and the reserve is zero. If they disagree, the gap is the part of the price that was never determined by the market — and it is better held as a number on the balance sheet than discovered later.

The rest of this chapter is about what to do once a model is chosen. This section was about not choosing one for a trade that never needed it.

[3] the two routes, checked against each other rather than against a third source, since one is definitionally what the other collapses to.

20.2 Two Jobs With One Name

A desk says "calibration" for both of the following, and they have almost nothing in common.

Definition 20.1 (Calibration)**.** Choosing parameters so that a model reproduces today's quoted prices. Cross-sectional, done under $\mathbb{Q}$, repeated every morning.

Definition 20.2 (Estimation)**.** Choosing parameters so that a model matches the behaviour of a historical time series. Longitudinal, done under $\mathbb{P}$, and a question of statistical inference.

Chapter 6 drew the measure distinction and this is where it bites hardest. Calibration has *no sampling error* — there is no sample. The quotes are what they are, and if the model reproduces them there is nothing left to be uncertain about in the statistical sense. Its difficulty is entirely that the map from parameters to prices may not be invertible.

Estimation has the opposite profile. The model may be inverted perfectly well; the trouble is that a finite history is a finite sample, and the estimator built on it can be badly biased in a direction nobody checks.

Remark (Why the confusion is expensive)**.** A parameter fitted one way is routinely used as though it had been fitted the other. Correlation is the standard offender: a correlation estimated from history is a $\mathbb{P}$ quantity, a correlation calibrated to quanto or spread option prices is a $\mathbb{Q}$ quantity, and they differ by a risk premium exactly as chapter 17's forward differed from the expected spot. Substituting one for the other is not an approximation. It is a category error that happens to produce a number.

20.3 A Perfect Fit That Determines Nothing

Take the SABR model of chapter 10 and the cleanest possible test of whether calibration recovers parameters: generate the quotes *from a known SABR model*, so the model is exactly right and there is no noise, no bid-offer and no misspecification. If the parameters cannot be recovered here, they cannot be recovered anywhere.

A five year expiry, a forward at 3%, seven strikes spanning two percent either side, and $(\alpha, \rho, \nu) = (0.30, -0.30, 0.40)$ with β fixed at 0.5. Starting the optimiser deliberately far away, at $(0.15, +0.40, 1.20)$, it converges and reproduces every quote to within 10^{-5} volatility points. A perfect fit.

Now ask the only question that matters: how far can each parameter be moved, re-optimising the others, before the fit degrades by a tenth of a volatility point — an amount several times smaller than any real bid-offer?[4]

Parameter	Fitted	Range that still fits	Width
α	0.300	[0.295, 0.304]	0.008
ρ	−0.300	[−0.999, −0.276]	0.72
ν	0.400	[0.070, 0.647]	0.58

[4] α against ρ, and the width of ρ.

The level parameter α is pinned to about one percent of itself. The correlation is not pinned at all: it can be anything from -0.28 to -1.0, and every one of those values reproduces the whole smile to inside a tenth of a point. The volatility of volatility can be anywhere in a nine-fold range.

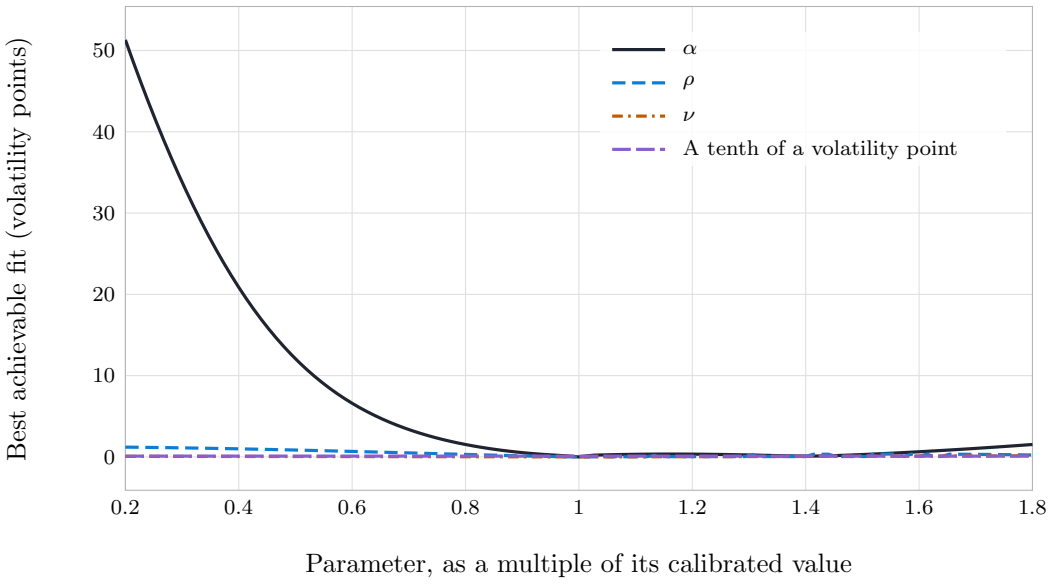

Figure 20.1: Each curve holds one parameter away from its calibrated value and re-optimises the other two, so it shows the best fit still achievable. The shape is the answer. α has a sharp minimum — move it and nothing recovers, because every quote constrains the level. ρ and ν have long flat floors running well below the tenth-of-a-point line, so a wide range of them is indistinguishable from the truth. A calibration report quoting $\rho = -0.30$ is reporting where the optimiser stopped, not something that was measured.

Drawn by `profile` and `identifiable_range`.

Remark (What is and is not being claimed)**.** Not that the optimiser failed — it returned the generating parameters exactly. And not that the minimum is non-unique: with noiseless quotes it is unique, sitting at the truth.

The claim is that the minimum is not identified *to the precision of the data.* Every quote in a real market arrives with a bid-offer around it, and once the objective is flat within that width, the parameters inside the flat region are indistinguishable on the evidence. Reporting one of them to two decimal places is a statement about the optimiser's stopping rule. With a realistic half-point bid-offer the regions above get wider still.

Remark (The identification problem, for the fourth time)**.** This is the shape these notes keep meeting. Chapter 16: a spread pins down $\lambda(1 - R)$ and neither factor, so recovery is fixed by convention. Chapter 18: the marginals are pinned and the copula is free. Chapters 10 to 12: the smile pins the marginals and leaves the dynamics open. Here: the quotes pin the level and leave the shape parameters loose.

Each time the resolution is the same and it is not statistical. Either bring more information — more instruments, of a kind that depends on the unidentified quantity — or fix the parameter by

convention and be honest that it was fixed. The one thing that never works is a better optimiser, because there is nothing wrong with the optimiser.

Example 20.3 (Bringing more information)**.** The constructive half, and it is measurable. Widening the quoted strikes, holding everything else fixed, changes the range of ρ that survives:[5]

Strikes quoted	ρ **still fitting**	**Width**
$\pm 0.5\%$	$[-0.999,\ -0.209]$	0.79
$\pm 1\%$	$[-0.999,\ -0.252]$	0.75
$\pm 2.5\%$	$[-0.314,\ -0.287]$	0.03

A thirtyfold improvement, from nothing but asking for quotes further out of the money. Correlation is a statement about the wings, so the wings are where the information about it lives, and a set of near-the-money quotes simply does not contain it. This is what "bring more information" means concretely, and it is a conversation with the broker rather than with the model.

Remark (Why β is fixed by convention)**.** Chapter 10 noted that β is conventionally set rather than fitted and left it there. The table above is the justification. A single expiry's smile already fails to pin ρ and ν; adding a fourth parameter that also controls shape makes the flat region larger still. β is chosen from a view about how the smile should move when the forward moves — which is dynamic information, and therefore precisely what a single day's quotes cannot contain.

20.4 The Bias Nobody Checks

Now the other job. The canonical estimation problem in these notes is the one chapter 22 will build its trades on: a spread is believed to mean revert, and the trade is sized by how fast.

Model it as an Ornstein-Uhlenbeck process,

$$dX_t = \kappa(\theta - X_t)\,dt + \sigma\,dW_t, \tag{20.3}$$

where κ is the reversion rate and $\ln 2/\kappa$ the half-life. Sampled at spacing Δ, (20.3) is an autoregression,

$$X_{i+1} = c + \phi X_i + \varepsilon_i, \qquad \phi = e^{-\kappa\Delta}, \tag{20.4}$$

so regressing each observation on the previous one recovers ϕ and hence κ. This is also the maximum likelihood estimator, since the transition is Gaussian. It is what every desk does.

It is biased, and severely.

Calculation 20.3 (Where the bias comes from and what it scales with)**.** The least squares estimate of an autoregressive coefficient is biased downwards — the classical result is

$$\mathbb{E}[\hat{\phi}] - \phi \approx -\frac{1 + 3\phi}{n}$$

for n observations.

[5] the narrowing, checked.

Least squares is unbiased under *strict exogeneity*: the errors must be uncorrelated with the regressor at every date, past and future, not merely at the same one. Regressing x_t on x_{t-1} satisfies the weaker condition and fails the stronger. It satisfies $\mathbb{E}[x_{t-1}\varepsilon_{t-1}] = 0$, since x_{t-1} is fixed before ε_{t-1} arrives — which is why the estimator is consistent, and why the problem disappears as the sample grows. But $x_{t-1} = \phi x_{t-2} + \varepsilon_{t-2}$ is built out of the earlier shocks, so it is correlated with ε_{t-2}, ε_{t-3} and all the rest. Consistency is what predeterminedness buys; unbiasedness is what strict exogeneity would have bought, and it is not available.

That says the estimator is biased without saying which way, and the sign has two sources. The estimate is a ratio,

$$\hat{\phi} = \frac{\sum (x_t - \bar{x})(x_{t-1} - \bar{x})}{\sum (x_{t-1} - \bar{x})^2},$$

and its denominator is not independent of its numerator: a shock raises x at its own date and, through the recursion, at every date after, so it inflates the sum of squares in a way that is tied to the errors above. The ratio is pulled down. That much happens even when the mean is known, and it gives a bias of about $-2\phi/n$.

Estimating the mean adds the rest, and is the easier half to picture. $\bar{x}$ is computed from the same path, and a sample mean drawn from one realisation of a persistent series sits closer to that realisation than the true mean does — the path drags its own average along with it. Measured against a centre that has followed it, the series appears to return sooner than it truly does. That is what takes the bias from $-2\phi/n$ to the $-(1+3\phi)/n$ above, and at ϕ near one it roughly doubles it. Both halves are measurable, and both are there: simulating an autoregression at $\phi = 0.95$ over two hundred steps, the two fits differ by a factor of about 2.4.[6]

A downward bias in ϕ is an *upward* bias in $\kappa = -\ln \hat{\phi}/\Delta$: the series looks less persistent than it is, so it looks to revert faster than it does.

Now propagate it. With Δ small, $\phi \approx 1$, and

$$\delta\kappa \approx -\frac{\delta\phi}{\phi\Delta} \approx \frac{4}{n\Delta} = \frac{4}{T},$$

where $T = n\Delta$ is the length of the history in years. The Δ has cancelled. The bias depends on the *span* of the data and not on how finely it was sampled, and it does not depend on κ either.

Remark (Read that again, because it is the practical point)**.** The instinct on discovering an estimate is noisy is to get more data, and in practice that means higher frequency data, because history has a fixed length and the tick archive does not. Equation above says that does nothing. Going from daily to hourly multiplies the number of observations by eight and leaves the bias exactly where it was.

The reason is that mean reversion is a statement about a timescale. A history informs you about it only by containing several of them, and sampling the same three years more finely does not contain more three-year periods. Only a longer history does.

Example 20.4 (The size of it)**.** Simulating (20.3) with $\kappa = 1$ — a half-life of eight months — and estimating it the standard way, in `quant/src/estimation.rs`:

[6] `measured`, against Kendall's two constants.

History	Daily	Hourly	Bias $\times T$
1y	6.10	6.04	5.1
2y	3.42	3.42	4.8
5y	1.91	1.93	4.6
10y	1.43	1.44	4.3
20y	1.21	1.21	4.2
40y	1.10	1.11	4.0

A single year of daily data reports a mean reversion of 6.1 against a truth of 1. The last column is the prediction $\delta\kappa \cdot T \approx 4$, and it holds across the range.

In half-lives, which is how a trader thinks: the truth is 8.3 months. Two years of data says 2.4 months. Ten years says 5.8. The estimate does not merely have error bars around the right answer — it is systematically, and largely, on the optimistic side.

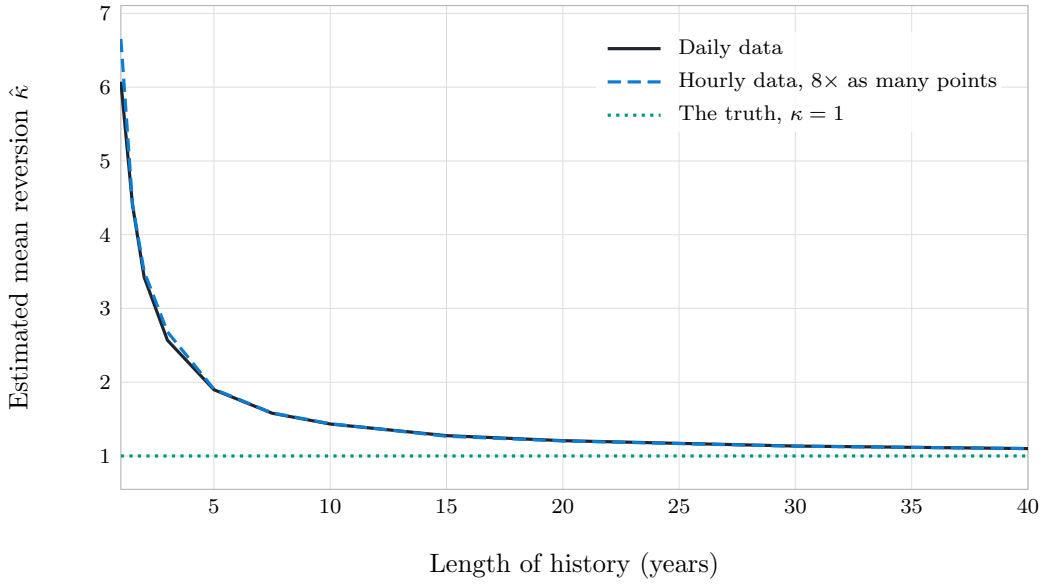

Figure 20.2: The estimated mean reversion against the length of history used, for a process whose true value is the flat line. Two things to read. The daily and hourly curves lie on top of each other, though one uses eight times as many observations — more data of the wrong kind is not more information. And the convergence is slow: it takes decades of history to get within ten percent of a reversion rate whose own half-life is under a year. Move the slider to a slower reverting process and the situation gets worse, because what matters is how many half-lives the window contains.
Drawn by `fit_ou` and `mean_reversion_estimate`.

Remark (What this does to a trade)**.** Chapter 22 will size relative value trades on the assumption that a spread returns to fair value on a known timescale, and this is the estimator that finds it. Suppose the true half-life is eight months and two years of data suggests two and a half.

Everything downstream inherits the error, and in the same direction. The expected holding period is understated by a factor of three, so the expected annualised return is overstated by the

same factor. The probability of the spread moving further against the position before it converges is understated, so the stop is set too tight and the position too large. And the capital charge, which depends on how long the exposure persists, is understated as well.

None of that is a failure of the strategy. It is one biased estimator propagating through every number that depends on it.

Exercise. Show from (20.4) that a series with no mean reversion at all, $\kappa = 0$, still produces a strictly positive $\hat{\kappa}$ with probability approaching one on a finite sample, using nothing beyond the bias already derived above.

20.4.1 Two Other Ways to Fit the Same Process

Least squares is one estimator, and the question is whether its bias is a property of least squares or of the problem underneath it. Two genuinely different estimators, on the same running example, answer that from different directions.

Calculation 20.4 (Method of moments)**.** The stationary distribution of (20.3) has mean θ, variance $\sigma^2/(2\kappa)$, and lag-one autocovariance at spacing Δ equal to the variance times $e^{-\kappa\Delta}$. Compute the sample mean, variance and lag-one autocovariance directly from the path, and invert:

$$\hat{\kappa} = -\frac{1}{\Delta}\ln\frac{\widehat{\mathrm{Cov}}}{\widehat{\mathrm{Var}}}, \qquad \hat{\theta} = \bar{x}, \qquad \hat{\sigma} = \sqrt{2\hat{\kappa}\,\widehat{\mathrm{Var}}}. \tag{20.5}$$

Nothing here is a regression: three numbers are read off the path and the three parameters solved for algebraically. Classical — Yule-Walker, carried over from discrete time — and a genuinely different computation from (20.4), sharing no arithmetic with it beyond the data.[7]

Example 20.5 (The same bias, from an unrelated computation)**.** Averaged over four thousand histories at each window, on the same $\kappa = 1$ process this chapter has used throughout:

History	Least squares	Method of moments
2y	3.48	3.83
5y	1.94	2.03
10y	1.44	1.47
20y	1.20	1.21
40y	1.09	1.10

The two estimators share nothing — one regresses, the other matches three moments to three equations — and they are biased upward by close to the same amount at every window, closing towards the truth at close to the same rate. That is the stronger claim this section has been building to: the upward bias in a short-window mean-reversion estimate is not an artefact of least squares. It is a property of estimating a slow timescale from a sample that does not contain many of them, and it survives a change of estimator entirely.[8]

[7] `estimation::fit_ou_moments`.

[8] `measured`.

Calculation 20.5 (Maximum likelihood with the actual tails)**.** Ordinary least squares is the maximum likelihood estimator under the assumption that the innovations in (20.4) are Gaussian. Chapter 18 gave every reason to doubt a real spread's innovations are: fitting a Student-t to daily changes there found four degrees of freedom typical of a name, not thirty. Replace the assumption and refit by maximum likelihood instead: at a fixed ν, weight observation i by $(\nu+1)/(\nu+(\hat{e}_i/s)^2)$, re-run the regression weighted, iterate to convergence, then search over ν for the value the likelihood prefers — the same Student-t machinery chapter 18 built for a copula, run here on a regression's own residuals rather than on a pair of marginals.[9]

Example 20.6 (What the weighting buys, and when)**.** Simulate the same process with Student-t innovations at four degrees of freedom, and compare the two estimators' root mean squared error in κ over eight hundred independent five year histories: least squares comes out at 1.45, the t-weighted fit at 1.11 — a quarter less error, from down-weighting exactly the days a fat-tailed shock would otherwise have dragged the fitted line towards. Run the same comparison with innovations close to Gaussian instead and the two agree to within a tenth of a percent, because the fitted degrees of freedom simply comes back large: nothing was given up by allowing for tails that were not there.[10]

Remark (What neither one touches)**.** Both are real improvements, and both are improvements to a different defect than the one this section opened with. Method of moments and the Student-t refit change how efficiently a given history is used — the second materially, whenever the innovations are actually fat-tailed. Neither changes how much a short window can say about a slow timescale, which is bounded by the number of half-lives the window contains and by nothing an estimator can do about it. Chapter 21's later distinction is the same one again: some defects are about the estimator, and some are about what the data was ever going to contain.

20.5 Testing Rather Than Estimating

Everything above estimates a parameter. This section asks the prior question, which a chapter called fitting and testing owes the reader and which the estimation above quietly assumes: does the data support mean reversion at all?

It matters because the answer is frequently no, and because chapter 22's trades rest on it. A convergence trade is a bet that a spread returns to a level. If the spread is a random walk the bet has no edge whatever, and telling the two apart is a hypothesis test rather than an estimation.

Definition 20.6 (The Dickey-Fuller test)**.** Regress the increment on the level,

$$x_t - x_{t-1} = c + b\,x_{t-1} + \varepsilon_t, \tag{20.6}$$

and test the null $b = 0$, a random walk, against $b < 0$, mean reversion. The statistic is the ordinary t ratio on b.

Remark (Why the critical value is not from a t table)**.** Equation (20.6) looks like a regression and its statistic does not have a t distribution.

[9] `estimation::fit_ou_student_t`.

[10] `the reduction, measured`; `both halves, measured`.

Standard regression theory needs the regressor to be well behaved as the sample grows. Here the regressor is the series' own lagged level, and *under the null* that series is a random walk, whose variance grows without bound. The usual limit theory does not apply, the statistic's distribution is skewed to the left, and using -1.65 from a normal table would reject far too often. In terms of (20.4), $b = 0$ here is exactly $\phi = 1$ there — a root of the AR polynomial $z - \phi$ sitting on the unit circle, which is what the field calls a unit root and where the whole family of tests gets its name. Dickey and Fuller tabulated the correct distribution by simulation; with an intercept and no trend the five per cent point is about -2.86, which is to say the test's *size* — the probability of rejecting a genuine random walk — is five per cent by that choice.

This is the standard trap in testing for mean reversion, and it arises because the null hypothesis is non-stationary. Testing a hypothesis under which nothing converges requires distribution theory in which nothing converges either.

Calculation 20.7 (How much data it takes)**.** Take a spread that genuinely reverts with a one-year half-life, observe it daily, and count how often the test finds the reversion that is really there.[11]

Sample	Rejects random walk	Median half-life	90% interval
2 years	6%	0.26	0.09 – 1.61
5 years	9%	0.49	0.19 – 1.64
10 years	19%	0.67	0.31 – 1.61
20 years	54%	0.81	0.45 – 1.52
40 years	98%	0.90	0.59 – 1.43

The test's size is five per cent, so the first row is barely distinguishable from rejecting at random. Two years of daily data on a spread that certainly reverts is powerless to say so.

Remark (Sampling faster does not help)**.** The natural response to the first row is to collect more observations, and it does not work. Observing the same five years ten times as often — hourly rather than daily, fifty times the data — leaves the power essentially unchanged.[12]

The reason is the one the remark *Read that again, because it is the practical point* already gave for the estimate's bias: what matters is elapsed *half-lives*, and fifty times the observations over the same calendar span produces no additional ones. The information a test can extract about κ, like the information an estimator can, is bounded by the span and not the sample size.

Separate this from the usual observation that more data is better. More data along the time axis is better. More data within a fixed window is, for this purpose, almost worthless — and the distinction is invisible if one thinks of a sample only as a number of rows.

Structure (What this does to a horizon)**.** Read the last two columns of Calculation 20.7 against chapter 22's use of them, because the consequence is stronger than that chapter claims.

There the argument was that mean reversion is estimated too fast, so a horizon is planned too short, with the illustration taken at thirty per cent fast. The table says thirty per cent is

[11] `estimation::inference_quality`, with the test's size checked against a near random walk first, since a power figure means nothing without it.

[12] `measured, within six percentage points`.

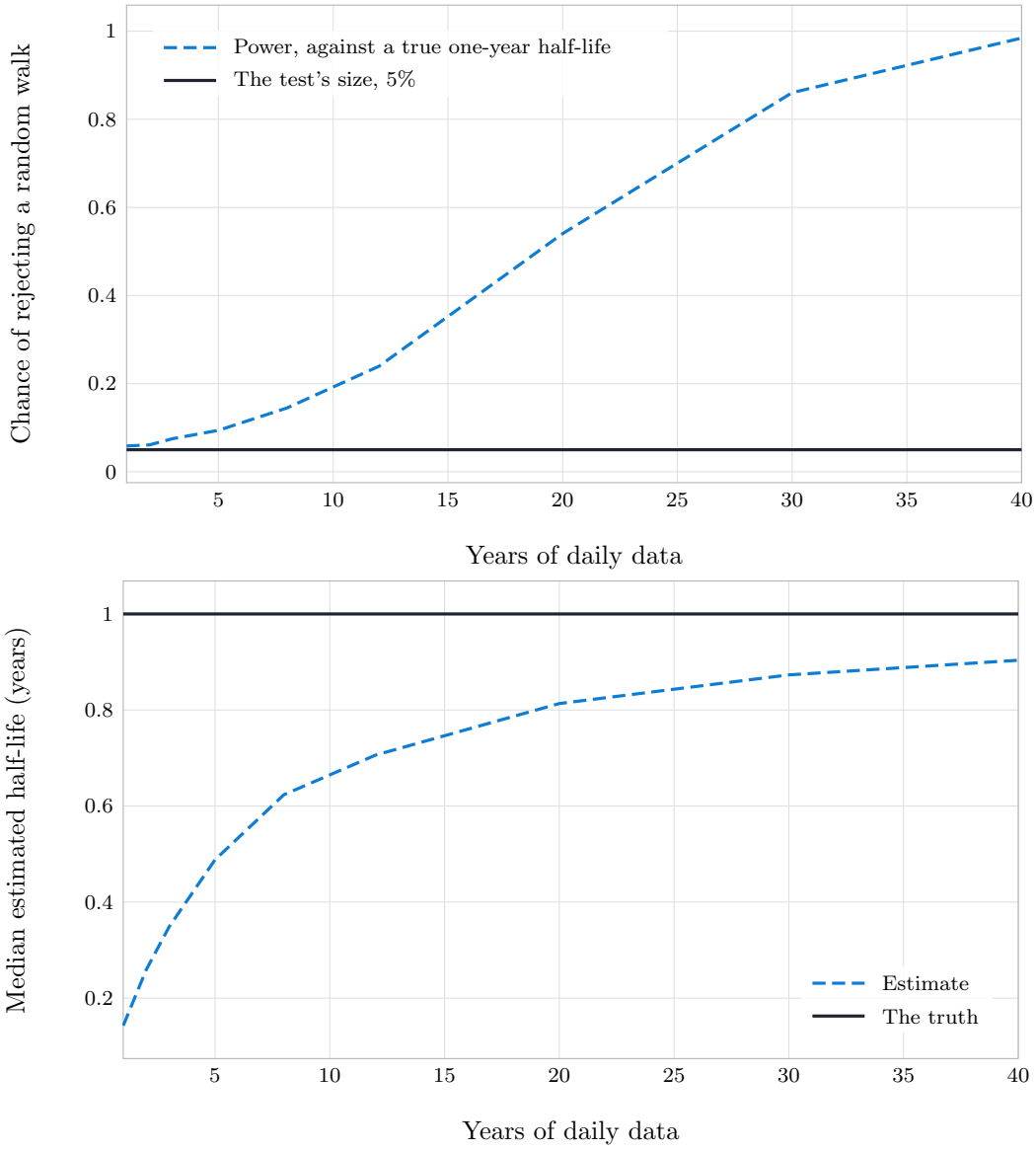

Figure 20.3: Above: the chance of rejecting a random walk, against the length of the record, for a spread that genuinely reverts with a one-year half-life. The flat line is the test's own size — the rate at which it rejects when there is nothing to find — and the two are hard to tell apart until the record is several years long. Below: the median estimated half-life against the truth, approaching it from beneath and still short of it after forty years. Both curves are functions of the calendar span and not of the number of observations, which is the subject of the next remark.

Drawn by `inference_quality`.

conservative: at five years of daily data the median estimated half-life is about half the truth, and at ten years about two thirds of it. A desk with a decade of history on a spread will typically believe it reverts one and a half times as fast as it does.

Worse is the interval. At five years the ninety per cent range for the half-life runs from about a fifth of a year to about a year and two thirds — a factor of eight. A horizon is not a quantity that data of that length determines, and a position sized on the point estimate is sized on a number the sample barely constrains.

Which leads to an uncomfortable but honest place. The structural reason chapter 22 requires for a trade is not merely a nice supplement to the statistics; it is doing most of the work, because the statistics on a realistic sample cannot establish either that the spread reverts or how quickly. A trader who says "this spread has a six-month half-life" is reporting a point estimate from a sample that could not distinguish six months from two years, and a trader who says "this spread is wide because a regulation forces somebody to hold the other side" is making a claim the data can actually support.

20.6 Does the Model Fit At All?

Both sections so far assume the model is right and ask what its parameters are. The prior question is whether to use it, and there is a general answer.

Theorem 20.8 (Probability integral transform)**.** *Let X have continuous distribution function F. Then $U = F(X)$ is uniform on $[0,1]$.*

Proof. $\mathbb{P}(U \le u) = \mathbb{P}(F(X) \le u) = \mathbb{P}(X \le F^{-1}(u)) = F(F^{-1}(u)) = u.$ □

Remark (Why this is the whole test)**.** Trivial to prove and unreasonably useful. If a model claims a conditional distribution for tomorrow's observation given everything known today, then feeding each realised observation through that day's claimed distribution function must produce numbers that are uniform on $[0,1]$ and independent of each other. Uniform because of Theorem 20.8; independent because each used all the information available at the time, so nothing predictable can be left.

Two properties, both testable, and *no reference to what the model is.* It applies to a diffusion, a jump process, a copula, a neural network, or a trader's intuition written down as a distribution. Anything that outputs a conditional distribution can be tested this way, and anything that does not output one cannot be tested at all — which is itself a useful criterion when someone offers you a model.

Example 20.7 (Catching a volatility that moves)**.** Take four thousand daily returns generated with Heston's own mean-reverting variance — the calibration chapter 10 already uses — and test them against a constant-volatility lognormal model fitted, generously, to the realised variance, so the overall level is right and only the shape is wrong.

The transformed values come out far from uniform in three places at once, not the clean hump or clean U a first guess might expect: a sharp excess at the centre, a genuine excess right at the extreme edges, and a deficit in the shoulders between them. Write R for the ratio of a day's actual

volatility to the fitted level s, so $\mathbb{E}[R^2] = 1$ by construction. The centre is provable outright. For a single fixed $R = r$ the transform is $\Phi(rZ)$ for a standard normal Z, and its density at $u = 0.5$ works out to exactly $1/r$: a quiet day, $r < 1$, produces a move too small for the wider reference model to read as extreme, so it lands close to the centre whatever it actually was. Averaging over the day-to-day randomness of R gives the density actually observed there, $\mathbb{E}[1/R]$, and this is where Jensen decides it: $1/r$ is convex, so an unusually quiet day pushes it arbitrarily high while an unusually loud one can only push it towards 0, never below — the quiet extreme has unlimited leverage on the average and the loud extreme does not. So $\mathbb{E}[R] \leq 1$ (Jensen once) forces $\mathbb{E}[1/R] \geq 1/\mathbb{E}[R] \geq 1$ (Jensen again), and the transform's density at $u = 0.5$ is at least one, strictly more whenever the volatility genuinely varies — the typical day is quieter than the level fitted to its mean square. The edge excess is a different fact, not Jensen's: whenever a day's volatility can exceed the fitted level at all, the transform's density diverges exactly at 0 and 1, from the minority of days loud enough to reach there, and it is invisible at any bin coarser than the tail itself. The Kolmogorov-Smirnov statistic rejects at any conventional level regardless of which piece a coarse histogram happens to show. Run the same test on data that really was generated with constant volatility and it passes, so the rejection means something.[13]

Note what the *shape* of the deviation tells you, and do not trust the first guess: a spike at the centre is Jensen's, and it says a single fitted level cannot track a volatility that genuinely moves — most days look calmer to it than they are. An excess right at the very edges, invisible in a coarse histogram, is the same mismatch seen from its rarer, louder days, not a separate defect. A tilt says the drift is wrong. The test does not merely reject; it says in which direction to look next.

Remark (Chapter 24's backtest will be a special case)**.** Value at risk backtesting, which chapter 24 comes to, counts the days on which the loss exceeded the reported quantile and asks whether there were too many. In this language: the model claims a conditional distribution, the transform of each day's loss should be uniform, and the exception indicator is just checking whether the transform exceeds 0.99 as often as it should.

Which is why that test is not allowed to stop at the count: the exceptions must also be independent. That is the second half of the transform's conclusion, and it is the half that catches the models that fail worst, since a model that is wrong about volatility clustering will produce the right number of exceptions arriving all in the same week.

Remark (In-sample fit is not evidence)**.** The deepest point in this chapter, and chapter 9 has already made it once. A local volatility model reproduces every European option price exactly. It is nonetheless wrong, in the specific sense that its predictions about how the smile moves are contradicted the next day.

A model with enough parameters fits anything. What distinguishes a model that works is out-of-sample and *dynamic* performance: not whether it reproduces today's surface but whether the hedge it prescribes actually removes the risk, and whether the surface it predicts for tomorrow is the surface that arrives. Chapter 10's decomposition of delta is the concrete form of that test, and it is the one an in-sample fit can never pass or fail.

[13] `estimation::cir_pit_shape` and `estimation::ks_statistic`; `the correctly specified model passes`, `the misspecified one shows all three, the edge excess checked at the exact percentile rather than by eye`, and `what speeding up the reversion does to both`.

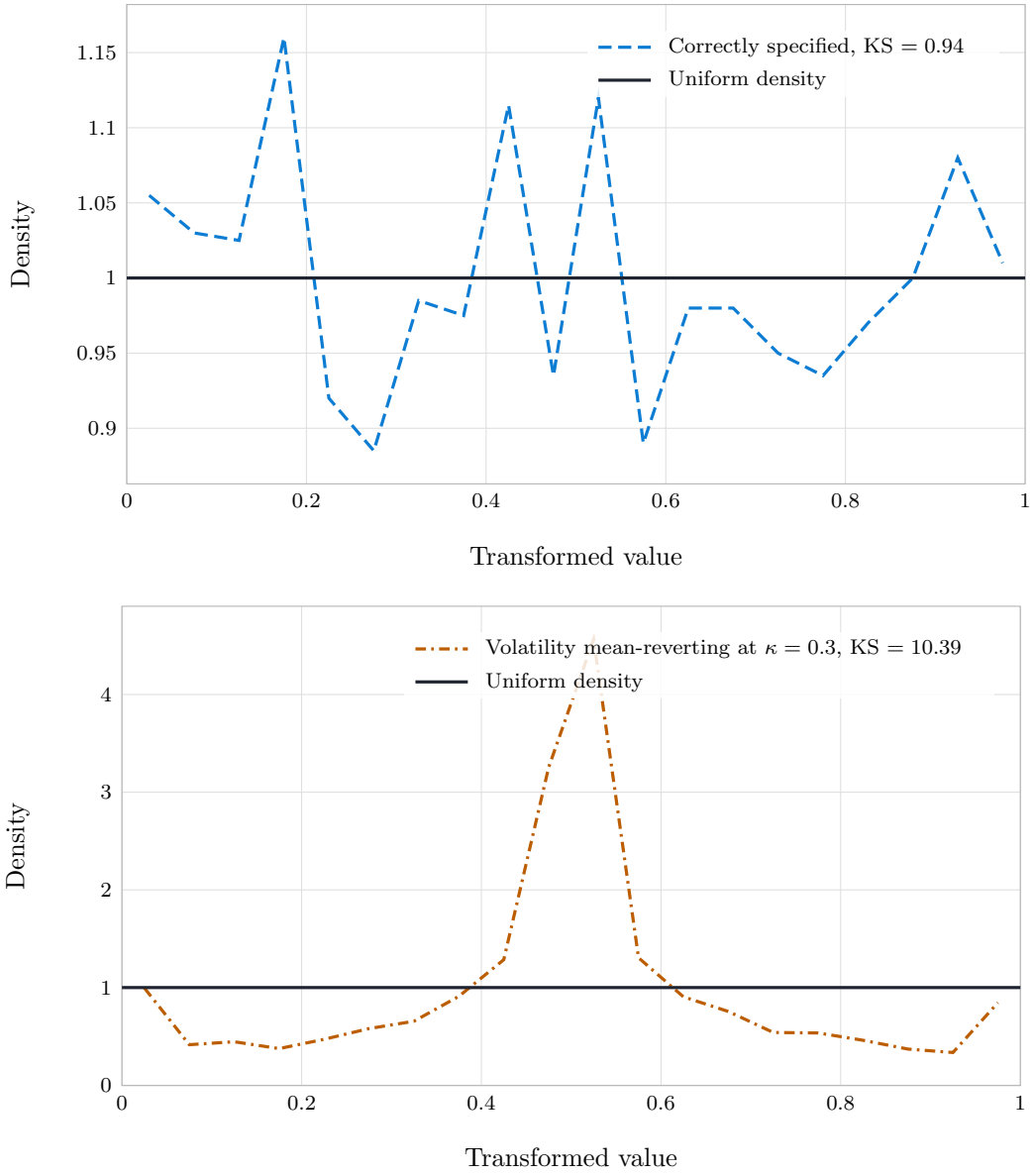

Figure 20.4: The transformed values of the same four thousand returns, histogrammed against the uniform density they should have. Above, the model tested against the data it actually came from: flat, and the Kolmogorov-Smirnov statistic sits well inside the rejection region. Below, the constant-volatility model tested against data with Heston's own mean-reverting variance, fitted with exactly the right average level: a sharp spike at the centre, a fainter excess right at the extremes, and a deficit between them — both a hump and a U, not a choice between them. Drag the mean-reversion speed and the whole shape converges towards uniform rather than trading one bump for the other: a faster-reverting volatility has a tighter stationary range for the same volatility-of-volatility, so the centre spike and the edge excess shrink together, and neither survives on its own.

Drawn by `cir_pit_shape` and `ks_statistic`.

20.7 What To Take Away

Three things.

First, a good fit is not a measurement. Ask of every fitted parameter how far it could move before anyone would notice, and report that width alongside the number. Most of the time nobody has computed it, and often it is embarrassing.

Second, when a parameter turns out to be unidentified, the fix is more information or an honest convention. It is never a better optimiser.

Third, the test of a model is out-of-sample and dynamic. The probability integral transform gives a general way to run that test for anything that emits a conditional distribution, and refusing to emit one should be treated as a property of the model rather than a detail of its implementation.

None of this is specific to derivatives, and none of it has changed in fifty years.

References

- Kendall, M. G. (1954). Note on bias in the estimation of autocorrelation. *Biometrika*, 41(3-4), 403–404.

- Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4), 863–883.

- Cont, R. (2006). Model uncertainty and its impact on the pricing of derivative instruments. *Mathematical Finance*, 16(3), 519–547.

Chapter 21

Estimation in Practice

Chapter 20 asked whether a fit means anything. This chapter asks the same question where the parameter is never observed at all — a variance, a regime, a factor driving a curve — and where the last decade of machine learning has made the boldest claims. Filtering recovers a latent state in principle and a network makes the recovery fast; neither changes what the data can identify, and telling the two apart is most of what follows. We work through what is genuinely hidden and what only looks that way, how long a regime takes to notice before it has already changed again, whether prices and a history can be fitted at once, the two parameters estimated constantly and understood least — a volatility and a correlation — and the oldest asymmetry in the subject: why implied volatility sits above realised, and what it costs to collect the difference.

21.1 What the Newer Methods Change

Everything in chapter 20 was classical. What the last decade of machine learning does and does not alter is worth setting out plainly, because the claims made for it are uneven.

21.1.1 Latent states: filtering, and the likelihood it hands over

Chapter 20 assumed throughout that the thing being estimated is observed. In fixed income it never is. Nobody sees the short rate, the variance, or the factors driving a curve; what is seen is a set of yields or option prices, each a known function of the state plus an error. That is a state-space model, and it changes what estimation means: there are now two problems, recovering the state and estimating the parameters, and the second cannot be attempted until the first is solved.

Definition 21.1 (State-space form)**.**

$$\begin{aligned} x_{t+1} &= f(x_t) + \text{noise} && \text{the state, unobserved} \\ y_t &= h(x_t) + \text{error} && \text{what is quoted} \end{aligned}$$

A *filter* computes $p(x_t \mid y_{1:t})$, the state given everything seen so far.

Remark (When it is exact: the Kalman filter)**.** If f and h are affine and both noises Gaussian, the filtering distribution is Gaussian and the recursion is exact and cheap.

Linearity is what makes the recursion *closed*: an affine map of a Gaussian is Gaussian, so a distribution described by a mean and a variance stays described by a mean and a variance, and the filter never has to represent anything else.

Gaussianity is what makes it *optimal among all estimators*. Drop it but keep linearity and the same recursion is still the minimum-variance estimator among linear ones, for any noise with finite variance — so the filter is not useless off its assumptions, it is merely no longer using everything in the data. What does break is the likelihood: the number the recursion emits is then a Gaussian quasi-likelihood rather than the true one, which is still usable for estimation but gives standard errors that are wrong unless corrected.

In the affine Gaussian case both hold. That case is not a toy here — it is precisely a Gaussian affine term structure model, where yields are affine in the state by construction, so chapter 8's models are estimated this way as a matter of course.

The recursion is two steps. *Predict* moves the state distribution forward by the dynamics, which widens it. *Update* conditions on the new observations, which narrows it, by a precision-weighted average of the prediction and the data — Bayes' rule for Gaussians and nothing more.

The by-product is what makes it valuable. Each step produces a predictive distribution for the next observation, and summing its log density over the history gives $p(y_{1:T} \mid \theta)$ *exactly*. So the filter does not merely estimate the state, it delivers the likelihood of the parameters — which is what makes maximum likelihood, or the posterior of the previous subsection, possible at all for a model whose state is never seen.[1]

Structure (The measurement error is not a nuisance term)**.** Five yields and a one-factor model is an over-determined system.

The difficulty is that the curve is not free. Here $y_t \in \mathbb{R}^5$ is the vector of the five yields quoted at date t, one per maturity, and h is affine: $y_{t,i} = a_i + b_i x_t$, with loading $b_i = (1 - e^{-\kappa\tau_i})/(\kappa\tau_i)$ and intercept a_i fixed by the parameters. The parameters are held still while the filter runs; the only thing that moves from one date to the next is the scalar state x_t. So the observation is not five points to be joined but a five-vector,

$$y_t \;=\; a + x_t\, b, \qquad a, b \in \mathbb{R}^5 \text{ fixed},$$

with $a = (a_1, \ldots, a_5)$ and $b = (b_1, \ldots, b_5)$ the fixed intercept and loading vectors, required to lie on a *line* in $\mathbb{R}^5$ — the set $\{a + s\, b : s \in \mathbb{R}\}$ that the scalar x_t sweeps out as it varies, one dimension inside five. Five numbers, one unknown, and generically no x_t reproduces them. The measurement error is what makes the problem well posed.

That also settles an apparent contradiction with chapter 8, which fits today's curve exactly with a one-factor model. It does, and it buys the fit with a time-dependent drift: a free function, in effect one number per maturity, spent entirely on the first date. It does nothing on any date after, where the same fixed loadings apply and only x has moved. A one-factor model can match any single curve and cannot match a panel of them, and it is the panel the filter sees.

[1] `estimation::AffineStateSpace`, with `the likelihood peaking at the true mean reversion` and the filtered state beating the best single yield by a third.

Two consequences. Assuming the yields are cleaner than they are makes the state estimate *worse*, because the filter then trusts each observation more than it deserves and chases the noise — which is measurable and is measured. And the fitted size of the error is a diagnostic worth reading: a model whose estimated measurement error comes out far above the bid-offer is telling you it cannot fit the cross-section, whatever its time series behaviour looks like. That is a specification test that costs nothing, since the number is estimated anyway.

Example 21.1 (Estimating the short-rate dynamics and the noise together, from yields alone)**.** The same five-yield model as above, now with nothing handed to the filter but 3,000 daily observations: the mean reversion κ, the state's own volatility σ, and the measurement error common to every yield, all three estimated at once by maximising the Kalman likelihood.[2]

It comes back with $\hat{\kappa} \approx 0.315$ against a truth of 0.3, $\hat{\sigma} \approx 0.0092$ against 0.01, and a measurement error of about 5.5 basis points against a true 5 — every one of the three within ten percent. The recovery is tighter than Heston's below, and for a specific reason: this likelihood is exact and smooth in every parameter, with no resampling to make it noisy, so a coordinate search has an easy target rather than a rough one.[3]

Remark (When it is not exact)**.** Affine and Gaussian is a strong assumption and the alternatives are ranked by how much of it they give up.

Extended and unscented filters keep the Gaussian approximation and handle a nonlinear h — the first by linearising, the second by propagating a small set of points through the nonlinearity and refitting a Gaussian. Both are cheap and both fail in the same circumstance, which is a nonlinearity strong enough that the filtered distribution is not Gaussian in shape. A square-root variance near zero is exactly that circumstance, so the models where one most wants a filter are the ones where these approximations are least safe.

Particle filters give up the Gaussian entirely: represent the distribution by a weighted sample, push each particle through the dynamics, reweight by how well it explains the new observation, and resample. Nothing is assumed about shape, which is why this is the method for a stochastic volatility model, where the state is a variance that is positive and skewed.

The cost is the usual one and it has a specific form. Weights degenerate — after a few steps one particle carries almost all of them — so resampling is required, and resampling makes the likelihood estimate noisy and discontinuous in the parameters, which breaks any optimiser that wants a gradient. The repair is to embed the particle filter's unbiased likelihood estimate inside a Markov chain Monte Carlo scheme rather than to optimise it, which is exact despite the noise and is why particle methods and the Bayesian treatment above are usually met together.

21.1.2 Particle filters, in more detail

Stochastic volatility has a problem the sections above sidestepped: v_t is not observed. Prices are, volatility is not, so the likelihood of a Heston model given a price history involves integrating over every possible volatility path — an integral of enormous dimension, which is why stochastic volatility models are so often calibrated to option prices instead of estimated from returns.

[2] `estimation::fit_affine_state_space`.

[3] `the recovery, measured against each of the three truths at once`.

Sequential Monte Carlo solves it. Carry a population of particles, each a hypothesis about the current v_t; propagate them through the model's dynamics; reweight by how well each explains the newly observed return; resample. What comes out is the filtered distribution of the latent state, and — the part that matters here — the likelihood of the data as a by-product, which makes maximum likelihood estimation possible for models where it previously was not.

Two things about it matter before using it, because both are places it fails quietly.

Weights degenerate. After a few steps one particle carries almost all the weight and the rest carry none, so the effective sample size collapses and the filter is representing a distribution by one point. Resampling is the standard repair — discard the low-weight particles and duplicate the high-weight ones — and it trades the degeneracy for impoverishment, since the surviving particles share ancestors and the population loses diversity in the distant past. Neither problem is visible in the output unless the effective sample size is monitored, and it should be.

The proposal matters more than the particle count. The naive scheme propagates particles through the model's own dynamics and then reweights, which is wasteful when the observation is much sharper than the dynamics: nearly every particle lands somewhere the data rules out, and the few that survive determine the answer. Proposing from a distribution that already accounts for the new observation — the auxiliary and guided variants — fixes this, and the improvement is far larger than the one bought by multiplying the number of particles.

This is the principled answer to a real problem, it is genuinely used, and it has no fashionable name.

Calculation 21.2 (A Heston variance, filtered from returns alone)**.** Simulate a Heston path — $\kappa = 2$, $\theta = 0.04$, $\eta = 0.5$, $\rho = -0.6$, daily steps over six years — and forget the variance was ever known. The bootstrap filter above needs one adjustment for the leverage correlation ρ: draw each particle's variance shock *before* scoring the day's return against it, so the return's likelihood and the particle's proposed next variance share the randomness the model says they must, rather than being drawn as though they came from two independent sources.[4]

Handed the true parameters, the filter's own estimate of v_t has a root mean squared error a bit over half the size of simply reporting the long-run average θ every day: real information, and far from all of it — which is exactly why these models are usually calibrated to option prices instead of estimated from returns, a fact stated earlier in this section and measured here.[5]

Example 21.2 (Estimating without being handed the answer)**.** Now the harder version of calculation 21.2: nothing but the six years of returns, and all four parameters — κ, θ, η, ρ — estimated by maximising the filter's own likelihood.[6] A coordinate search rather than a gradient method, deliberately: resampling makes the likelihood surface noisy and discontinuous in the parameters, for the reason given above, and a textbook implementation would go further and embed the filter inside a particle Markov chain Monte Carlo scheme, which tolerates that noise rather than being defeated by it. This is the cheaper version — common random numbers held fixed across the search to tame the discontinuity rather than remove it — enough to show the mechanism recovers something real.

[4] `estimation::heston_particle_filter`.

[5] `measured, against the constant-θ baseline`.

[6] `estimation::fit_heston_particle_filter`.

It comes back with $\hat{\kappa} \approx 1.76$ against a truth of 2, $\hat{\theta} \approx 0.037$ against 0.04, $\hat{\eta} \approx 0.58$ against 0.5, and $\hat{\rho} \approx -0.50$ against -0.6 — every parameter within thirty percent of the truth, the leverage correlation within a tenth. Not exact, and not nothing: four parameters of a model whose entire state is unobserved, recovered from the one thing that was ever going to be observed.[7]

Remark (Heston is affine too — so why doesn't the Kalman filter above apply to it?)**.** Chapter 19's transform methods rest on Heston's variance being affine: its drift and its diffusion coefficient squared, $\eta^2 v_t$, are both affine functions of v_t. That is a real fact about the model, and it is not the fact a Kalman filter needs.

A Kalman filter needs the map from state to observation, and the map from one state to the next, to be affine *with additive, state-independent Gaussian noise* — so that an affine map of a Gaussian stays Gaussian, and the filtering distribution never leaves the family it started in. Vasicek and Hull-White satisfy this: their diffusion coefficient is a constant σ, the corner of the affine family where the state-dependent part of the diffusion vanishes entirely. Heston's diffusion coefficient is $\eta\sqrt{v_t}$, genuinely state-dependent, so the noise entering the transition is neither additive nor of fixed scale, and the state's own stationary law is non-central chi-squared rather than Gaussian. The affine family that makes chapter 19's transforms work is strictly larger than the Gaussian corner of it that makes this chapter's filter exact, and the two conditions are easy to conflate precisely because both go by the same name.

Example 21.3 (What linearising it costs)**.** Apply an extended Kalman filter to the same simulated path as calculation 21.2, given the same true parameters. Two simplifications are needed to get there at all: the leverage correlation is dropped from the observation equation, and the CIR transition is linearised around the current mean rather than carried through exactly.[8]

The particle filter's RMSE against the true variance sits 42% below the constant-θ baseline; the extended Kalman filter's sits only 5% below it — barely better than not filtering at all, and about 63% higher than the particle filter's own error on the identical path. The gap is not an implementation shortfall on the linearised filter's part; it is the measured cost of the two simplifications the remark above made necessary, on a calibration where the Feller condition is violated and v_t genuinely visits the region — near zero — where the linearisation is worst.[9]

21.1.3 Reporting the valley: Bayesian calibration

Chapter 20 ended with a complaint: a calibration reports $\rho = -0.30$ when the evidence supports anything from -0.28 to -1. The Bayesian answer is direct — do not report a point. Put a prior on the parameters, condition on the quotes with a likelihood reflecting the bid-offer, and report the posterior.

The output is then the flat region itself rather than an arbitrary point inside it, and every downstream number inherits the uncertainty honestly. A risk figure computed across the posterior is a risk figure that knows the model was not fully determined, which is exactly the information a point estimate throws away.

[7] `the recovery, measured against each of the four truths at once.`

[8] `estimation::ekf_heston_variance.`

[9] `the two filters, measured on the same path.`

The cost is computational and it is real, since the posterior must be sampled rather than optimised. But note that nothing has been added to the model. The information was always this thin; the Bayesian version merely declines to hide it.

Structure (A posterior over parameters is a posterior over prices)**.** The consequence is larger than better reporting of a calibration.

A price is a function of the parameters, $V = V(\theta)$. A posterior $p(\theta \mid \text{quotes})$ therefore pushes forward to a distribution of prices, obtained by pricing under each draw. That distribution is the object several later chapters have been approximating by hand.

It is the model reserve. Chapter 24 computes model risk by fixing a plausible range for an undetermined parameter, revaluing at each end, and taking the difference. That is a two point approximation to a push-forward, and it is crude in a specific way: it ignores the shape of the valley, ignores that two parameters may be jointly undetermined along a diagonal while each looks fine alone, and produces a number whose size depends on how wide somebody drew the range. A posterior quantile is the same quantity computed rather than sketched.

And it nets correctly across several jointly undetermined parameters at once, which the two-point version cannot. Chapter 24 already computes its reserve at the level of the book rather than trade by trade, for exactly this reason: two trades whose exposure to a single parameter offsets have no joint model risk even though each has plenty on its own. But that repair works only because there is one parameter to revalue against. Once several are jointly undetermined, as the diagonal case above shows, a separate two-point range in each has no notion of the joint region they occupy together and nothing to revalue against it. The push-forward has this for free: one draw of the whole parameter vector θ prices every trade in the book, so the distribution of the portfolio value already reflects every cancellation — across trades sharing one parameter, and across parameters that are jointly rather than separately undetermined — inside the same computation that produced the trade-level number.

It gives relative value a decision rule. Chapter 22 insists that a model saying a spread is wide is not a signal. A posterior says how wide the spread would have to be to mean anything: if the observed spread lies inside the band the model's own uncertainty generates, the model has not identified a mispricing, it has failed to determine a price. Only a spread outside the band is a candidate — and even then the structural reason chapter 22 insists on is still required. This is a sharper filter than the usual one and it is available at no extra modelling cost.

Remark (What the band is really a statement about)**.** Along a direction the quotes constrain, the posterior is narrow and is driven by the data. Along a flat direction — and chapter 20's whole point is that these exist — the posterior is the *prior*, because conditioning on uninformative data returns what was assumed. So the width of the band in those directions measures the prior rather than the market.

That is not a defect, provided it is said. It converts an assumption that was previously invisible into a number somebody has to write down and defend, which is precisely what chapter 24 says the reserve process is for. A band that is wide because the prior was wide is an honest statement that nobody knows; a point estimate in the same situation is a dishonest statement that somebody does.

The practical obstacle is cost, since sampling a posterior requires thousands of pricings of a model that may take minutes each. Which is where the next subsection's neural approximation

earns its place: a fast surrogate for the pricing map does not resolve the identification problem, but it makes the calculation that *reports* the problem affordable. Speed does not buy knowledge; it buys the ability to say how little one has.

21.1.4 Speed, not knowledge: neural approximation

The honest and successful use of neural networks in derivatives is unglamorous: learn the map from parameters to prices offline, once, and then evaluate it in microseconds.

That pays off because the map is expensive for models like rough volatility where each price is a simulation, and calibration needs thousands of evaluations. Learning it turns an overnight calibration into an interactive one. The model is unchanged, the prices are unchanged, and the network is a lookup table with good interpolation — which is why it works and why it is safe: the approximation error can be measured against the true pricer on as many test points as one likes.

Remark (The limit, which is the point of this chapter)**.** A neural network approximating the pricing map does not, and cannot, resolve chapter 20's problem. The flat valley is a property of the map itself — of the fact that quotes near the money contain almost nothing about ρ. Learning that map faster reproduces the flatness faithfully, at speed. No architecture recovers a parameter the data does not contain.

Machine learning changes what is computationally feasible. It does not change what is identifiable, and identifiability is a property of the instruments quoted.

21.1.5 Learned models, and the two questions they do not answer

The more ambitious programme replaces the model rather than approximating it: neural stochastic differential equations, market generators, deep hedging, and network solutions of high-dimensional pricing equations. Those belong to chapter 19, since they are about computing a price rather than fitting one.

Two cautions belong here rather than there, because they are about fitting rather than about what the model can be. The first needs stating carefully, because the usual version of it is wrong.

Whether a learned model is arbitrage-free depends entirely on what is learned. A network fitted directly to the price surface has no structural guarantee: chapter 4 established that a price system without a martingale measure admits a money pump, and nothing in a training loss enforces one — a constraint can be added as a penalty, but a penalty is a preference and not a guarantee. A network placed inside the coefficients of a stochastic differential equation is in a completely different position, and chapter 19 works out why in full: chapter 2's representation theorem fixes the drift and leaves only the diffusion coefficient free, so writing

$$\frac{dS_t}{S_t} = r\,dt + \sigma_\theta(t, S_t, v_t)\,dW_t \tag{21.1}$$

with σ_θ a network produces an arbitrage-free model for every value of θ — the model class itself rather than a restriction on it, with the constraint carried by the equation rather than by a penalty. The remaining problem is the one this chapter is about, which no amount of structure removes: a finite set of quotes does not determine a function.

And chapter 9's warning transfers exactly. A generative model that reproduces the historical distribution of surfaces perfectly is in the position local volatility was in: excellent fit, unconstrained dynamics, and no guarantee that the hedge ratios it implies are the ones that work. Fitting more flexibly does not turn an in-sample fit into evidence.

21.2 What Is Actually Latent

The previous section filtered a state on the grounds that it is not observed, and how much of that is true deserves asking.

Remark (Adapted to what?)**.** "Adapted" is not a property of a process but a relation between a process and a filtration, and three of them are in play whenever a latent state is posited: $\mathcal{G}_t$, generated by what is observed; $\mathcal{F}_t$, the model's own, containing whatever latent factors it carries; and the one generated by the driving Brownian motions. A process can be $\mathcal{F}$-adapted and not $\mathcal{G}$-adapted, and that gap is what the word "latent" is pointing at.

Two separate things force adaptedness, and they are usually conflated. The Itô integral needs a predictable integrand to exist at all — chapter 1's construction fails outright if the integrand sees the increment it multiplies — and since σ is that integrand, it must be predictable in whatever filtration the integral is taken in. That is well-posedness, not economics. Separately, no-arbitrage requires the price to be a semimartingale in the *market's* filtration, which is chapter 4's use of the Bichteler-Dellacherie theorem, and requires trading strategies to be predictable.

Put those together and positing a latent state does not escape adapted volatility, it only relabels it. If S is a semimartingale in $\mathcal{G}$, decompose it there; the continuous martingale part is $\int \hat{\sigma}\, d\hat{W}$ with $\hat{\sigma}$ *$\mathcal{G}$-adapted*, by chapter 2's representation theorem. Whatever hidden machinery is imagined upstream, the observable description always has an adapted volatility. What changes is which one — $\sqrt{v}$ becomes something like $\sqrt{\mathbb{E}[v \mid \mathcal{G}]}$.

Structure (Incompleteness is not unobservability)**.** There is a sharper statement, and it cuts against the usual picture of stochastic volatility as a model with something hidden in it.

Quadratic variation is a limit of sums of squared increments of the *observed* price, so $\langle S \rangle$ is measurable with respect to the price's own filtration. Where it has a density, $v_t = d\langle S \rangle_t / (S_t^2 dt)$ is therefore a function of the path being watched. And once v is known, its own innovations $dZ = (dv - \kappa(\theta - v)dt)/(\eta\sqrt{v})$ are known too. So continuous observation of a single price reveals the entire state of a standard stochastic volatility model. Nothing in it is hidden.

What remains true is that the market is *incomplete*: one traded asset cannot span two Brownian motions, so Z-risk cannot be hedged with S alone. That is a statement about spanning, and it is routinely confused with a statement about observation.

Which relocates the question. Genuine latency does not come from the model. It comes from the observation being discrete, from the microstructure noise that Calculation 21.6 shows drives realised variance away from the truth rather than towards it, and from jumps confounding the estimate. The filtering literature is about imperfect observation, not about a metaphysically hidden state — and that is a more tractable problem, because the imperfection can be measured.

Remark (Two projections, and only one of them is a local volatility model)**.** Chapter 9 projects a stochastic volatility model onto $\sigma^2_{\text{loc}}(t,x) = \mathbb{E}[v_t \mid S_t = x]$, conditioning on the current level and nothing else, and matches one-dimensional marginals. The projection onto what an observer knows conditions on the whole observed path,

$$\hat{\sigma}^2_t = \mathbb{E}\big[v_t \,\big|\, \mathcal{G}_t\big],$$

and these are not the same object. The second is not a local volatility model at all: $\hat{\sigma}_t$ depends on the entire history, which makes it a *path-dependent volatility* model.

That distinction explains a fact both chapter 10 and chapter 18 arrive at from other directions. Matching marginals leaves the joint law free, so two models with the same σ_{loc} agree on every European price and disagree completely about a forward smile. The path projection keeps what the level projection throws away, which is exactly the dependence across time.

It also names the family that modern work has moved towards. If the honest observable model is a functional of the path, then write one. Guyon and Lekeufack's volatility, for instance, is a function of two exponentially weighted moments of past returns — a trend and an activity level — fitted directly to an index's own history rather than to a hidden factor.[10] In full generality, volatility can be written as a functional of the path signature, which is dense in continuous path functionals, so which features of the history matter is learned rather than hand-picked — the construction Perez Arribas, Salvi and Szpruch use to build an entire pricing model this way.[11] Such a model is fully observable by construction, and it has no latent state to filter.

Remark (What a neural stochastic differential equation does and does not escape)**.** A network in the coefficients of

$$dS_t = \mu_t\, dt + \sigma_\theta(t, S_t, v_t)\, S_t\, dW_t$$

is still an Itô diffusion with adapted volatility. It does not go beyond the form — chapter 2 showed the form is not a restriction — it goes beyond the *parametric families* within the form, which is a real gain and a different one.

The genuine extensions are three, and none is a change of equation. Richer path dependence, as above. Rough volatility, where v is driven by a fractional Brownian motion with H near 0.1: the volatility is then neither Markov nor a semimartingale, while the price remains a perfectly good semimartingale and no-arbitrage is untroubled. And jumps, which is the only one that changes the shape, and chapter 2 says what is obtained.

There is also a practical asymmetry to weigh, and it is this chapter's argument in a new setting. A latent-state neural model needs filtering inside the training loop, since the likelihood integrates over unobserved paths — which is why such models are fitted by variational approximation and why their parameters are so poorly determined. A path-dependent model has no latent state, so every conditioning variable is observed, the likelihood is direct, and the parameters are identified by data one actually has. If identification rather than compute is the binding constraint, flexibility is better spent on the path dependence than on a hidden state.

[10] Guyon, J., & Lekeufack, J. (2023). Volatility is (mostly) path-dependent. *Quantitative Finance*, 23(9), 1221–1258.

[11] Perez Arribas, I., Salvi, C., & Szpruch, L. (2020). Sig-SDEs model for quantitative finance. *Proceedings of the First ACM International Conference on AI in Finance.*

21.3 Regimes, and Whether They Can Be Caught

One observation motivates latent states more than any other: past prices are frequently a poor guide to future ones, in a way that looks less like noise than like something changing. A policy stance, a liquidity condition, information held by somebody else — none of it visible in the price history until it acts. The standard response is a regime: volatility takes one of two values, the state switches at random times, and the state is not observed.

It is tempting to describe this as a parameter changing rather than a state being hidden, and the distinction does not survive inspection. A regime indicator *is* a state process; calling σ a parameter that changes is declining to model it, and modelling it turns the parameter into a state.

Remark (What actually separates a regime from a diffusive variance)**.** Something does separate them, and it is not what the word "latent" suggests.

Neither is hidden under continuous observation. If the regimes differ in volatility then $v_t = \sigma^2(R_t)$ is recoverable from the bracket by the argument above, so the regime is recoverable too — a two-state model is no more opaque than a square-root variance. The difference is in the *predictability of the change* rather than the observability of the level. A diffusive variance has continuous paths, so its next value is close to its current one and the near future is forecastable from the present. A regime switches at a totally inaccessible stopping time in the sense of chapter 2: no announcing sequence exists, so nothing in the observed past gives warning. It is fully observable and completely unforecastable, and those two properties are usually assumed to be incompatible.

Which leaves the practical difficulty exactly where the next section puts it. With continuous noiseless observation neither object is hidden, so everything that makes a regime hard is the discreteness and the noise of real observation, and (21.2) is the measure of what that costs.

There is a third case. Suppose the transition law is unknown, or the states cannot be enumerated — which is what people usually mean when they say the world changed. Then no filter applies, because a filter needs a likelihood for each state and a prior over the set of them, and neither exists. That is not a latent state but an unspecified model, and the honest response is chapter 24's reserve rather than an estimate. Distinguishing the two is worth more than distinguishing parameters from states: one is a computation and the other is an admission.

Remark (The regress, and where to stop)**.** Modelling a parameter as a process does not remove an arbitrary choice, it relocates one. A constant σ becomes a variance process with its own κ, θ and η; making those stochastic in turn introduces theirs. Each level pushes the arbitrariness up rather than eliminating it, and there is no level at which the regress terminates on its own.

So the stopping rule has to come from outside the modelling, and this chapter supplies it. Add a level only where the data can see it move faster than it moves — which is (21.2) again, applied to the new state rather than to a regime. A parameter promoted to a process whose variation the observations cannot resolve has not been modelled; it has been given somewhere to hide, and the fit will improve while nothing has been learned. That is the whole of this chapter in one criterion.

Remark (The Bayesian version is the natural one)**.** A regime is a discrete state, so the posterior over it is a probability rather than a distribution over a continuum, and the filter is correspondingly cheap. It needs a prior over the two states to start the recursion — the stationary distribution implied by the switching probability is the natural default, so the filter does not open biased toward

either regime. Predict by the switching probability, update by Bayes against the two observation densities, renormalise: two multiplications per observation, exactly.

What that buys is the object the Bayesian calibration above asked for. The output is not a regime but a probability of a regime, so every price computed from it is a mixture, and the width of that mixture is a confidence band that updates as each new price arrives. A desk holding a position can therefore report not only its value but how much of the value is a bet on the state estimate, which is the number that ought to size the position.

The question that decides whether any of this is worth building is not accuracy but *speed measured in data*. A regime that is detected reliably after two hundred observations, and lasts fifty, has been correctly identified and is of no use whatever. That comparison has an answer.

Calculation 21.3 (How long it takes to notice)**.** Sequential detection theory gives the expected delay to declare a change, subject to a false alarm rate α, as

$$\text{delay} \;\approx\; \frac{\ln(1/\alpha)}{D} \quad \text{observations}, \qquad D = \tfrac{1}{2}\big(r - 1 - \ln r\big), \quad r = \sigma^2_{\text{high}}/\sigma^2_{\text{low}}, \tag{21.2}$$

with D the Kullback-Leibler divergence per observation between the two regimes' return densities. The numerator is the evidence that must be accumulated; the denominator is the rate at which each observation supplies it.

Running the filter on simulated switches, against the decision rule that declares a switch once the new regime's posterior crosses 0.9 — a threshold, not itself the false-alarm rate, which the simulation measures rather than assumes:[12]

Regime shift	D per observation	Measured delay	Predicted
15% → 30%	0.807	9.4	8.5
15% → 25%	0.378	16.3	17.3
15% → 18%	0.038	97.9	152.2

The measured delay tracks (21.2) for the two wide shifts and beats it for the narrow one, which is expected — the bound is asymptotic in small α and the filter is running at a false alarm rate of a third of a per cent.

Read the last row. A volatility regime moving from fifteen to eighteen takes about a hundred trading days to establish at ninety per cent confidence, which is five months. If such a regime lasts weeks, it is real, it is undetectable in time to act on, and a model containing it is describing something nobody can trade.

Structure (The screening criterion comes before the model)**.** Equation (21.2) turns a modelling question into an arithmetic one, and it should be asked before anything is built.

Is the expected life of the regime longer than $\ln(1/\alpha)/D$*?* If not, stop. No filter, Bayesian or otherwise, recovers information the observations do not contain, and the delay above is not a property of the algorithm — it is a property of the two distributions being told apart. A better estimator cannot beat it, because it is the rate at which the data distinguishes the hypotheses.

[12] `estimation::RegimeFilter`, which measures the delay against (21.2) rather than assuming it.

Three consequences follow.

The gap matters quadratically, not linearly. Expanding D for a small shift gives $D \approx (r-1)^2/4$, so halving the difference between the regimes quarters the information and quadruples the delay. Regimes that are subtle are not slightly harder to detect; they are out of reach.

Sampling faster helps, in calendar time, and only inside the model. The interval cancels out of D entirely — a return carries the same information about a volatility *ratio* whatever span it covers — so the delay in observations is fixed by the regimes alone and the delay in calendar time is that divided by the observation frequency.

The cancellation, however, needs the returns to be independent draws of variance $\sigma^2\Delta$, so that halving Δ doubles the count of observations and leaves each one's information intact. Three things go wrong with that as Δ shrinks.

Microstructure noise floors the estimate, by (21.4), so past some frequency the extra observations are measuring the spread.

The returns also stop being independent, since the same noise induces autocorrelation in them — and n dependent observations do not carry n times one observation's information, so the delay in observations stops falling as fast as the count rises.

The third is the one that is easiest to miss and hardest to repair. *The quantity being estimated changes with the scale.* Intraday volatility carries a pronounced daily pattern, opening and closing far above the middle of the session, and it responds to liquidity events that leave no mark on a daily series. So a regime in minute-by-minute volatility is not a finer measurement of the daily regime; it is a different object with its own dynamics. Sampling faster to detect a shift in daily volatility can converge quickly and confidently on a shift in something else.

Which gives the honest version of the criterion. The frequency that optimises the volatility measurement is a ceiling and not a target: sample as fast as (21.4) allows, but no faster than the scale on which the regime you care about is defined. If the regime is a monthly change in daily volatility, minute data does not shorten the delay for it, however many observations it supplies.

And the criterion generalises past regimes. Any latent structure — a switch, a jump in a parameter, a change in correlation — is worth carrying only if the data separates its states faster than the states change. That is a sharper version of this chapter's recurring complaint. Elsewhere the trouble was that the calibration set did not determine a parameter; here it is that the data cannot determine it *in time*.

Chapter 20's remark Sampling faster does not help reaches the same verdict for mean reversion, and the two are the same criterion at the level of the general statement — but not at the level of mechanism, which is worth keeping straight. There, sampling faster never helps, even under noiseless continuous observation: what a history identifies is bounded by elapsed half-lives alone, and slicing the same span more finely adds none. Here, sampling faster genuinely does help, without limit, in the idealisation — D is interval-invariant, so more observations in the same span is strictly more evidence — and it is only the three practical obstructions above that cap the benefit. Mean reversion fails to be identifiable in time as a theorem; a regime fails to be identifiable in time as a fact about noise, dependence, and scale that a cleaner market would not share.

Remark (What a subtle regime excuses, and what it does not)**.** A discrete regime that sits close to its neighbour is the coarse shadow of a continuous state: a chain with many closely spaced levels

converges, under the same kind of scaling that turns a random walk into Brownian motion, to a diffusion. Recasting the problem that way trades one difficulty for another rather than removing it. There is no longer a threshold to cross or a false alarm to control — filtering v_t in a stochastic volatility model reports a posterior mean and variance continuously, with no decision forced on any date — but the posterior is left just as imprecise as the discrete version already said it must be. A vol-of-vol too small, or a mean reversion too slow, is filtered with the same poverty of information that here reads as an undetectable regime; calculation 21.4 measures the continuous version of the same fact, that only a longer history pins down a slow coefficient, no faster sampling.

The other question — does an undetectable regime matter — has two different answers depending on which quantity is asked about, and they diverge by more than one might expect. A price computed today, Bayes-weighted across both regimes, is close to either regime's own price whenever the regimes are close and the pricing map is smooth in the parameter — true from the first day, with nothing to wait for, since it is a statement about the mixture of two close numbers rather than about how much evidence has accumulated. A position already held and hedged on the wrong assumed volatility the whole time is a different question, and here subtlety excuses nothing: the gap in realised variance after N observations grows in proportion to $N(r-1)$, linear in the very quantity whose square sets the detection delay, so the realised divergence crosses any fixed materiality threshold after about $1/(r-1)$ observations while confident detection needs about $1/(r-1)^2$. The ratio of the two diverges as the regimes converge, so for a subtle enough pair the cost is already material long before anyone could have told the two regimes apart with any confidence at all. Not knowing which regime is in force is not the same as its not mattering; it only fails to matter for the number that gets recomputed fresh each morning.

What can break that first claim is a payoff whose price is not smooth in the parameter where the two regimes sit — a barrier or a digital positioned exactly where they diverge — and there the Bayes-mixed price can be sensitive to a regime probability the data barely moves. That is a statement about the curvature of the payoff, not about the difficulty of the estimation, and it is a third and different thing from the other two: one says the model risk is invisible in the number recomputed daily, the other says it accumulates silently in the number nobody recomputes until the position is closed.

21.4 Calibrating to Prices and to History at Once

The programme suggested by (21.1) is more ambitious than fitting a surface, and it is the natural one to attempt: learn σ_θ so that the model reprices the liquid instruments *and* reproduces the dynamics the price history actually exhibits. One model, two data sources, two terms in the loss.

The question is whether that is coherent, since the two datasets live under different measures. It is, and the reason is exact rather than approximate — which makes this one of the few places in the chapter where the news is good.

Remark (Girsanov decides what is shared)**.** A change of measure alters the drift of a diffusion and leaves its diffusion coefficient untouched. That is chapter 6's theorem, and read as a statement about estimation it says: σ is the same object under the historical measure and the pricing measure, while μ is not.

The reason runs deeper than Girsanov's formula happening to spare σ: a change of measure is a reweighting of path probabilities and nothing else, so it never touches a path itself. Quadratic variation is a pathwise limit of sums of squared increments of the observed price, computed without ever taking an expectation, so it cannot see that reweighting at all: $\langle S\rangle_t$, and with it $v_t = d\langle S\rangle_t/(S_t^2\,dt)$ and $\sigma = \sqrt{v_t}$, are the same function of the same path under $\mathbb{P}$ and $\mathbb{Q}$, not two numbers that merely happen to agree. That is what licenses reading σ off a single history and carrying it into a pricing model with no translation required, and the same argument runs one level down in a stochastic volatility model: v_t and $\sqrt{v_t}$ are themselves fixed once the path is, so they too are the same pathwise object under both measures.

What a change of measure moves is v_t's own drift, and through it which of v_t's paths are the likely ones — never the value v_t takes on a path already fixed. The confusion, if there is one, is forgetting that the Brownian motion driving v_t has changed too: Girsanov relates the two by $dW_t^{\mathbb{Q}} = dW_t^{\mathbb{P}} - \lambda_t\,dt$, and substituting one for the other inside v_t's equation exactly cancels the drift difference, leaving $v_t(\omega)$ untouched on every path. $\mathbb{P}$ and $\mathbb{Q}$ do not disagree about where v_t goes; they disagree about which process counts as the driving Brownian motion. §21.7 works out exactly which drift parameters move, and by how much.

Under the pricing measure the drift is not a free parameter at all — it is r for a tradeable, or the drift condition of chapter 8 for a curve. So of the two coefficients in (21.1), one is determined by no-arbitrage and the other is shared between the two measures. The only genuinely free object linking them is the market price of risk $\lambda = (\mu^{\mathbb{P}} - \mu^{\mathbb{Q}})/\sigma$.

That much is standard. What makes the combination pay off is a fact about what a price history can and cannot measure.

Calculation 21.4 (What a history identifies, and what no amount of it does)**.** Observe a diffusion with drift 5% and volatility 20% over a fixed window of one year, sampled n times, and estimate both coefficients. Averaged over four hundred independent histories:[13]

Samples in the year	Volatility error	Drift error
250	3.64%	0.158
2 500	1.12%	0.155
25 000	0.36%	0.159

The two columns do not merely differ in rate. The volatility error falls like $\sqrt{2/n}$ and can be driven as low as the data allows. The drift error does not move at all — not slowly, but exactly not at all — because the estimator $(X_T - X_0)/T$ is a function of the endpoints and the intermediate samples are not merely uninformative about the drift, they do not appear in it. Its standard error is $\sigma/\sqrt{T}$ regardless of n.

Only a longer window helps: sixteen times the history quarters the drift error, the same $1/\sqrt{T}$ law applied to T instead of n.

One caveat, which the next section is about: the volatility column says what a *diffusion* gives up under sampling, and a price is not a diffusion when looked at closely enough. The improvement stops, and then reverses.

[13] `estimation::identification_by_frequency`.

Structure (The division of labour is exact)**.** Put the two facts together and they interlock neatly.

The parameter a price history estimates *well* — arbitrarily well, given enough sampling frequency — is the diffusion coefficient. The parameter it estimates *not at all* over a fixed window is the drift. And by Girsanov the diffusion coefficient is precisely the one shared with the pricing measure, while the drift is precisely the one no-arbitrage has already determined there.

The payoff is a constructive answer to this chapter's own complaint. Every under-identified parameter these notes have found is a parameter of the *diffusion* structure — β in chapter 10, the mixing weight in chapter 11, the mean reversion in chapter 12, the correlation between forwards in chapter 13. All of them are measure-invariant. So all of them are in principle visible in the time series, which is a different dataset from the one that failed to determine them, and Calculation 21.4 says the time series is the good direction to look. When this chapter says the fix for an unidentified parameter is more information, this is where the information is.

Remark (Three reasons it is harder than that)**.** The argument above is correct and it is not a recipe. Three things stand between it and a working joint calibration.

The processes are latent. Quadratic variation identifies the volatility of something *observed*. The parameters that are unidentified belong mostly to processes that are not: the variance process itself in a stochastic volatility model, which regime is in force, the state of a multi-factor curve. Estimating those from prices is a filtering problem rather than a sum of squared increments, which is why the particle filter appears earlier in this chapter, and the rates degrade accordingly.

Some of them are drift parameters of a latent process. A mean reversion is the drift of the variance or of the curve factor, so it inherits Calculation 21.4's bad column: it needs a long history rather than a finely sampled one, and this chapter has already measured that a mean reversion fitted to a finite sample comes out biased fast. Sampling faster does not touch it.

The over-identifying restriction can fail, and its failure is information. If both sources speak about σ, they can disagree — and they routinely do, because implied volatility exceeds realised volatility on average. That gap is not an error in either estimate; it is §21.7's subject. But it means a joint calibration that assumes the two agree will corrupt a σ the option prices had identified correctly, and one that does not assume it must model the discrepancy.

So the honest summary is narrower than the structure above suggests. Joint calibration adds genuine information about the diffusion structure, which is where these notes keep finding the gaps. It buys that at the price of requiring a model for the risk premium, and it should be run as a *test* before it is run as a fit: fit to prices alone, then ask what the historical dynamics of the fitted model would look like and whether they resemble the ones observed.

21.5 Estimating a Volatility

The previous section leaned on volatility being the easy parameter. Four decisions have to be made before a number comes out, and each of them moves it by more than the precision anybody quotes.

Over what window

Volatility is not constant, which is the premise of five chapters of these notes, so a sample average over a window is an average of something that moved. A long window has small sampling error and estimates a blend of regimes; a short one tracks the current regime and is noisy. There is no correct answer, only a stated one, and the honest practice is to report the window alongside the number and to check that the conclusion does not depend on it.

The choice interacts with the use. A volatility feeding a risk model over a ten day horizon and a volatility feeding a five year swaption are not the same quantity even if both are called annualised volatility, and using one for the other is the commonest way a risk number and a pricing number come to disagree for reasons nobody can locate.

At what resolution, and the answer is the hedging frequency

The sampling interval looks like a purely statistical choice and it is not. Chapter 5 showed that the profit and loss of a delta hedged position over an interval is

$$\tfrac{1}{2}\Gamma S^2\left(\left(\frac{\Delta S}{S}\right)^2-\sigma^2\Delta t\right),$$

so what a hedger is paid, interval by interval, is the difference between the *squared move over their rebalancing interval* and the variance they sold. Accumulated over the life of the option, the quantity that determines whether the hedge made or lost money is the realised variance sampled at the rebalancing frequency.

That answers the question. A desk rebalancing daily is exposed to daily realised variance and should estimate it from daily data, and one rebalancing hourly to hourly variance. The two are not the same number — if they were, the signature plot below would be flat — and the difference is not an estimation error to be minimised but a genuine difference in what is being hedged.

The square root is not free

Almost every estimator computes a variance and roots it. The sum of squared increments is unbiased for the variance; the square root is concave, so by Jensen the volatility that comes out is biased *low*.

Calculation 21.5 (The correction, in closed form)**.** For n increments of a driftless Gaussian, $n\hat{\sigma}^2/\sigma^2$ is chi-squared with n degrees of freedom, so

$$\mathbb{E}[\hat{\sigma}]=\sigma\sqrt{\frac{2}{n}}\,\frac{\Gamma\!\left(\frac{n+1}{2}\right)}{\Gamma\!\left(\frac{n}{2}\right)}\;\approx\;\sigma\left(1-\frac{1}{4n}\right). \tag{21.3}$$

At twenty increments — a month of daily data — the factor is 0.98758, so the volatility comes out 1.24% too low, which on a 20% volatility is a quarter of a point.[14]

[14] `estimation::sqrt_bias_correction`, with `a simulation confirming that dividing by it removes the bias` rather than merely that the formula is what it is.

The bias is one-signed, so averaging across months does not remove it, and it is available in closed form, so there is no reason to carry it. It is the smallest of the four difficulties here and the only one with an exact answer.

And a price is not a diffusion when looked at closely

The serious difficulty is the one that contradicts the previous section's arithmetic. What is observed is not the efficient price but the price at which a trade happened, which alternates between the bid and the offer. Write the observation at each sampled time as $Y_i = X_i + \epsilon_i$, with ϵ_i independent noise of scale η, roughly the half spread. An observed increment is then $\Delta Y_i = \Delta X_i + (\epsilon_i - \epsilon_{i-1})$, a difference of two independent noise draws rather than one, so its noise variance is $2\eta^2$ regardless of how short the sampling interval is:

$$\mathbb{E}\big[\text{realised variance over } n \text{ samples}\big] = \sigma^2 T + 2n\eta^2. \tag{21.4}$$

The bias is *linear in the sampling frequency*. Sampling faster does not converge on the truth, it diverges from it. Consecutive increments also share a noise draw with opposite sign — ϵ_i enters ΔY_i positively and ΔY_{i+1} negatively — which induces a negative autocorrelation between adjacent returns, the same dependence that later stops sampling faster from helping a regime be detected any sooner.

Calculation 21.6 (The volatility signature plot)**.** A 20% volatility over one trading day, observed with a one basis point half spread.[15]

Samples in the day	Measured volatility	Error
6 (hourly)	19.09%	−4.5%
26 (quarter hourly)	19.86%	−0.7%
78 (five minutes)	20.05%	+0.3%
390 (one minute)	20.48%	+2.4%
1 560 (fifteen seconds)	21.87%	+9.3%
7 800 (two seconds)	28.16%	+40.8%

At two second sampling the measured volatility is half as large again as the truth, and it is not converging — it is measuring the spread.

Structure (Two biases pointing opposite ways)**.** The table is not monotone, and the reason is the useful part.

At coarse sampling (21.3) dominates: few increments, so the square root pulls the estimate down, and hourly sampling is four and a half per cent low. At fine sampling (21.4) dominates and the estimate runs away upwards. The two biases point in opposite directions and there is a frequency where they cancel — here around a quarter of an hour, and the five minute sampling that the literature recommends is within a fraction of a per cent of the truth for the same reason.

So the standard advice to sample every five minutes is not a rule of thumb about market conventions. It is where two known biases of known size happen to cross for typical parameters,

[15] `estimation::realised_variance_with_noise`, measured against (21.4) rather than against intuition.

and (21.3) and (21.4) say how it moves: a wider spread pushes the optimum coarser, a shorter horizon pushes it finer. A desk trading an instrument with a ten basis point spread and using five minute sampling because that is what the papers say is making an error of a hundred times the size the papers were addressing.

It also settles the previous section's overreach. The claim that a fixed history identifies volatility arbitrarily well is a statement about a diffusion. Against real observations the error falls, reaches a floor set by the spread, and then rises, so the identification argument holds down to the noise and no further — which is still a great deal better than the drift, whose error never falls at all.

Remark (What is done about it)**.** Three responses, in increasing order of effort. Sample at the frequency where the biases cross and accept the residual, which is what most desks do and which Calculation 21.6 shows is defensible. Subsample and average — compute the realised variance on several offset grids and average them, which reduces the variance without changing the bias. Or use an estimator built for the problem, and there are two standard ones, both exploiting the same fact: the bias in (21.4) is exactly proportional to the increment count, so two realised variances computed on the *same* path at two different frequencies carry the same noise process in a known, fixed ratio.

Two-scale realised volatility takes that ratio literally. Average the realised variance over k coarse, offset subsamples of the path to estimate the low-frequency side; take the ordinary realised variance over the whole path for the high-frequency side; and subtract the two in the proportion (21.4) predicts. The noise term cancels algebraically rather than merely being averaged over.[16]

A realised kernel works on the same increments directly. Ordinary realised variance is the increments' own sample autocovariance at lag zero; the noise contributes a matching autocovariance at lag one, of the opposite sign, because adjacent increments share a noise draw with opposite signs, as noted above. Adding the higher lags back in, at a weight that shrinks smoothly to zero by some bandwidth, cancels the bias instead of leaving it in — the negative autocorrelation used constructively rather than merely diagnosed.[17]

The third route is the correct one: the noise is a known functional form with an unknown scale, so it can be estimated and removed rather than avoided. Equation (21.4) has η in it, and anything with a parameter in it can be fitted.

Calculation 21.7 (What the two estimators buy, on the same path)**.** The same simulated day as Calculation 21.6, now measured three ways at once: naive realised variance, two-scale realised volatility, and a realised kernel.[18]

Samples in the day	Naive	Two-scale	Kernel
26 (quarter hourly)	20.06%	19.31%	19.95%
78 (five minutes)	20.12%	19.82%	20.08%
390 (one minute)	20.50%	19.93%	19.99%
7 800 (two seconds)	28.16%	20.00%	20.07%

[16] Zhang, L., Mykland, P. A., & Aït-Sahalia, Y. (2005). A tale of two time scales. *Journal of the American Statistical Association*, 100(472), 1394–1411.

[17] Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2008). Designing realised kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6), 1481–1536.

[18] `estimation::noise_robust_estimators`, with `both staying within a few percent of the truth where the naive estimator does not`.

At the finest sampling, where the naive estimator is measuring the spread rather than the volatility, both alternatives are within a tenth of a percentage point of the truth — the whole point of building an estimator around the bias rather than merely choosing a frequency to tolerate it at.

21.6 Estimating a Correlation

Chapter 13 produced an uncomfortable result: the correlation between rates moves a swaption volatility by percentage points, while the choice between two whole model families moves it by hundredths. If that is right — and it is measured — then the correlation estimate is the most consequential number in the whole apparatus, and it deserves more care than it usually gets.

It is also the hardest thing in this chapter to estimate, for a reason that is dimensional rather than statistical.

Calculation 21.8 (A correlation matrix of pure noise)**.** Take forty independent series — the forwards of a ten year quarterly structure, if they had no correlation at all — and estimate their correlation matrix from one year of daily observations. The truth is the identity, so every eigenvalue is one. The estimate is not close.[19]

Series	Observations	Largest	Smallest	Ratio
40	250	1.84	0.39	4.7×
40	1 000	1.40	0.66	2.1×
40	4 000	1.19	0.83	1.4×

At a year of data the largest apparent factor looks nearly five times the smallest, in a matrix with no factors in it whatever. Anyone reading a scree plot of those eigenvalues would report a rich factor structure and would be reporting nothing.

Remark (The spread is predictable, which is what makes it usable)**.** The numbers above are not accidents of the simulation. Estimating p series from n observations means extracting $p(p-1)/2$ numbers from pn data points, and the ratio $q = p/n$ controls everything. The Marchenko-Pastur law says the sample eigenvalues of independent series fill the interval

$$\left[(1-\sqrt{q})^2,\ (1+\sqrt{q})^2\right], \tag{21.5}$$

and the measured columns above sit against those edges to within a few per cent.

That is more useful than a warning, because it supplies a null hypothesis. An eigenvalue *inside* (21.5) is consistent with noise and carries no information; one above the upper edge is a factor that is really there. So a scree plot becomes a test rather than an impression: draw the upper edge on it, and count what pokes out. On the US Treasury curve — the only one this book's committed data can put through the test — exactly two eigenvalues clear the edge: level and slope, on every panel tested, including four decades of daily observations spanning several rate cycles rather than one calendar year.[20] That is chapter 24's level and slope, now established rather than asserted.

[19] `estimation::sample_correlation_spectrum`, against the Marchenko-Pastur edges rather than against an impression.

[20] `risk::only_level_and_slope_clear_the_noise_edge_on_the_full_panel` and `risk::only_level_and_slope_clear_the_noise_edge`

The third shape a curve's principal component analysis always returns, curvature, sits inside the noise bulk on every panel tested; it is a real direction to trade, as chapter 22's butterfly shows, but not one this test can distinguish from what forty-four years of pure noise would also produce. Whether a curve with more cross-sectional structure — a swap curve segmented by credit and liquidity tier, say — would put a third eigenvalue above the edge is a real question this book's data cannot answer; the count of two is a property of the government curve, not asserted as a property of curves in general.

The remedy follows the diagnosis. Keep the eigenvalues above the edge, replace the bulk with its average, and rebuild the matrix; or shrink the sample matrix towards a structured target, which is the same operation performed continuously rather than by a cutoff. Both are ways of refusing to estimate what the data cannot support.

Structure (Three routes, and the market model takes the third)**.** There are only three things one can do about Calculation 21.8.

More data. It works and it works slowly: the excess of the top eigenvalue over one falls like $\sqrt{q}$, so sixteen times the history roughly quarters it. Sixteen years of daily data also spans several regimes, so what is gained in sampling error is lost in the assumption that the correlation was constant — which returns to the window problem of the previous section, now with sharper consequences.

Cleaning. Estimate the matrix, then remove the part attributable to noise, by eigenvalue clipping or by shrinkage towards a target. This is the honest general-purpose answer and it makes no structural assumption beyond the choice of target.

Parametrise. Assume a form. Chapter 13 uses $\rho_{ij} = e^{-\beta|T_i - T_j|}$, one parameter for the whole matrix, which cannot overfit because there is nothing to overfit with. The cost is that it is wrong in a specific way — real curve correlations do not decay to zero at long separations, they flatten out at a positive level, so the exponential understates the coupling between the long end and the very long end. A two parameter form $\rho_{ij} = \rho_\infty + (1 - \rho_\infty)e^{-\beta|T_i - T_j|}$ fixes exactly that and is what is generally used.

The third route is the right one here and the reason is Calculation 21.8 rather than parsimony for its own sake. Forty forwards from a year of data cannot support a free matrix; the estimate would be noise wearing the shape of structure, and chapter 13 showed that this is the parameter the price is most sensitive to. When the data cannot identify a parameter and the answer depends on it, imposing a form is not laziness — it is the only alternative to letting the noise choose.

Calculation 21.9 (Checking the form against the same test that counted the factors)**.** A two parameter form is chosen for parsimony, but Calculation 21.8's own machinery can ask a sharper question: does it match the *shape* a principal component analysis finds, or only survive by having too few parameters to fail? An equicorrelation floor ρ_∞ is, by itself, a rank-one matrix with a flat eigenvector — a level factor. An exponential decay on top of it contributes a second, maturity-ordered direction — a slope. So the two parameter form's own structure predicts exactly two real dimensions, which is either a coincidence or a genuine match to the remark above's finding of exactly two, on four decades of Treasury data.

Fit ρ_∞ and β to the sample correlation matrix of daily changes across the committed eight tenors, by least squares on the off-diagonal entries, then compare the fitted matrix's own eigenvalues

to the real ones.[21]

	Real	**Fitted, $\rho_\infty = 0.51$, $\beta = 0.19$**
Level	6.28	6.40
Slope	1.12	0.80
Third (noise)	0.27	0.34

The level eigenvalue is matched to within two per cent, and neither matrix puts a third eigenvalue above the noise edge of 1.05 — the rank is right, not merely small enough to be safe. The slope eigenvalue is not matched: the fit understates it by about a quarter. A single exponential has one parameter left after ρ_∞ pins the floor, and that parameter cannot simultaneously set the decay's shape and match how much independent variance the real slope carries — so the form gets the number of real factors right and gets one of their sizes wrong, which is a more precise criticism than "it is only an approximation" and a more precise defence than "it captures the two real factors."

Remark (And it is not constant, in the direction that hurts)**.** One further problem, which no amount of cleaning addresses. Correlations move, and they move *up* in stress: rates that normally decouple stop decoupling when everything is being sold at once.

So an estimate from a calm sample understates the correlation in exactly the scenario a risk calculation is about, and chapter 18 has already made the general form of this point — the coupling in the tail is not the coupling in the body, and a single correlation number cannot carry both. For pricing, the calm number is arguably the right one, since it is the average over the option's life that matters. For risk it is not, and chapter 24's stress work has to use something else. Reporting one number for both uses is the error, rather than any particular estimate of it.

21.7 Why Implied Exceeds Realised

The gap is large, persistent and one-signed, so it is not a misfit to be calibrated away. Two explanations are usually offered and they are frequently presented as alternatives. They are not, and separating them matters for §21.4, because they imply different things about whether a joint calibration is well posed.

It is a risk premium, and it sits where Girsanov says it must

The first explanation is that variance is a priced risk factor. A position that is long volatility pays off when volatility spikes, and volatility spikes when everything else is going wrong, so it is a hedge — and a hedge earns a negative expected return, which is to say it costs a premium to hold. Options are that hedge, so they are expensive relative to what the underlying goes on to do.

This is formally the same object as the equity risk premium and belongs in the same place. Under the historical measure a stock drifts faster than a bond because holding it is uncomfortable; under the pricing measure it does not. Here the variance process drifts differently under the two measures for the same reason.

[21] `risk::fit_maturity_correlation`, with the comparison measured against the same panel Calculation 21.8 and the noise-edge remark use.

Calculation 21.10 (The premium, written down)**.** Take the Heston variance of chapter 10 under the historical measure, and let the market price of variance risk be proportional to volatility, so that changing measure adds $-\lambda v\,dt$ to the variance's drift. Then

$$\kappa^{\mathbb{Q}} = \kappa + \lambda\eta, \qquad \theta^{\mathbb{Q}} = \frac{\kappa\theta}{\kappa^{\mathbb{Q}}}, \tag{21.6}$$

and η is unchanged. Take a 20% long-run volatility, mean reversion $\kappa = 2$, vol-of-vol $\eta = 0.3$, and the variance starting at its own long-run level so that nothing below is confused with a transient:[22]

λ	$\kappa^{\mathbb{Q}}$	**One year implied**	**Premium**
-1	1.70	20.90%	0.90 pts
-2	1.40	21.89%	1.89 pts
-3	1.10	23.00%	3.00 pts

against a realised expectation of exactly 20%. The observed gap on an equity index is one to three volatility points, so a market price of variance risk of a few units reproduces it — the explanation is the right size, which is the first thing to check about any explanation.

Structure (The premium cannot be in the vol-of-vol)**.** If implied volatility carries a risk premium, it is tempting to suppose that vol-of-vol carries one too — that η is higher under the pricing measure the way θ is. It is not, and it cannot be. η is a diffusion coefficient, and a change of measure moves drifts and leaves diffusion coefficients alone. The premium has nowhere to go except into κ and θ.

That is §21.4's division of labour applied one level down, and it has a consequence that is directly useful. Vol-of-vol is measure-invariant, so it is *jointly identified*: an η estimated from the time series of realised volatility and an η implied by the curvature of the smile are estimates of the same number, and they can be compared. If they disagree, something is wrong with the model rather than with the premium, because no premium can explain a discrepancy in a diffusion coefficient. This is one of the few genuinely testable cross-measure restrictions in derivatives pricing, and it is almost never tested.

It is one-sided flow, and dealers cannot arbitrage it away

The second explanation makes no reference to risk aversion. Banks are structurally short options, because their clients — corporates hedging liabilities, insurers hedging guarantees, funds buying protection — want to own them. A dealer who cannot lay the position off must hold it, and chapter 23 derives what happens next: a market maker holding inventory it did not choose quotes to be paid for holding it. The premium is then a price for absorbing one-sided demand, in exactly the sense chapter 22 requires of a trade — somebody is on the other side for a reason that is not an opinion.

This explanation is not a rival to the first. It is a second premium, paid for a different service, and both are collected by the same position. A single calibrated λ cannot tell them apart, though: matching the whole observed gap makes λ absorb whatever share of it is flow, not only the risk premium it is meant to represent, and the remark on telling them apart below says what to check before trusting it.

[22] `heston::variance_premium`, with the expected integrated variance in closed form rather than simulated.

And liquidity, which is why neither gets arbitraged away

A premium of one to three volatility points invites the obvious question: why does nobody sell it until it is gone? The answer is the third component, and it is not a separate explanation so much as the reason the first two survive.

Harvesting the premium means selling options and delta hedging to expiry. The round trip costs the option's bid-offer, which on a liquid index swaption is a fraction of a volatility point and on anything else is more, plus the accumulated cost of rebalancing the delta, which chapter 23 prices and which grows with how often one rebalances — and Calculation 21.6 has just shown that rebalancing frequency is not a free choice either. A premium smaller than the cost of collecting it is not an arbitrage; it is a fee for a service that is genuinely expensive to provide.

That produces a floor. *A model error smaller than the bid-offer spread is not a model error anyone can trade*, so it cannot be arbitraged away and there is no force driving it to zero. Chapter 13 used exactly this to settle the choice of coordinates: an approximation two orders of magnitude below the spread is free. The same reasoning says the variance premium, being larger than the spread but not by much, is exactly in the range where it persists — big enough to be real, small enough that collecting it is a business rather than an arbitrage.

Liquidity also enters as a premium in its own right, and in the same direction. An instrument that cannot be exited cheaply trades at a discount, which is to say a higher expected return, and chapter 22's on-the-run spread is that discount measured directly. For options the effect concentrates where it is least convenient: far strikes and long tenors, which have the widest spreads, the thinnest volume and the least reliable marks — and which are also where a model is least constrained by quotes. So the part of the surface where calibration has the least to say is the part where the price is least trustworthy, and the two failures compound rather than offset.

Remark (Volume is a poor proxy and it is the one everybody uses)**.** Since liquidity is not directly observable, it is proxied — usually by volume, sometimes by quoted spread, occasionally by the price impact measures of chapter 23. Volume is the worst of the three and the most used, because it is high in two opposite circumstances: when an instrument is easy to trade, and when everybody is trying to get out of it at once. A liquidity signal built on volume therefore reads its highest exactly when liquidity is about to be worst, which is the failure mode chapter 22's six trades all share.

Quoted spread is better and still incomplete, since a quoted spread is for a quoted size and the size is what matters in a stress. The honest measure is the one chapter 23 builds, price impact per unit of quantity, and it requires data most participants do not have.

Remark (How to tell them apart, and why it matters here)**.** The two have different signatures, and the distinction is testable rather than philosophical.

A risk premium is compensation for an exposure, so it should be roughly proportional to the quantity of variance risk borne and should have a term structure fixed by the model — Calculation 21.10 shows it emerging over the mean reversion time, so it grows with horizon in a shape κ determines. It should be present in every market where variance is risky, in proportion to how risky.

A flow premium is compensation for absorbing inventory, so it should track *dealer capacity* rather than risk: largest where client demand concentrates, which is downside strikes and specific

tenors rather than uniformly across the surface; widening when balance sheet is constrained, at quarter ends and after dealer losses; and varying across markets with who happens to be buying rather than with how volatile they are.

Both signatures are visible in the data, which is the honest answer to which explanation is right: there is a persistent baseline consistent with a risk premium, and a strike- and tenor-dependent component that moves with demand.

For §21.4 that distinction is decisive rather than academic. If the gap is a risk premium, λ is a smooth function of the state with few parameters, it can be modelled, and a joint calibration to prices and history is well posed — one estimates the shared diffusion structure and the premium together. If the gap is flow, then the quantity separating the two measures depends on dealer positioning, which appears in *neither* dataset. A joint fit will then attribute a flow effect to dynamics, and the dynamics is what it was trying to learn. So the programme's viability rests on how much of the gap is which, and that question has to be answered before the fit rather than by it.

Some of that positioning is not entirely unobservable, only absent from these two datasets specifically. Exchange-reported open interest and volume by strike and tenor proxies for where client demand concentrates; dealer CDS spreads and balance-sheet disclosures proxy for the capacity constraint directly, since a dealer's own cost of capital is what widens at quarter ends and after losses; and CFTC positioning reports do the same on the futures side of the same trade. None of it substitutes for prices and history — open interest in particular inherits the remark above on volume, reading highest exactly when it is least informative — but each is a genuinely third source, carrying the one thing prices and history structurally lack, and answering the risk-versus-flow question before the fit has to be built from data like this rather than from more of what the fit already has.

The split is worth making because the two carry different consequences, not only different causes. A risk premium should persist wherever variance is priced, so a book built to harvest it can be sized and stressed like any other risk exposure; a flow premium can vanish with the client demand or dealer capacity that created it, so the same one to three points of income calls for a liquidity stress rather than a volatility one, and for sizing that tracks the business rather than the risk. The genuine way to do the split is to regress the observed gap, by strike and tenor, on the risk premium's own term structure alongside the capacity and demand proxies above, and let the two explain what they actually explain.

21.8 What To Take Away

Three things carry over from chapter 20, sharpened by what a latent state adds.

First, speed is not knowledge. A neural surrogate for a pricing map, a particle filter, a posterior sampled instead of optimised — each turns a slow calculation into a fast one, and none of them extracts information the quotes or the history did not contain. The flat valley in a SABR fit is exactly as flat once the fit is instant.

Second, not everything travels between the two measures, but more does than is usually assumed. A diffusion coefficient is the same object under $\mathbb{P}$ and $\mathbb{Q}$, so it can be estimated from

history and checked against what is implied; a drift, a mean reversion, or the risk premium in the gap between implied and realised volatility cannot make that trip, because a change of measure is precisely what moves them.

Third, a latent state is usually a smaller problem than it is made out to be, and a real one is a race. Continuous, noiseless observation of a price would reveal its own quadratic variation and, with it, the variance a model called hidden — but no real market grants that for free. Microstructure noise floors the estimate long before the sampling interval reaches zero; a jump adds its own squared size to quadratic variation and must be separated from the diffusive part rather than counted as more of it; and what quadratic variation converges to is itself a different object at every scale, so sampling faster answers a question about a finer regime rather than the coarser one more precisely (Calculation 21.3's third consequence, and the hedging-frequency argument earlier in this chapter). What remains genuinely unknown is what discreteness, noise, jumps, and scale destroy — and a change worth modelling is one the data can tell apart from noise faster than the change itself unfolds, at the resolution the change is actually defined at.

The tools for computing have changed enormously. The questions have not.

References

- Doucet, A., de Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer.

- Horvath, B., Muguruza, A., & Tomas, M. (2021). Deep learning volatility. *Quantitative Finance*, 21(1), 11–27.

- Buehler, H., Gonon, L., Teichmann, J., & Wood, B. (2019). Deep hedging. *Quantitative Finance*, 19(8), 1271–1291.

- Han, J., Jentzen, A., & E, W. (2018). Solving high-dimensional partial differential equations using deep learning. *PNAS*, 115(34), 8505–8510.

- Guyon, J., & Lekeufack, J. (2023). Volatility is (mostly) path-dependent. *Quantitative Finance*, 23(9), 1221–1258.

- Perez Arribas, I., Salvi, C., & Szpruch, L. (2020). Sig-SDEs model for quantitative finance. *Proceedings of the First ACM International Conference on AI in Finance*.

- Zhang, L., Mykland, P. A., & Aït-Sahalia, Y. (2005). A tale of two time scales. *Journal of the American Statistical Association*, 100(472), 1394–1411.

- Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2008). Designing realised kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6), 1481–1536.

Part V

Trading and Risk

Chapter 22

Relative Value and the Basis Trade

We look at what the other side of the industry does with the same machinery. A relative value desk is not trying to price a derivative correctly; it is trying to find two things that should be worth the same and are not. We derive the decomposition of a fixed income return into carry, roll-down and yield change, construct the curve trades that isolate one factor from another, and work through the cash-futures basis in enough detail to see why the trade exists and why it periodically destroys the people doing it. Who is on the other side, and why they stay there, is the question the rest of the chapter returns to: the same ideas cover selling volatility and what that position is short, the retail versions of these trades, and how much a signal about a future release is worth, which is a mutual information measured in nats and sized by a log-optimal rule.

22.1 What This Chapter Is

Everything so far has been derivation. A model was posited, a measure found, a price computed, and the result was true given the assumptions. This chapter describes trades that people put on to make money.

The subject here is not whether these trades work. It is what they *are*: what a position consists of, how its profit and loss decomposes, and — the part that is genuinely mathematical and genuinely useful — where the risk actually sits, which is frequently not where it appears to sit. A trade whose P&L looks like a small steady income and whose risk is a rare enormous loss is a recognisable object, and recognising it is worth more than any view about whether it is currently cheap.

The reader who wants to know whether to put a trade on will not find it here. The reader who wants to know what they are holding will.

Structure (What makes a spread a trade)**.** A model tells you a spread is wide. That is not a signal, and treating it as one is the characteristic error of the subject. Two further things are required before a wide spread is a trade, and neither is a modelling output.

A structural reason. Somebody has to be on the other side for a reason that is not an opinion — a regulatory constraint that forces a pension fund to hold a particular instrument, an index rule that requires a bond to be sold on a date, a balance sheet cost that makes a dealer unwilling to warehouse a position, a settlement convention that nobody can arbitrage away. When such a reason exists, the spread is a price being paid for a service and can be collected. When it does not, the wide spread is more likely to be the model's error than the market's, and §22.9's negative number is the case where telling the two apart is the whole problem.

A horizon. Convergence is not a date. A spread that is two standard deviations wide and genuinely mean reverting still takes longer to close than the intuition suggests, and goes further against the position first. That is a question about first passage times rather than about forecasting, it is answerable, and the next section answers it — because the usual way of getting it wrong is not misjudging direction but misjudging how long being right takes.

The model's job in all of this is narrower than it looks. It supplies the spread and the decomposition of the P&L; it has nothing to say about either of the two conditions above. Which is why the discipline in relative value work lives in chapter 20's estimation rather than in the pricing.

22.2 Who Is On the Other Side

The structural reason the structure above demands is not an abstraction. It belongs to a specific institution, with a specific mandate and a specific regulator, and knowing who that is — and what they are required, rather than persuaded, to do — is most of what separates a genuine relative value trade from a model's own noise.

Repo, which is how balance sheet actually gets priced

Almost everything below is financed the same way, so the mechanism comes first, before the players.

Definition 22.1 (Repurchase agreement)**.** A repo is a sale of a security today, combined with an agreement to repurchase the same security at a fixed later date for a fixed, slightly higher price. Economically it is a collateralised loan — the seller borrows cash and posts the security as collateral — structured as two trades rather than one so that, in a bankruptcy, the lender already owns the collateral rather than standing in a queue for it. The difference between the two prices, annualised, is the *repo rate*: the interest paid on the cash. The lender advances less than the collateral's market value, the *haircut*, to stay protected if the collateral's price falls before it can be sold.

Most repo is *general collateral*: the cash lender does not care which bond is posted, any similar one will do, and the rate sits close to a short-term benchmark. A bond becomes *special* when enough people need that specific security — to deliver into a short, to meet a settlement obligation, because it is the benchmark issue hedges are executed in — that they will accept a lower return on their cash purely to secure it. The repo rate on a special bond falls below the general collateral rate, and in a genuine squeeze it can fall to zero or below: the cash lender is then paying to lend, because what they are actually buying is the bond, not the yield. That is the mechanism behind §22.11's on-the-run richness and what happened to Salomon's target note in 1991: a repo rate is not a property of the borrower's credit, it is a price for a specific piece of collateral, and it moves with demand for that collateral exactly like any other price.

The players

Pension funds owe payments decades into the future, frequently linked to inflation or wages, and are marked against those liabilities' own discounted value rather than against a market index. That creates a structural demand for very long duration and for inflation exposure, in quantities a fund's own assets rarely supply directly — which is why liability-driven investing uses swaps and long gilts or Treasuries, often levered through repo, to close the gap cheaply rather than buying enough of the underlying bond outright. The lever is exactly what turns a slow, structural trade into a fast one under stress: in September 2022 a rise in UK gilt yields fell on levered liability-driven investing positions as a fall in collateral value, triggering margin calls that could only be met by selling gilts, which pushed yields higher and triggered the next round of calls — the same feedback loop the mortgage convexity hedging entry of §22.11 describes, with a margin call standing in for a rebalancing trigger and the Bank of England's emergency purchases standing in for the intervention that finally broke it.

Insurers face a related but differently shaped constraint. Solvency II in Europe and risk-based capital rules in the US charge capital against a mismatch between the timing of assets and the timing of liabilities, so an insurer's own capital cost falls when it buys assets that match liability cashflows closely — long-dated credit, structured or illiquid assets that qualify for favourable treatment under the rules, specifically because they match rather than because they are attractively priced. The constraint creates demand for an asset *shape*, and a security that fits that shape can trade rich to an otherwise identical one that does not, for the same reason the on-the-run bond does.

Banks and dealers are constrained by capital and leverage-ratio rules, most acutely at quarter ends when those ratios are reported, and by the balance sheet a repo book requires to warehouse anything. Chapter 23 works out the consequence for quoting; here the consequence is that a dealer is structurally reluctant to hold inventory, structurally short the optionality clients want to buy, and the natural supplier of exactly the balance sheet covered interest parity's cross-currency basis prices.

Hedge funds and relative value desks are the balance sheet that steps into the gap dealers leave — unconstrained by the same regulatory capital rules, but constrained instead by financing that can be pulled and by investors who can redeem, both on a timescale shorter than the trade's own horizon. That asymmetry is the horizon condition named above, in institutional form: the trade is right, and the capital available to hold it is not guaranteed to survive as long as being right takes.

Corporates issue the debt the rest of this list trades around, and hedge the interest rate and currency exposure that issuing it creates — paying floating and receiving fixed on newly issued debt, or the reverse, and buying the caps that bound a floating liability. That one-directional demand — borrowers wanting caps, issuers wanting the option to call — is exactly the flow "Selling Volatility, and What That Is Short" identifies as what leaves the dealer community structurally short optionality and paid to be.

Asset managers run mandates benchmarked to an index rather than to a liability, and are marked to market continuously rather than against a discounted cashflow, which pushes them towards the more capital-efficient instrument for a given exposure — futures over cash bonds, a swap over a bond where the mandate allows it — exactly the preference behind the cash-futures basis below.

None of this is a claim that any of these institutions is behaving irrationally. Each is optimising something — a funding ratio, a capital ratio, a tracking error — that is not the price of the instrument it trades, and §22.11 is a catalogue of what is left over when several of them optimise different things against the same market.

22.3 How Long Being Right Takes

Take the simplest possible model of a converging spread: an Ornstein-Uhlenbeck process, as in chapter 20,

$$dX_t = -\kappa X_t\,dt + \sigma\,dW_t, \tag{22.1}$$

entered when X sits two stationary standard deviations from zero, held until it returns. The mean reversion is certain — this is not a case where the trade might be wrong about direction. Everything below is what happens when it is right.

Calculation 22.2 (First passage is not the half-life)**.** The half-life $\ln 2/\kappa$ is the number everyone quotes and it answers a different question. It says how fast an *expectation* decays: $\mathbb{E}[X_t] = X_0 e^{-\kappa t}$ halves in that time. It does not say how long a path takes to reach zero, and the two are not close.

Simulating (22.1) from two standard deviations, with a one-year half-life:[1]

half-life $\ln 2/\kappa$	1.00 years
mean time to first reach the mean	2.10 years
mean worst level reached first, in deviations	2.48

So the trade takes more than twice the half-life, and before converging it goes about half a deviation further against the position than the level it was entered at. Neither number is available from the half-life, and both are what size the trade.

Remark (Where a stop turns a winning trade into a loss)**.** The mean worst excursion is an average, so a stop placed at it is hit about half the time. Placing it closer is worse:

Stop, in deviations	**Fraction stopped out before converging**
2.5	39%
3.0	12%
4.0	under 1%

Entered at two deviations, a stop half a deviation away loses a trade that was right about direction two times in five. At twice the entry width the stop stops binding altogether, and the trade's risk becomes the holding period rather than the loss.

That is the regime to be in, and reaching it is a sizing decision rather than a stop-placement decision: the position has to be small enough that sitting through twice the entry width is tolerable. A desk that sizes to its stop rather than to its horizon has built a trade that loses when it is right.

[1] `estimation::ConvergenceTrade`.

Structure (The same first passage problem as the doubling strategy)**.** This is chapter 4's doubling strategy in respectable clothing.

There, a strategy on a martingale was certain to reach its target and the drawdown before it had a tail so heavy that its mean was infinite — which is why admissibility has to bound the loss rather than its expectation. Here the process is mean reverting rather than a martingale, so the tail is far lighter and the mean excursion is finite. But the structure is the same: a bet that is certain to win, whose risk lives entirely in what happens before it does, and whose sizing is therefore governed by a first passage distribution rather than by an expected return.

The difference is what makes relative value a business and doubling a fallacy. Mean reversion makes the excursion's distribution thin enough to survive with finite capital, and (22.1)'s κ is exactly what controls how thin. Which puts an uncomfortable amount of weight on knowing κ.

Remark (And κ is the parameter that is estimated badly)**.** Chapter 20 measures that mean reversion estimated from a finite sample comes out too fast, and chapter 14 explains why from the spectrum: a deviation decays more quickly than the slowest mode at first, so fitting one exponential to a whole sample returns a rate above the true gap. The bias is one-signed.

Follow that through to the trade. An overstated κ is an understated half-life, which is an understated horizon. With the true half-life at one year and the estimate thirty per cent fast — comfortably inside what a decade of data delivers — the planned holding period is 1.63 years against a realised 2.10.[2] The desk expects to be out in twenty months and is still in the position at twenty-five.

Every part of that error points the same way. The horizon is longer than planned, so the funding cost is larger than budgeted; the position is held through more of the excursion distribution, so the drawdown is deeper than modelled; and the capital is committed longer, so the return on it is lower than advertised. There is no compensating error in the other direction.

22.4 Whether the Spread Converges at All

The section above began by granting that the mean reversion is certain, and everything in it followed from that. What happens when it is not granted?

Chapter 20 has already ruled out the obvious approach. Fitting (22.1) to a history returns a positive $\hat{\kappa}$ whether or not there is any mean reversion to find — on a random walk the estimate is positive with probability approaching one — so "the regression found convergence" establishes nothing. What is needed is not a better estimate but a *test*, with the random walk as the null hypothesis rather than as an alternative nobody considered.

Chapter 20's Dickey-Fuller test is the natural tool here rather than an imported one, and its critical value of about -2.86 rather than -1.65 is exactly the correction that a non-stationary null requires.[3]

Calculation 22.3 (What the test can actually see)**.** The test is correctly sized — a genuine random walk passes for mean reverting about five per cent of the time, as asked. The question is the other error: how often does the test find mean reversion that is really there?

[2] `measured`.

[3] `estimation::unit_root_test`.

Half-life	1 year	3 years	5 years	10 years
6 months	6%	11%	17%	52%
1 year	6%	7%	8%	18%
2 years	6%	5%	6%	8%

Read the middle row. A spread that genuinely reverts with a one-year half-life, watched for five years, is identified as reverting eight times in a hundred.[4] The test is not broken; it is being asked to separate two hypotheses that a sample of that length barely distinguishes.

And the pattern across the table is the same shape chapter 20 found for the estimation bias. Power depends on the number of *half-lives* the sample spans and not on its length or its resolution: a one-year half-life over five years and a two-year half-life over ten give the same answer, though one history is twice the other.[5] Sampling the spread hourly would add observations and no power at all.

Structure (Which is why the structural condition does the work)**.** Put Calculation 22.3 beside the two conditions this chapter opened with and the two halves of its argument fit together.

A test that finds real convergence eight times in a hundred cannot be what a desk relies on. Worse, it is not merely uninformative but dangerous under search: given n instruments there are many combinations to weight, and looking through enough of them will produce one that passes at five per cent whether or not anything converges. A screen that tests a thousand spreads finds fifty by construction. The statistic that was weak as evidence becomes actively misleading as a filter.

So the structural reason demanded earlier is not a piece of good practice sitting alongside the statistics. It is what the statistics cannot supply. A butterfly is a curvature by construction, a cash-futures basis is tied to financing by an arbitrage that must close at delivery, an on-the-run spread is a liquidity premium with a named mechanism — and each of those is a reason to expect convergence that does not come from the sample, so it is not consumed by having searched the sample. The test is then a check on a prior rather than a way of forming one, which is the only role its power supports.

That is also the honest difference between this and the equity pairs trading the same mathematics is usually taught with. There the combination is typically *discovered* by search over a universe, and the discovery is exactly the procedure Calculation 22.3 says will manufacture false positives. Here the combination is usually written down first, for a reason, and the data is asked only whether it disagrees.

22.5 Carry, Roll-Down, and the Rest

Start with the simplest possible question. I buy a bond, fund it, and hold it for three months. Where does the money come from?

Write $y(T)$ for the yield of a T-maturity bond, D for its duration and h for the horizon. Over the horizon two things happen: the bond gets older, and the curve moves. Expanding the price

[4] `measured`.

[5] `measured`.

change,

$$\underbrace{\frac{\Delta P}{P}}_{\text{total return}} \approx \underbrace{y(T)\,h}_{\text{coupon}} \underbrace{-\,r\,h}_{\text{funding}} \underbrace{-\,D\big(y(T-h)-y(T)\big)}_{\text{roll-down}} \underbrace{-\,D\,\Delta y}_{\text{yield change}} + \tfrac{1}{2}C(\Delta y)^2. \tag{22.2}$$

The first two terms are the *carry*: what the bond pays less what the funding costs. The third is the *roll-down*: even if the curve does not move at all, a five year bond becomes a four-and-three-quarter year bond, and on an upward sloping curve that means its yield falls and its price rises. The fourth is the only term involving an actual change in the market, and the fifth is the convexity of chapter 7.

Calculation 22.4 (How the known part compares with the unknown)**.** Equation (22.2) separates what is known from what is not and says nothing about their sizes, which is the comparison that decides whether a carry trade is a harvest or a bet. Both are available: the known part from today's curve, and the risk from the realised volatility of the same maturity's yield over the same horizon.

Taking the Treasury curve and a quarter's holding period, with the one-month bill as funding: the known part is exact discount-factor arithmetic against today's curve, and the risk is realised yield volatility over four decades of history where a tenor has been quoted that long, and one year where it has not.[6]

Maturity	yield	carry + roll	risk over the quarter	ratio
1y	4.43%	96bp	37bp	2.57
2y	4.71%	28bp	89bp	0.32
5y	4.83%	−13bp	230bp	−0.06
10y	4.96%	−4bp	400bp	−0.01
30y	5.29%	−269bp	444bp	−0.61

Remark (One year is an exception, and the long end has inverted)**.** From two years out the pattern is the one the two conditions above lead you to expect: the known part is a small fraction of the risk, a directional bet with a small tilt in its favour rather than an income stream with noise attached.

One year does not fit it. A ratio of 2.57 is not a small tilt, and the reason is visible in the curve itself: the one month bill sits close to half a point below the one year bill, a genuinely steep front end that even four decades of yield volatility does not dominate. Whether that persists is a question about tomorrow's curve, not a property asserted of curves in general — the same calculation answers it again whenever it is asked.

And the long end has inverted rather than merely underperformed. The thirty year known return is comfortably negative, not the largest number on the curve: the twenty year point sits above both its neighbours, so a bond ageing from thirty towards twenty years ages into a *higher* yield and a lower price, roll-down running backwards. The worst risk-adjusted place to hold duration on this curve is not "the longest maturity" as a rule; it is wherever the curve's own local slope happens to be working against you, and on this curve that is squarely the long end.

The important structural feature of (22.2) is that the first three terms are *known today*. They are properties of today's curve and the passage of time, not forecasts. Only the fourth is uncertain.

[6] `risk::what_the_curve_currently_pays_to_hold_by_maturity`, against the panel in `public/marketdata`.

That is what makes carry-and-roll trades attractive and what makes them dangerous. A position with positive carry and roll makes money on every day the market does not move. It loses money when the market moves against it, and the losses are proportional to duration, which is to say much larger than the daily income. A carry trade is therefore short a large, rare loss and long a small, steady gain — which is the payoff diagram of a sold option, assembled without buying or selling one.

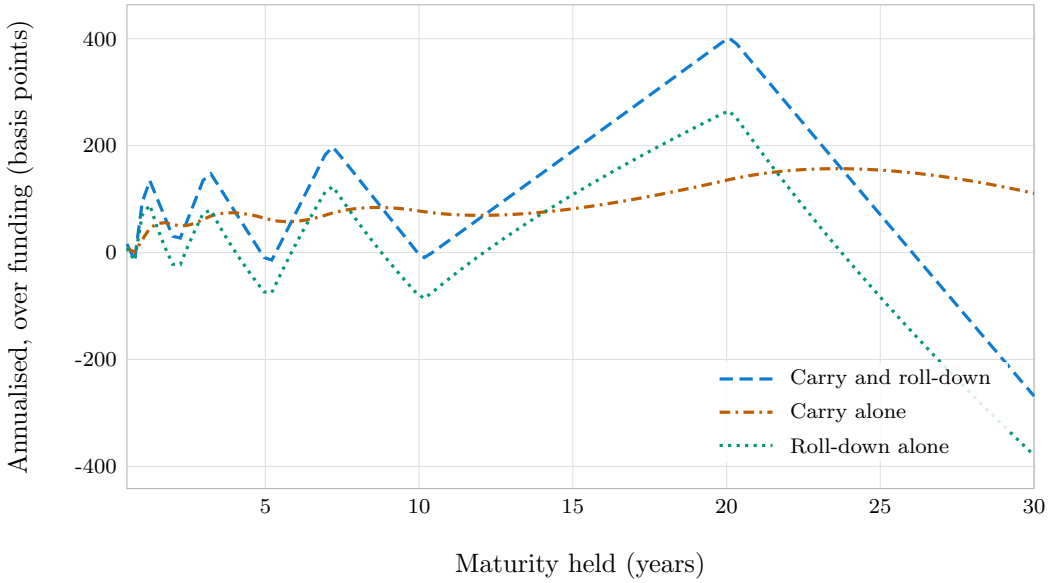

Figure 22.1: Carry and roll-down along the US Treasury curve, annualised, over the cost of funding for three months. Carry generally rises with maturity — a longer bond typically yields more over the funding rate — but it is not immune to the curve's own shape: it peaks in the twenties and eases back by thirty, tracking the same local inversion roll-down responds to more sharply. Roll-down depends on the *slope* of the curve where the bond sits, so it is largest where the curve is steepest and turns negative where the curve inverts, which on this curve is between twenty and thirty years. The total is the sum, and it is not monotone in maturity: the point of the curve that pays best to hold is not the longest one.

Drawn by `bootstrap_par` and `Curve::df`. US Department of the Treasury, daily par yield curve rates, as of 2026-09-22 (par yield, semiannual coupon, actual/actual). Retrieved from https://home.treasury.gov/interest-rates-data-csv-archive.

Remark. Note how much of this survives without any model at all. The decomposition is arithmetic and the figure is the curve of chapter 7 with a subtraction applied. No volatility, no measure, no calibration. That is characteristic of relative value work: most of the analysis is careful bookkeeping, and the modelling enters only when an option is involved.

22.6 Curve Trades

A directional position in one bond is a bet on the level of rates, which is a macroeconomic view rather than a relative value one. The relative value version isolates the shape.

Chapter 12 noted that a principal component analysis of curve changes finds three factors: a level factor moving all rates together, a slope factor steepening or flattening, and a curvature factor moving the middle against the ends. Between them they explain almost all of the variance.

This immediately suggests what to trade. Pick three points on the curve, and choose notionals w_1, w_2, w_3 so that the position has no exposure to the first two factors:

$$\sum_i w_i D_i \, \ell_i = 0, \qquad \sum_i w_i D_i \, s_i = 0,$$

where ℓ and s are the loadings of the level and slope factors on each point and D_i the durations. Two equations in three unknowns leave a one-dimensional family, and fixing the overall size picks one. The result is a *butterfly*: long the middle against the wings, or the reverse.

What is left is exposure to curvature and to nothing else — so if the middle of the curve is dear relative to the two ends, this position monetises exactly that, whatever the level and slope do.

That is the claim, and on a real curve it is only approximately true.

Calculation 22.5 (How neutral the neutral position is)**.** Take the equal-weighted $1 : -2 : 1$ fly on the two, five and ten year points — the version that is used because it needs no estimate at all, being exact for any move that is linear in maturity — and measure how much of its daily variance the first two principal components still explain, on the committed Treasury history.[7]

	explained by level and slope
Outright five year	98.1%
$1 : -2 : 1$ butterfly	26.4%

So the weighting does most of what it claims and not all of it. An outright position is the level factor and essentially nothing else; the fly removes three quarters of the factor exposure, and a quarter of what it does is still level and slope. Its own volatility is around two basis points a day, so the residual is not a rounding error — it is a directional position of real size sitting inside a trade that is described as non-directional.

The gap is the difference between the weighting that is exactly right for a straight-line move and the weighting that is right for the moves this curve actually makes. Fitting the weights to the estimated loadings closes it, at the cost of the next remark.

Remark (The weights are a modelling choice)**.** The loadings come from a covariance matrix estimated over some window, and the estimate depends on the window. A butterfly hedged with loadings from a calm period is not hedged in a volatile one, because the factors themselves rotate.

This is chapter 7's lesson about interpolation, and chapter 24's about model risk, arriving in a third place: the hedge is a property of an estimate, and the estimate is a choice. A desk running these trades is exposed to the estimation window in a way that no risk report shows, because the risk report uses the same window.

Calculation 22.6 (Removing the third factor too)**.** A fourth point turns the fly's two constraints into three: choose weights on four points so the position is null to level, slope *and* curvature at

[7] `risk::factor_neutrality`, on the same daily curves as the factor decomposition above.

once, leaving one degree of freedom for size. Add the thirty year point to the existing two, five and ten year fly and solve for the weights that null all three fitted loadings at once — a *condor*.[8]

	2y	**5y**	**10y**	**30y**
Weight	1.00	−0.03	−0.63	0.27

Where the naive fly left a quarter of its variance in level and slope, this position's exposure to all three factors together is indistinguishable from zero: nothing measurable is left of level, slope or curvature, only whatever the curve's own three-factor decomposition does not span.

Remark (What is left cannot be told from noise)**.** Chapter 6 called a position like this one a statistical arbitrage: a factor-neutral portfolio with a genuine expected return, compensation for whatever the factors do not span rather than for bearing level, slope or curvature risk. Whether *this* residual has one is a different question, and chapter 6's own point that volatility can be estimated from one realisation and drift cannot answers it before the position is even priced. Its mean daily move over the committed year is about a tenth of a basis point against a volatility of eight basis points a day — a t-statistic of about 0.2, indistinguishable from the zero a position with no edge at all would produce.[9] A year of daily data was never going to settle this. The condor's neutrality is exact and measured; its edge, if it has one, is not — which is why a structural reason, not this regression, is what would actually make a residual like this one collectible.

Calculation 22.7 (Not choosing a window)**.** There is no window that is right, and that can be shown rather than argued. Take a hedge ratio that genuinely wanders and estimate it two ways, over a month and over a year. When the ratio moves slowly the long window wins comfortably; when it moves five times faster the short window wins instead.[10] The better choice depends on a rate of rotation that is not observed, so choosing a window is guessing at it while pretending to be doing something else.

The alternative is to stop choosing. Write the ratio as a state that drifts and the regression as a noisy view of it,

$$\beta_t = \beta_{t-1} + \eta_t, \qquad y_t = \beta_t\, x_t + \varepsilon_t,$$

and this is chapter 21's filter with a different state in it — no new machinery, and the same recursion. Its gain settles at a level fixed by the ratio of the two variances, so the estimate is exponentially weighted with a memory implied by how fast the ratio is thought to move rather than declared in advance. Told the truth about that rate, it beats both windows at every rate.[11]

Remark (What the filter has actually changed)**.** Not as much as the last paragraph suggests.

The filter needs the rate of rotation, which is the window question in different clothes. Give it a rate within about a factor of two of the truth and its advantage survives; tell it the ratio moves five times more slowly than it does and it is *worse* than having simply taken the long window.[12]

[8] `risk::condor_neutrality`, the null vector of the same three-equation system the butterfly's own weights above solve with one fewer point, here on fitted loadings rather than asserted as $1 : -2 : 1$.

[9] `risk::the_condor_is_genuinely_neutral_and_its_own_return_is_unmeasurable`.

[10] `measured`.

[11] `measured`.

[12] `measured`. Being told too fast is the more forgiving of the two errors, which is the usual asymmetry: a filter that distrusts its own history recovers, one that will not update does not.

So the gain is not that a parameter has been removed. It is that the parameter has been replaced by a better posed one. "How many days of history should the weights use" is a question about the estimator with nothing outside the data to answer it; "how fast do the factors rotate" is a question about the curve, and chapter 12's factor analysis is the sort of thing that could answer it. That is a real improvement and a modest one, and it is the honest description of what filtering buys here.

22.7 The Cash-Futures Basis

Now the trade that made the news, and the one where the mathematics is richest.

A Treasury futures contract does not reference a single bond. The short may deliver any bond from a defined basket, and to make bonds of different coupons and maturities comparable the exchange applies a *conversion factor* CF_i to each. Delivering bond i against a futures priced at F earns the short the invoice amount

$$\text{Invoice}_i = \mathrm{CF}_i \times F + \text{accrued interest}_i.$$

Definition 22.8 (The basis)**.** For a deliverable bond with clean price P_i,

$$\text{Basis}_i = P_i - \mathrm{CF}_i \times F.$$

The short delivers whichever bond costs least to buy and deliver, so the futures price is set by the *cheapest to deliver* — the bond minimising the basis. That is an option held by the short, and like every option in these notes it has value, is worth more when volatility is higher, and changes hands as the market moves: a large enough rally changes which bond is cheapest, so the futures contract's effective underlying switches.

The standard way to see whether the basis is rich or cheap is to ask what financing rate makes the trade break even.

Definition 22.9 (Implied repo rate)**.** The implied repo rate is the return earned by buying the bond today, funding it to delivery, and delivering it into the futures:

$$\text{IRR} = \frac{\text{Invoice at delivery} - \text{Purchase price today}}{\text{Purchase price today}} \times \frac{1}{\tau},$$

with τ the time to delivery.

Compare it with the actual repo rate at which the bond can be financed. If the implied repo rate exceeds the actual, the package — buy the bond, sell the future, fund in repo — earns more than it costs, and that difference is the basis trade.

Calculation 22.10 (What the basis trade actually is)**.** Set it out as a portfolio, which is the only way to see the risk.

- Buy N face of the cheapest to deliver bond, at price P.

- Finance it in the repo market: post the bond as collateral, receive $NP(1-h)$ in cash, where h is the haircut. Put up NPh of your own money.

- Sell $N \times$ CF of futures — the amount that delivers exactly the bonds held, since the invoice pays CF per unit of futures price. Selling futures requires initial margin rather than the notional, so this leg costs little cash.

The futures position is pinned by the bonds held and by nothing else: it is set so that the delivery obligation is exactly what is owned.

With the legs matched, the position is close to hedged: a rise in yields loses money on the bond and makes it on the future. What remains is the basis itself, which converges to zero at delivery — so the profit is the difference between the implied and actual repo rates, earned with near certainty if the position is held to delivery.

So the leverage does not come from the futures at all. It comes entirely from h, the haircut on the bond leg, which is what lets a large holding of bonds be carried against a small amount of capital; the futures margin is a further call on capital rather than a multiplier of the position. With a haircut of 1%, \$100 of bonds requires \$1 of equity to hold, so a basis of two basis points annualised becomes a return on that equity of two hundred. *The trade is not attractive because the edge is large. It is attractive because the edge is small and the leverage is enormous.*

Why does the gap exist at all? Because the two sides of it are wanted by different people. Asset managers who want duration without balance sheet buy futures, pushing them rich. Someone must take the other side, hold the cash bond, and finance it, and that is a use of balance sheet that banks have been progressively less willing to supply since capital rules began charging for it. Hedge funds stepped into that gap. The basis is, in effect, the price of balance sheet.

22.8 Why It Breaks

The trade is hedged against the thing it looks exposed to, and exposed to two things it does not look exposed to. Both are about funding rather than about rates.

The margin on the futures leg is settled daily; the gain on the cash leg is not. Chapter 8 derived this asymmetry to compute the future-forward basis, where it was worth a basis point or two. Here it is a cash flow: if yields fall, the short futures position loses and must post variation margin *today*, while the matching gain on the bond is unrealised. A hedged position with no economic exposure can therefore demand cash without limit in the interim.

The haircut is not constant. It is set by the lender and rises when markets are volatile. A haircut moving from 1% to 3% triples the capital the position requires, and a fund that does not have it must sell. Selling the cash leg widens the basis, which marks every other holder of the trade to a loss and raises their margin calls too.

That is the whole mechanism. At 1% haircut the leverage is a hundred to one. A move in the basis of ten basis points — unremarkable in a stressed week — is ten percent of capital. A move of a hundred is the fund.

Remark (This is chapter 4's admissibility condition, in the wild)**.** There is a genuine and not merely rhetorical connection to the theory.

Chapter 4 met a strategy that was an arbitrage under the definition then in use: the doubling strategy, which reaches a certain profit almost surely by being willing to sustain unbounded losses along the way. The repair was to require *admissibility* — that the wealth process stay above a fixed level — and we observed that without it no market is arbitrage-free, because the mathematics permits a trade that no counterparty would fund.

The basis trade is that condition being tested by reality. Held to delivery with no financing constraint it converges, and the profit is nearly certain. Held with a hundred times leverage and daily margin, the interim losses are exactly what admissibility rules out, and the strategy fails not because the arbitrage was not there but because surviving to collect it was not financed.

The lesson chapter 4 drew from a piece of measure theory is the same one: a profit you cannot fund your way to is not a profit.

Exercise (Sizing the spiral)**.** A fund holds \$100 of the basis trade at a 2% haircut, so \$2 of capital. The haircut rises to 4%. Show that, holding the position, the fund needs \$2 more capital; and that if instead it deleverages to its existing capital it must sell half the position. Then argue that if enough funds do this at once the basis widens, and show that a widening of 20 basis points on the remaining position wipes out a further tenth of the capital. This is the loop, and its speed is set by the haircut, not by rates.

22.9 Swap Spreads, and a Negative Number That Should Not Exist

One more spread, because it makes a point about what these trades measure.

The swap spread is the swap rate of chapter 7 less the yield of the matched-maturity government bond. Naively it should be positive: a swap has bank credit behind it and a government bond does not, so the swap should pay more.

At long maturities it has been persistently *negative*, which appears to say that lending to a bank is safer than lending to the government. It does not. The swap is collateralised, and chapter 7 derived what that means: a fully collateralised trade carries essentially no credit exposure, and is discounted at the collateral rate. The bond is not collateralised — it must be bought, funded and held on a balance sheet that is charged for it.

So the swap spread is not a credit spread at all. It is the price of balance sheet again, the same quantity the basis measures, seen from another direction. Once the capital rules made holding bonds expensive, the spread that compensated for that could and did go negative, and stayed there.

Remark. The general lesson: when a spread that theory says should have one sign persistently has the other, the usual explanation is not that the market is wrong. It is that the theory omitted a constraint that binds — here, that balance sheet is finite and charged for. Chapter 7's collateral derivation and chapter 24's capital reserves are two faces of that constraint, and this spread is its price.

22.10 Selling Volatility, and What That Is Short

Everything so far has been a trade in a rate or a spread between rates. There is a second family, at least as large, in which the instrument is an option and the quantity being traded is volatility itself.

The position is a swaption or a cap sold and then delta hedged, so that the direction of rates is removed and what is left is the difference between the volatility charged and the volatility that arrives. The structural reason is on the other side and is the one chapter 23 describes: the flow is one-directional. Borrowers want caps and issuers want the option to call, so the dealer community is structurally short optionality and is paid to be. Chapter 10 locates the same premium inside a model, where it sits in the drift parameters κ and θ and not in the vol-of-vol η, which is measure-invariant.

Calculation 22.11 (What a hedged short option actually earns)**.** Sell a call at an implied volatility σ_i, hedge it with the delta computed at σ_i, and let the world realise σ_r. Over one rebalancing interval the position is delta neutral, so by chapter 5's expansion the profit and loss over that interval is the theta collected less the gamma paid,

$$-\Theta\,dt - \tfrac{1}{2}\Gamma\,(dS)^2,$$

and the theta-gamma identity of that chapter, $\Theta = -\frac{1}{2}\sigma_i^2 S^2\Gamma$ at zero rates, substitutes for the first term. Since $(dS)^2 = \sigma_r^2 S^2 dt$ along the realised path, the two collapse into

$$\text{P\&L} \;=\; \int_0^T \tfrac{1}{2}\,\Gamma_t\,S_t^2\left(\sigma_i^2 - \sigma_r^2\right)dt. \tag{22.3}$$

Simulating the hedge and accumulating (22.3) along the same paths are different computations — the first forms the payoff and uses the delta, the second never forms a payoff and uses the gamma — and they agree.[13]

Read (22.3) carefully, because it does not say what the trade is usually described as saying. It is not a bet on realised variance. It is a bet on realised variance *weighted by gamma*, and gamma is large only near the strike and near expiry. A move of a given size contributes according to where the underlying was when it happened, so two paths delivering identical realised variance pay different amounts.

How different is worth measuring. Selling a one-year at-the-money option at 22% into a world realising 18% earns 1.59 on average, which is exactly the difference between the two Black-Scholes premiums and is the sense in which the edge is real.[14] But restricting to the paths that realised 18% to four decimal places, the outcomes still run from 0.49 to 2.94.[15] The dispersion remaining after the forecast has been made *exactly right* is wider than the entire average edge.

Structure (Short the path, not the variance)**.** A desk selling volatility describes itself as short variance and forecasts variance accordingly, but (22.3) says the exposure is to a gamma weighted

[13] `measured`.
[14] `measured`.
[15] `measured`.

average, and the weighting is a fact about the path rather than about its total. Being right about volatility and wrong about where the moves land is a losing position.

So the trade is short the *shape* of the realised path, and the chapter's usual question has the usual answer: it is short a constraint, namely that the hedge is discrete and the gamma is concentrated. Chapter 5 has already drawn the same object: two markets with identical total variance and different jump content produced identical premiums and entirely different distributions of outcome. This is that figure's trading counterpart.

It also explains the instrument. A variance swap pays realised variance with no gamma weighting at all, which is exactly the exposure the hedged option fails to give, and that — not any view on volatility — is what it is for. Chapter 15 has in fact already built the object, though not under that name: the second moment of the swap rate comes out as an integral of swaption prices with *equal* weight at every strike, which is a variance swap on the rate in all but the contract. The uniform weighting is not an accident — expanding a payoff in options gives each strike the weight $g''(K)$, and the second moment has $g'' = 2$ everywhere.

The familiar $1/K^2$ weighting is the same construction asking for a different variance. Buying the variance of the *logarithm* means replicating $-2\ln(F_T/F_0)$, whose second derivative is $2/K^2$, and that is the equity convention because equity variance is quoted in proportional terms. A rates desk buying basis points squared weights uniformly. The two are not competing formulas for one instrument; they are the same formula for two.

Three further trades in this family:

The volatility term structure. Selling short-dated and buying long-dated volatility is a trade on how implied volatility is interpolated between expiries, and what it is short is the event calendar. Chapter 9 shows why: a smooth interpolation through a payroll or a meeting date reads the lump as a term structure and marks every expiry between the quotes wrong. The trade collects the marking error and is short the possibility that the event is bigger than the lump priced for it.

Skew. Selling a receiver and buying a payer against it is a position in the smile's slope, and chapter 10 says the slope is a statement about the correlation between the rate and its volatility. What that is short is not a rate move but a change in that correlation — and the mortgage hedging flow of the next section is precisely a mechanism that changes it, which is why the two trades are less independent than a risk report shows.

Conditional and curve volatility. A swaption on a spread is worth an amount that depends on the correlation between the two rates, and chapter 13 measures how much: marking correlation moves a price by percentage points where the choice between two well-specified models moves it by hundredths. A trade in spread volatility against outright volatility is therefore short a correlation marking, which is an estimate from chapter 21 rather than a quote, and the least observable input in the book.

22.11 Six More, and What Each One Is Short

The two trades worked through above are not a representative sample, and a reader who saw only them would take away that relative value is about bonds against futures. It is not. What follows

is a catalogue rather than a derivation — each entry is a trade that has been put on at scale, with the structural reason that made it collectible and the exposure that made it dangerous.

On-the-run against off-the-run. The most recently auctioned Treasury of a given maturity trades rich to an almost identical bond issued a few months earlier. The reason is not credit and not cash flows, which are nearly the same; it is that the new bond is the benchmark, is what hedges are executed in, and finances more cheaply in repo because everyone wants to borrow it. The spread is therefore the price of liquidity and of financing, and it can be collected by buying the old bond and selling the new one. What the position is short is liquidity itself — and liquidity premia widen precisely when a position needs to be unwound. The trade was a signature of Long Term Capital Management, and in 1998 the spread went from rich to richer while the leverage that made it worth doing forced the unwind. Chapter 4's admissibility condition is the abstract version: a trade financed by an unbounded drawdown is not a trade, and the drawdown here was a mark-to-market on a position whose thesis was correct.

The squeeze. A short position in a specific deliverable bond is not short a price. It is short the ability to borrow that bond, and if someone corners the supply, the cost of maintaining the short is set by them. Salomon Brothers demonstrated this at the two year Treasury auction of May 1991 by acquiring a dominant share of the issue, after which the repo rate on the note collapsed and the shorts paid. The general point is structural and worth carrying: for a deliverable instrument the funding leg has a different risk profile from the price leg.

Inflation-linked against nominal. A Treasury inflation-protected bond plus an inflation swap that converts its cash flows to fixed replicates a nominal Treasury exactly — a static replication of the kind chapter 15 builds, with no model in it at all. The replication has been persistently cheaper than the nominal bond it replicates, at times by a wide margin, and the mispricing survived for years rather than minutes. It is as close to a textbook arbitrage as the market offers, and it persisted for the reason everything else in this section persists: capturing it requires holding a levered, balance-sheet-intensive position to maturity, and the institutions with the balance sheet had better uses for it.

Covered interest parity. Borrowing dollars directly and borrowing yen and swapping them into dollars should cost the same, since the forward rate is a ratio of numeraires and chapter 17 derives it with no freedom in it. Since 2008 they have not cost the same, the gap widens reliably at quarter ends when balance sheets are reported, and it is a cross-currency basis rather than a mistake. The trade is to supply the scarce thing — dollars, on a balance sheet — and what it is short is the regulatory constraint that made them scarce, which is to say a rule that can change.

Mortgage convexity hedging. A homeowner holds an option to prepay — refinancing the loan when rates have fallen enough that a new, cheaper one is worth the cost of arranging — so a mortgage portfolio is short that option. Its duration shortens when rates fall, because more of the pool is then expected to prepay and return principal early, and lengthens when rates rise, because refinancing slows and the expected life extends: the opposite of an ordinary bond, whose duration moves the other way. Whoever holds that exposure must hedge it, and the hedge is mechanical: rates fall, duration has shortened below target, so the hedger receives fixed to buy duration back; rates rise, duration has lengthened, so the hedger pays fixed to sell it off. Buying duration into a rally and selling it into a sell-off is trading with the move rather than against it, so this flow — proportional to the change in the aggregate duration of a very large portfolio, and forced rather than discretionary — amplifies whatever rate move triggered it rather than damping it. And because the

amplification is large, mechanical and recurring, it supports a standing bid for the instrument that lets a hedger own the same exposure outright instead of chasing it with rebalancing — a swaption — which is why this flow supports swaption implied volatility. Two trades follow: buying that optionality ahead of a move expected to trigger the flow, a direct bet that realised volatility will run ahead of what is priced; or trying to anticipate the flow itself, the rate level at which a wave of refinancing starts and the rebalancing it forces, rather than volatility in general. Both are short the same thing: the stability of prepayment behaviour, which is a model output and one of the least reliable ones in the industry.

The auction cycle. Dealers must absorb each new issue, and their capacity to warehouse it is finite. Yields have tended to rise into an auction and fall after it, a pattern with a named mechanism rather than a curve-fit, and it is the closest thing in this chapter to a pure statistical arbitrage with a structural reason attached. What it is short is dealer capacity: the pattern is strongest when balance sheet is scarce, which is also when the position cannot be financed.

Structure (The same short, six times)**.** Set the six side by side and one thing is common to all of them. None is short a price. Each is short a *constraint* — financing, borrowing, balance sheet, regulation, a behavioural model, warehousing capacity — and each pays a premium that exists because the constraint binds on somebody.

That is not a coincidence and it is the reason these trades exist at all. A genuine mispricing of a price, in a liquid market with many participants, is arbitraged away in the time it takes to notice. What is not arbitraged away is a premium paid for absorbing a constraint, because absorbing it requires the very thing that is scarce. So the surviving trades are exactly the ones where the return is compensation for a service rather than a reward for being right.

Why is a mispriced *price* removed so fast?

Not because it is hard to see. A cash-futures basis, a triangular inconsistency in currencies, a bond trading through its own strip — each is arithmetic on quoted numbers, computed identically by everyone with the same feed. Detection is not the scarce input and confers no advantage, which is why no desk earns anything from noticing.

It is removed fast because it is a *race for a finite quantity.* A mispricing is not a pool of free money but a bounded amount available at a stale price, and it goes to whoever reaches it first. That makes the return to speed winner-take-most rather than proportional, and a payoff of that shape is competed to whatever the physical limit happens to be — which is why the participants who do this have spent their capital on the distance between two machines rather than on models.

And that is exactly what separates the two classes. A mispricing removable by a fast trade needs no balance sheet: the position is opened and closed, nothing is carried, and the only scarce input is time, which is bought once and then owned. A premium for absorbing a constraint cannot be competed away by speed at all, because the scarce input is the ability to *hold* — collateral, capital, warehousing, the willingness to sit through Calculation 22.2's excursion — and no amount of speed supplies it. Speed removes the arbitrages that can be closed in a moment and is powerless against the ones that must be carried for months, which is why the surviving trades all look alike.

The fast trader's profit is the slow quoter's adverse selection as seen in chapter 23: a quote that can be picked off before it is pulled is exactly the toxic flow that chapter measures and prices into the spread.

It follows that the risk of the whole class is one risk. All six lose money in the same circumstance — when the constraint tightens rather than relaxes, which is when financing is withdrawn, balance sheets contract and everyone holding the premium needs to stop. That is why relative value books that appear diversified across trades are not, and why chapter 24's stress work has to be run on the constraint rather than on the instrument.

22.12 What Retail Actually Trades

Everything above is desk-scale: bonds financed in size, spreads held to maturity, balance sheet an individual account does not have. A retail account trades a narrower set of instruments, built mostly around two things — a rate ETF with a liquid, exchange-listed options market, and a callable or structured note — and the question this chapter has asked of every trade above applies to these too: what is the position short, and is there a reason to expect it is paid for.

Selling optionality on a rate ETF

A long-dated Treasury ETF such as TLT has a liquid, listed options market, which an individual bond or a swap does not — and because it trades in the same equity-market structure, the size check chapter 21 runs against an equity index, implied over realised by one to three volatility points, applies to it directly rather than needing translation the way a swaption would. Selling a covered call against a holding of it — or, equivalently, a cash-secured put to acquire it below today's price — is the retail-accessible version of "Selling Volatility, and What That Is Short" above: short gamma, financed by the premium, and earning it for the same underlying reason a dealer's short-swaption book does — variance is a priced risk, and a hedge against it costs a premium to hold. The edge, where there is one, is the same edge; the position is smaller and the counterparty is a market maker rather than a client hedging a liability, but the source of the premium has not changed.

What has changed is everything working against collecting it. A retail account pays the ETF's own expense ratio and the option's own bid-offer, and cannot warehouse the position the way a dealer's balance sheet can if the premium on offer is not enough for the risk taken — the same one to three points, collected in smaller size against larger frictional costs.

Buying optionality: leverage, not edge

Buying a call or a put on the same ETF, or on a Treasury future for more leverage and less liquidity, is the mirror position, and it inherits the mirror problem. The premium a covered call collects is paid by whoever is on the other side of it: a bought option is, on average, betting against the same variance risk premium the seller of one collects. That does not make the trade wrong. The premium is compensation for a real, one-signed risk, and a trader with a genuine view about direction can be right about it while still paying a structural headwind on the instrument used to express that view. It means the honest accounting for a leveraged directional options position is the view's own expected payoff, net of a premium priced against the buyer before the position is even opened — a structural reason is what would make a position collectible, and leverage on its own is not one.

Trading the event calendar

The event-lump mispricing "The volatility term structure" trades institutionally, as a calendar spread across two expiries, is executable directly on a single name or a single ETF: buy a straddle or a strangle expiring just after a scheduled event — an FOMC meeting, a payrolls print, an earnings date — and the position is a bet that the actual move exceeds what the option's own implied volatility already prices for it. Chapter 9's point about smoothing through the event applies here exactly as it does institutionally: a market maker who has not modelled the event separately from the surrounding calendar marks every expiry near it off an interpolated volatility that is neither the event's own size nor the quiet days around it, and the mispricing — too large or too small — is what a straddle bought or sold around the date is short. Selling the same structure is the mirror position, financed by the same over-pricing an event lump can carry when the interpolation is too generous, and it is short exactly what the trade is built to collect from: a move bigger than the market left room for, on the one day size is least forgivable.

Why the retail version is usually defined-risk

A dealer selling volatility warehouses the tail because chapter 23 shows a partial hedge is the arithmetic optimum once hedging costs money, and a dealer's balance sheet is deep enough to survive being wrong before it is right again. A retail account, or a small fund, does not have that — §22.2 already made the point that the capital available to hold a position is not guaranteed to survive as long as being right takes, and for an individual account "not guaranteed" is closer to "essentially never." So the retail version of every short-volatility position above is usually built with the loss capped from the outset — an iron condor rather than a naked short strangle, a credit spread rather than a naked short option — paying away part of the premium collected to buy back the tail a dealer would simply hold. That is not a different trade; it is the same premium, minus the price of the one thing a small account cannot self-insure.

Structured notes, and who actually prices the option

The third common vehicle is not a strategy an investor builds but one sold pre-built: an autocallable, a reverse convertible, a note marketed as "enhanced income," each of which is a bond plus a short option position packaged as a single security with a headline coupon. The economics are the covered call above, with one difference that matters: the issuer prices the embedded option once, keeps a spread for structuring and distribution, and the buyer never sees a premium to compare against a quoted market price the way the ETF option above has one sitting next to it. A structural reason is present here too, just on the other side of the trade from where it usually sits in this chapter — it is the issuer, not the buyer, who is the natural seller of a service, structuring and distribution and balance sheet, and is paid for it. That the product is marketed as an investment rather than sold as an option is exactly what makes the issuer's side of the premium collectible without ever being quoted against anything the buyer could check.

22.13 Where an Edge Lives: Information as a Filtration

Everything so far in this book has fixed a filtration $\mathcal{F}_t$ at the outset and asked what prices must be, given it. That construction has been so uniform that it is easy to miss what it assumes: that everyone sees the same thing. An edge is what happens when they do not, and the machinery to say so precisely is already in place.

Let $\mathcal{G}_t$ be the market's filtration, the one under which the discounted price is a martingale by chapter 4, and let $\mathcal{F}_t \supseteq \mathcal{G}_t$ be a trader's, containing something extra. A martingale in the smaller filtration need not be a martingale in the larger one. Under a technical condition it remains a semimartingale, and acquires a drift:

$$\underbrace{dS_t = \sigma_t\, dW_t}_{\text{in } \mathcal{G}} \qquad \longrightarrow \qquad \underbrace{dS_t = \sigma_t \alpha_t\, dt + \sigma_t\, d\tilde{W}_t}_{\text{in } \mathcal{F}}. \tag{22.4}$$

That drift α is called the *information drift*. The alpha a trader is looking for is not a metaphor borrowed from regression. It is a drift term, it exists only relative to a filtration, and its magnitude is the amount by which one information set disagrees with another about where the price is going. The literature that works this out is the literature on enlargement of filtrations, developed to study insider trading; the mathematics is indifferent to how the extra information was obtained.

Example 22.1 (A signal, and what it is worth)**.** Take the tractable case, where the extra information is a single number seen at the outset. Let W drive the price over $[0,1]$ and let the trader observe

$$L = \rho W_1 + \sqrt{1-\rho^2}\, Z,$$

with Z independent noise — a view on the terminal value, correlated ρ with it.

Fix t and work conditionally on $\mathcal{F}_t$, which is to say on W_t. Split the signal at t,

$$L \;=\; \underbrace{\rho W_t}_{\text{known}} \;+\; \underbrace{\rho\,(W_1 - W_t) + \sqrt{1-\rho^2}\, Z}_{\text{not}},$$

so the unknown part has variance $\rho^2(1-t)+(1-\rho^2) = 1-\rho^2 t$. Write $V = L - \rho W_t$ for it: the piece of the signal the market's own filtration cannot yet see, though the trader — who has observed L itself — already knows its value exactly. Its variance is what the market would still assign it, and that shrinks as t runs on: the price path itself increasingly confirms what L already implied.

Now take the next increment ΔW over $[t, t+h]$. Given $\mathcal{F}_t$, the pair $(\Delta W, V)$ is jointly Gaussian with

$$\operatorname{Var}(\Delta W) = h, \qquad \operatorname{Var}(V) = 1-\rho^2 t, \qquad \operatorname{Cov}(\Delta W, V) = \rho h,$$

the covariance because L loads on every increment of W with weight ρ. Gaussian conditioning is then one line,

$$\mathbb{E}\big[\Delta W \,\big|\, \mathcal{F}_t, L\big] \;=\; \frac{\operatorname{Cov}(\Delta W, V)}{\operatorname{Var}(V)}\, V \;=\; \frac{\rho\, h}{1-\rho^2 t}\,\big(L - \rho W_t\big),$$

and dividing by h gives the drift the enlarged filtration sees,

$$\alpha_t \;=\; \frac{\rho\,(L - \rho W_t)}{1-\rho^2 t}. \tag{22.5}$$

So the information drift is a *regression coefficient* — the slope from projecting the next increment on the part of the signal not yet realised. That is the same object chapter 4 found when it wrote the hedge ratio as $d\langle V,S\rangle/d\langle S\rangle$ and read it as a regression slope. Alpha and a hedge ratio are the same kind of thing.

The second moment follows without further work. Unconditionally $V \sim N(0, 1-\rho^2 t)$, so squaring (22.5) and taking expectations,

$$\mathbb{E}[\alpha_t^2] \;=\; \frac{\rho^2\,(1-\rho^2 t)}{(1-\rho^2 t)^2} \;=\; \frac{\rho^2}{1-\rho^2 t},$$

which rises as $t \to 1$: the signal is worth most at the end, when little of it is left unconfirmed and the remaining uncertainty has collapsed onto it. What that drift is worth needs a criterion for sizing a position, and the criterion is not the obvious one.

Remark (Why maximise $\mathbb{E}[\ln X_T]$ and not $\mathbb{E}[X_T]$)**.** Put a fraction π of wealth in an asset with $dS/S = \mu\,dt + \sigma\,dW$ and the rest in cash at zero, so $dX/X = \pi\mu\,dt + \pi\sigma\,dW$ and $X_T = X_0 \exp\big((\pi\mu - \tfrac{1}{2}\pi^2\sigma^2)T + \pi\sigma W_T\big)$. Taking the expectation directly,

$$\mathbb{E}[X_T] = X_0 \exp\big((\pi\mu - \tfrac{1}{2}\pi^2\sigma^2)T\big)\,\mathbb{E}\big[e^{\pi\sigma W_T}\big] = X_0 \exp\big((\pi\mu - \tfrac{1}{2}\pi^2\sigma^2)T\big)\,e^{\frac{1}{2}\pi^2\sigma^2 T} = X_0\,e^{\pi\mu T},$$

the Itô correction and the lognormal mean adjustment cancelling exactly. That is linear and unbounded in π: maximising $\mathbb{E}[X_T]$ says take infinite leverage, an answer produced entirely by a vanishing set of paths with an astronomical outcome, while the leverage needed to reach them puts almost every other path into ruin. $\ln X_T$ is what is additive across time — a sum of log-returns — so its average over many periods converges to its expectation, and maximising $\mathbb{E}[\ln X_T]$ maximises the almost-sure long-run growth rate of compounded wealth rather than a mean dominated by outcomes that are never actually reached.

The point sharpens because $\ln X_T$ is Gaussian, and a Gaussian is symmetric about its mean: $\Pr\left[[]\ln X_T < \mathbb{E}[\ln X_T]\right] = \tfrac{1}{2}$ exactly, so

$$\operatorname{median}(X_T) \;=\; e^{\mathbb{E}[\ln X_T]}.$$

Maximising $\mathbb{E}[\ln X_T]$ is therefore exactly maximising the wealth level half of all paths beat — a typical outcome, not a hypothetical average one. The gap between the two objectives is now a number rather than a slogan: for a lognormal, mean equals median times $e^{\operatorname{Var}(\ln X_T)/2}$, so

$$\frac{\mathbb{E}[X_T]}{\operatorname{median}(X_T)} \;=\; e^{\frac{1}{2}\pi^2\sigma^2 T},$$

growing without bound in π. That ratio measures precisely how far $\mathbb{E}[X_T]$ has drifted from what almost every path actually delivers.

By Itô,

$$d\ln X_t = \Big(\pi\mu - \tfrac{1}{2}\pi^2\sigma^2\Big)dt + \pi\sigma\,dW_t,$$

so the long-run growth rate is the bracket, and maximising it over π gives $\pi^* = \mu/\sigma^2$ and

$$g^* \;=\; \tfrac{1}{2}\Big(\frac{\mu}{\sigma}\Big)^2. \tag{22.6}$$

The achievable growth rate is half the squared Sharpe ratio, and nothing about the horizon or the wealth enters.

Structure (A hard boundary on leverage, invisible to $\mathbb{E}[X_T]$)**.** The growth rate $g(\pi) = \pi\mu - \frac{1}{2}\pi^2\sigma^2$ is a downward parabola, zero at $\pi = 0$ and again at $\pi = 2\mu/\sigma^2 = 2\pi^*$, positive strictly between them and negative outside. So wealth grows without bound almost surely for $0 < \pi < 2\pi^*$, and collapses to zero almost surely for $\pi > 2\pi^*$ — a hard ruin boundary at exactly twice the growth-optimal leverage, not a soft or probabilistic one. $\mathbb{E}[X_T] = X_0 e^{\pi\mu T}$ never sees it: it keeps rising for every $\pi > 0$, describing paths that, past $2\pi^*$, almost surely do not occur. Maximising $\mathbb{E}[\ln X_T]$ instead sits at π^*, in the interior of the region that actually grows.

Remark (Why this objective, and not some other utility)**.** The choice of ln here is not a risk-aversion parameter fitted to taste. For repeated bets at a *constant* leverage, Kelly (1956) and Breiman (1961) proved something stronger: fix $\pi^* = \mu/\sigma^2$ and compare against any other constant leverage $\pi' \neq \pi^*$. Then

$$\frac{X_T(\pi^*)}{X_T(\pi')} \;\to\; \infty \qquad \text{almost surely as } T \to \infty.$$

Not on average — on almost every path, the log-optimal strategy eventually overtakes any fixed competitor, by a factor that grows without bound. Among constant-leverage strategies, $\mathbb{E}[\ln X_T]$ is therefore not one reasonable objective among several: it is the one whose maximiser almost surely dominates every other fixed one in the long run. The adapted, time-varying case — which is what π_t^* below actually is — needs a separate argument, given after it.

Remark (The optimum is myopic, which is why π_t can track α_t)**.** $\pi^* = \mu/\sigma^2$ was derived for constant μ, σ, and (22.5)'s α_t is not constant — it moves with t and W_t. The substitution below is licensed by a specific property of log utility. Since

$$\ln X_T = \ln X_0 + \int_0^T \left(\pi_t\mu_t - \tfrac{1}{2}\pi_t^2\sigma_t^2\right) dt + \int_0^T \pi_t\sigma_t\, dW_t,$$

and the second integral has expectation zero, maximising $\mathbb{E}[\ln X_T]$ over adapted π_t reduces to maximising the first integral's integrand pointwise at every t, since nothing couples the choice at one time to the choice at another. The instant-by-instant optimum, using whatever μ_t, σ_t are known at t, is therefore exactly globally optimal. That is what licenses $\pi_t^* = \mu_t/\sigma_t^2$ with a time-varying drift below, continuously rebalanced as α_t itself evolves — a property specific to log utility, not a general fact about portfolio choice.

Remark (The almost-sure dominance carries over to the adapted case)**.** Algoet and Cover (1988) extend Kelly and Breiman's result exactly this far, with no restriction on how μ_t, σ_t move: the myopic strategy π_t^* almost surely dominates every other adapted strategy π_t' in the same sense, $X_T(\pi^*)/X_T(\pi') \to \infty$ almost surely. The mechanism is the pointwise inequality already used above, $g(\pi_t^*) \geq g(\pi_t')$ at every t — so the drift of $\ln\left[X_T(\pi^*)/X_T(\pi')\right]$ accumulates without bound wherever π_t' differs from π_t^*, while what is left is a martingale whose fluctuation grows only like $\sqrt{T}$, too slowly to catch a drift growing like T. So tracking α_t with $\pi_t^* = (\theta + \alpha_t)/\sigma$ carries the same almost-sure dominance as the constant-leverage case, not merely the weaker property of maximising an expectation.

Now read (22.4) through (22.6). In the enlarged filtration the price acquires an extra drift of $\sigma\alpha_t$, so its Sharpe ratio rises from θ to $\theta + \alpha_t$ and the optimal position becomes $\pi_t^* = (\theta + \alpha_t)/\sigma$. The

informed investor's growth is $\frac{1}{2}(\theta+\alpha_t)^2$ against the uninformed investor's $\frac{1}{2}\theta^2$, and the difference is $\theta\alpha_t + \frac{1}{2}\alpha_t^2$. The signal is unbiased — $\mathbb{E}[\alpha_t] = 0$, since L and W_t both have mean zero — so the cross term vanishes in expectation, and the informed investor's advantage over the uninformed one's $\frac{1}{2}\theta^2$ is $\frac{1}{2}\mathbb{E}[\alpha_t^2]$ exactly, whatever θ itself happens to be.

The value of the signal is second order in it: it comes from the *variability* of the perceived drift and not from its level, which is why what emerges below is an information quantity rather than a return. Integrating over the horizon,

$$\frac{1}{2}\int_0^1 \mathbb{E}[\alpha_t^2]\,dt = \frac{1}{2}\int_0^1 \frac{\rho^2\,dt}{1-\rho^2 t} = -\tfrac{1}{2}\ln\big(1-\rho^2\big), \tag{22.7}$$

which is exactly the mutual information of the jointly Gaussian pair (W_1, L), checked here rather than taken on faith. Both are standard Gaussians with correlation ρ, so $I(W_1;L) = h(W_1) - h(W_1 \mid L)$. The unconditional entropy is $h(W_1) = \frac{1}{2}\ln(2\pi e)$; conditioning on L leaves W_1 Gaussian with variance $1-\rho^2$, the same conditioning fact used throughout this example, so $h(W_1 \mid L) = \frac{1}{2}\ln\big(2\pi e(1-\rho^2)\big)$. Subtracting,

$$I(W_1;L) \;=\; \tfrac{1}{2}\ln(2\pi e) - \tfrac{1}{2}\ln\big(2\pi e(1-\rho^2)\big) \;=\; -\tfrac{1}{2}\ln(1-\rho^2),$$

matching (22.7) exactly.

Kelly proved the general statement in 1956, for a gambler receiving tips over a noisy wire: the growth rate a log-optimal bettor can achieve from a side channel equals that channel's mutual information, whatever the channel is. The derivation above is the continuous-time instance — half a squared Sharpe ratio, integrated — and it arrives at an information quantity because that is what the theorem says it must. Which also fixes the units in which a signal should be valued. Not in basis points of edge, which depend on how the position is sized, but in nats, which do not.[16]

Example 22.2 (What the datasets actually are, in rates)**.** In fixed income the enlargement usually has a specific and mundane shape: the market is waiting for a scheduled government statistic, and something observable earlier is correlated with it.

Payrolls and unemployment are released monthly and are among the largest scheduled moves in the front end of the curve. Job postings, payroll-processor records and staffing volumes are observable continuously and are correlated with the release. Consumer spending arrives in retail sales and in the consumption component of output; card transaction panels and point-of-sale feeds see much of it first. Inflation is released monthly and scraped prices, freight rates and commodity fixings move before it. Wages appear in the employment cost series and in earnings, and are visible earlier in posted salaries. Activity shows up in satellite imagery of shipping, in electricity load, in traffic. Geopolitical and policy events are read from filings, transcripts and news well before they are quantified anywhere official.

Each of these is an enlargement $\mathcal{F}_t \supseteq \mathcal{G}_t$ of exactly the kind (22.4) describes, and (22.7) says what each is worth. That reframing does real work, because it replaces the question people usually ask — is this dataset novel, is it large, is it hard to obtain — with the only question that determines the value: how much does it move the conditional distribution of the release, *given* everything already

[16] `estimation::information_value` simulates the drift and confirms both (22.7) and the pointwise second moment, which is where an algebra error would hide.

public. A card panel covering a small and unrepresentative slice of spending may be genuinely novel and worth nothing. A widely available series may still carry information if nobody has bothered to condition on it correctly.

Two features specific to this setting are worth naming. The information has a known expiry: the release date is when $\mathcal{G}$ catches up completely and the drift goes to zero by construction, so the horizon in (22.7) is not a modelling choice but a published calendar. And the release itself is a measurement of the same underlying quantity with its own error, so a signal can be right about the economy and wrong about the print — the tradeable object is the statistic, not the world, and the conditional mutual information that matters is with respect to the number that gets published.

Structure (The value of a dataset is a mutual information)**.** Equation (22.7) is a special case of a theorem: the additional expected logarithmic utility available to a trader whose filtration is enlarged by a random variable equals the mutual information between that variable and the market.[17] The question "what is this dataset worth" therefore has an answer, in nats, and the answer does not depend on what the dataset is made of.

Three consequences.

Novelty is not the criterion; conditional information is. What (22.7) measures is the information in L *given* what is already known. A dataset that is a function of the observable price history has zero conditional mutual information and is worth exactly nothing, however unusual its provenance — and this is a testable condition rather than a matter of judgement. Most datasets sold as alternative fail it.

The value is quadratic in the correlation at the low end. Expanding (22.7), a signal with $\rho = 0.1$ is worth about a quarter of one with $\rho = 0.2$, not a half. Weak signals are worth very much less than their correlation suggests, and the last increment of correlation is worth the most, which is the mathematical form of the observation that being slightly better informed than the market is nearly useless.

Alpha decays by ceasing to exist, not by being crowded. The drift in (22.4) is defined relative to $\mathcal{G}$. As others acquire the same data, $\mathcal{G}$ grows to contain it, and α does not shrink — it becomes zero, because the information is no longer information. That is a sharper statement than the usual one about crowding, and it predicts the right shape: the decay is driven by who else has the data rather than by how much capital is deployed against it.

Remark (The chapter's other alpha, and what it shares with this one)**.** The projection above is not the only place this chapter's regression appears. Chapter 6 showed that in an arbitrage-free curve, a portfolio spanned by the level, slope and curvature factors earns exactly the factor premia and no more — its own market price of risk θ, read from the portfolio side. Regress the excess returns of the bonds on a curve against their loadings on those three factors and the residual is a portfolio with no factor exposure at all; the butterfly earlier in this chapter, which neutralises level and slope and bets on curvature, is the same projection stopped one factor short.

Where that residual carries a drift — and deciding whether it does is the hard part — it is also called an alpha, and the construction is (22.5)'s again: there, the next return increment

[17] Amendinger, Imkeller & Schweizer (1998) identify the value with mutual information in general, under a condition on the enlargement that guarantees the price remains a semimartingale in the larger filtration; Pikovsky & Karatzas (1996) solve the enlarged-filtration portfolio problem the value function comes from. Kelly's horse race and this chapter's Gaussian signal are both special cases of the same theorem, not two different results.

was projected onto the unconfirmed part of a signal, run through time; here, each bond's return is projected onto its factor loadings, run across bonds at one moment. Same operation, cross-sectional instead of temporal.

The economics are not the same, and the mutual-information valuation above does not transfer. The residual is not information about anything — it still holds what the factors do not span, a bond's liquidity, its specialness in repo, the flow behind it, so it is a statistical arbitrage rather than an actual one. It is compensation for absorbing a constraint, in the sense the six trades earlier in this chapter used the word, not a signal about a future realisation. What does carry over is the sizing mathematics alone: once the residual has an estimated mean and variance, $\pi^* = \mu/\sigma^2$ prices and sizes a position in it exactly as it does a position in S, applied to a different asset rather than derived differently. The pricing theory behind the factor premia is Ross's Arbitrage Pricing Theory; relative value inverts it, taking the residual to be not quite zero and measuring by how much.

Two honest limits on all of this. The mutual information is an upper bound, attained by an investor with no transaction costs, no position limits and no financing constraint, and the two conditions above's structural conditions are exactly the reasons that investor does not exist; the earlier sections of this chapter are a catalogue of what the bound loses to reality. And the binding difficulty in practice is not acquiring information but estimating α from it, which is chapter 20's identification problem in yet another guise. Equation (22.7) tells you the size of the prize. It does not tell you that you can find it, and chapter 19 argues that this — rather than compute — is what the constraint has always been.

References

- Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, 35(4), 917–926.

- Breiman, L. (1961). Optimal gambling systems for favorable games. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 65–78.

- Algoet, P. H., & Cover, T. M. (1988). Asymptotic optimality and asymptotic equipartition properties of log-optimum investment. *Annals of Probability*, 16(2), 876–898.

- Burghardt, G., & Belton, T. (2005). *The Treasury Bond Basis*. McGraw-Hill.

- Barth, D., & Kahn, J. (2021). Hedge funds and the Treasury cash-futures disconnect. *OFR Working Paper* 21-01.

- Schrimpf, A., Shin, H. S., & Sushko, V. (2020). Leverage and margin spirals in fixed income markets during the Covid-19 crisis. *BIS Bulletin* 2.

- Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3), 341–360.

- Litterman, R., & Scheinkman, J. (1991). Common factors affecting bond returns. *Journal of Fixed Income*, 1(1), 54–61.

- Klingler, S., & Sundaresan, S. (2019). An explanation of negative swap spreads. *Journal of Finance*, 74(2), 675–710.

- Krishnamurthy, A. (2002). The bond/old-bond spread. *Journal of Financial Economics*, 66(2–3), 463–506.

- Jegadeesh, N. (1993). Treasury auction bids and the Salomon squeeze. *Journal of Finance*, 48(4), 1403–1419.

- Fleckenstein, M., Longstaff, F. A., & Lustig, H. (2014). The TIPS-Treasury bond puzzle. *Journal of Finance*, 69(5), 2151–2197.

- Du, W., Tepper, A., & Verdelhan, A. (2018). Deviations from covered interest rate parity. *Journal of Finance*, 73(3), 915–957.

- Perli, R., & Sack, B. (2003). Does mortgage hedging amplify movements in long-term interest rates? *Journal of Fixed Income*, 13(3), 7–17.

- Lou, D., Yan, H., & Zhang, J. (2013). Anticipated and repeated shocks in liquid markets. *Review of Financial Studies*, 26(8), 1891–1912.

- Amendinger, J., Imkeller, P., & Schweizer, M. (1998). Additional logarithmic utility of an insider. *Stochastic Processes and their Applications*, 75(2), 263–286.

- Pikovsky, I., & Karatzas, I. (1996). Anticipative portfolio optimization. *Advances in Applied Probability*, 28(4), 1095–1122.

Chapter 23

Market Making

Everything so far has priced instruments. This chapter quotes them, which is a different problem: a price is a number and a quote is a pair of numbers offered to somebody who gets to choose which side to take. We derive the one genuinely solvable quoting model, and find that the substitution which solves it is the same that solved the affine models — exponential utility makes a nonlinear control problem linear. Then we look at what that model cannot see, which is that the counterparty may know something, and at how a rates desk measures whether its flow is toxic. The chapter ends where a desk's day ends: with a position nobody chose, because hedging it completely was not worth what it cost, which is what chapter 24 has to carry.

23.1 What the Spread Is Payment For

A market maker undertakes to quote both sides of a price on request, and is paid the difference between them. The undertaking is the product: a client who wants to hedge a liability in size, now, does not want to wait for a natural counterparty to appear. Immediacy is the service and the spread is its fee.

That much is uncontroversial and explains nothing about how wide the spread should be. Two separate costs set it.

- *Inventory risk.* Having bought, the maker holds a position it did not choose, and the position has variance. This cost is symmetric in the sign of the flow, grows with volatility, and would exist even if every counterparty were a coin flip.

- *Adverse selection.* The counterparty chose to trade, and chose which side. If they know something, the maker is on the wrong side of it. This cost is asymmetric by construction — it exists precisely because the client selects — and would exist even if the maker could hedge instantly and hold no inventory at all.

The two point in the same direction on the spread and for opposite reasons, which is why a desk that models one and neglects the other can be badly wrong while appearing to have a theory. The next section builds the model of the first. The one after that is about the second, which no tractable model of the first contains.

23.2 A Solvable Quoting Problem

The canonical model is Avellaneda and Stoikov's. It is the only quoting model in this book that is solvable in the sense of chapter 14, and *why* it is solvable turns out to be a result we already have.

Definition 23.1 (The quoting problem)**.** A mid price diffuses with no drift, $dS_t = \sigma\, dW_t$. The maker continuously chooses distances $\delta_t^a, \delta_t^b \geq 0$ and stands ready to sell at $S_t + \delta_t^a$ and buy at $S_t - \delta_t^b$. Fills of one unit arrive as Poisson processes whose intensity falls with the distance,

$$\lambda(\delta) = Ae^{-k\delta}, \tag{23.1}$$

independently on the two sides. Writing X_t for cash and q_t for inventory, the maker maximises exponential utility of terminal wealth,

$$u(t,x,q,s) = \sup_{\delta^a,\delta^b} \mathbb{E}\Big[-\exp\big(-\gamma(X_T + q_T S_T)\big)\Big]. \tag{23.2}$$

Equation (23.1) says the further from the mid one quotes, the less one trades, with a sensitivity k. It is the only place the market's willingness to deal enters, and it is exogenous: the arrival rate depends on where the maker quotes and on nothing else. In particular it does not depend on where the price is about to go, and §23.3 is about the consequences.

The exponential utility in (23.2) looks like a taste and is a tractability assumption, and a specific one: constant absolute risk aversion means $\partial_x u \propto u$, so wealth cancels out of (23.3) entirely, which is what collapses the four-dimensional problem to the two-dimensional linear system below. A power utility's derivatives do not scale with u itself and would leave wealth in the equation. That is not the only tractable formulation of the quoting problem — Guéant, Lehalle and Fernandez-Tapia solve it with a running inventory penalty in place of a terminal utility, linearising by a related but different route — exponential utility is simply what makes *this* reduction go through.

And the mid price is a martingale, so the maker has no view. That is the right convention: a maker with a view should express it as a position, not by leaning the quote, and the two decisions are cleanly separable in this model precisely because the drift is zero.

The equation

Calculation 23.2 (The Hamilton-Jacobi-Bellman equation)**.** Over $[t, t+dt]$ three things can happen: nothing, a sale, or a purchase. Collecting the diffusion of s and the two jump terms, the value function satisfies

$$\begin{aligned} 0 = \partial_t u + \tfrac{1}{2}\sigma^2 \partial_{ss} u &+ \sup_{\delta^a} \lambda(\delta^a)\Big[u\big(t, x+s+\delta^a, q-1, s\big) - u\Big] \\ &+ \sup_{\delta^b} \lambda(\delta^b)\Big[u\big(t, x-s+\delta^b, q+1, s\big) - u\Big]. \end{aligned} \tag{23.3}$$

A sale moves cash up by the price received, $s+\delta^a$, and inventory down by one; a purchase does the reverse. The suprema sit *inside* the equation, which is what makes it a control problem rather than a valuation.

Equation (23.3) is unpromising. It is nonlinear, it contains two optimisations, and the state has four dimensions. What rescues it is that the objective was chosen to make a particular family invariant.

Theorem 23.3 (The substitution that solves it)**.** *Write*

$$u(t,x,q,s) = -\exp\big(-\gamma(x+qs)\big)\cdot v(t,q)^{-\gamma/k}. \tag{23.4}$$

Then (23.3) *holds if and only if* v *satisfies the* linear *system*

$$-\frac{\partial v}{\partial t}(t,q) = -\alpha\, q^2\, v(t,q) \;+\; \eta\Big(v(t,q-1)+v(t,q+1)\Big), \tag{23.5}$$

with $v(T,q)=1$ *and constants*

$$\alpha = \frac{k\gamma\sigma^2}{2}, \qquad \eta = A\left(1+\frac{\gamma}{k}\right)^{-(1+k/\gamma)}. \tag{23.6}$$

The optimal distances are then

$$\delta^a(q) = \frac{1}{k}\ln\frac{v(t,q)}{v(t,q-1)} + \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{k}\right), \qquad \delta^b(q) = \frac{1}{k}\ln\frac{v(t,q)}{v(t,q+1)} + \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{k}\right). \tag{23.7}$$

Proof. Substitute (23.4) and divide through by the positive quantity $\exp(-\gamma(x+qs))$, which appears in every term. Writing $\theta(t,q)=\frac{1}{k}\ln v(t,q)$, so that $u=-\exp(-\gamma(x+qs+\theta))$, the derivatives are

$$\partial_t u = \gamma\,\partial_t\theta\cdot|u|, \qquad \partial_s u = \gamma q\,|u|, \qquad \partial_{ss}u = -\gamma^2 q^2\,|u|.$$

A sale takes u to $-\exp(-\gamma(x+qs+\delta^a+\theta(t,q-1)))$, since the cash gained is $s+\delta^a$ and the inventory term loses one s. So the jump bracket is

$$u(\text{after}) - u = |u|\Big[1-e^{-\gamma(\delta^a+\Delta^a)}\Big], \qquad \Delta^a = \theta(t,q-1)-\theta(t,q).$$

Dividing (23.3) by $|u|$ leaves an equation in θ alone — the cash, the price and the utility scale have all cancelled, which is the point of the exponential form:

$$\gamma\,\partial_t\theta - \tfrac{1}{2}\gamma^2\sigma^2 q^2 + \sup_{\delta^a} G(\delta^a,\Delta^a) + \sup_{\delta^b} G(\delta^b,\Delta^b) = 0, \tag{23.8}$$

where $G(\delta,\Delta) = Ae^{-k\delta}\big[1-e^{-\gamma(\delta+\Delta)}\big]$.

Now do the optimisation, which is one line. Setting $\partial G/\partial\delta = 0$ and writing $z=e^{-\gamma(\delta+\Delta)}$,

$$-k(1-z)+\gamma z = 0 \quad\Longrightarrow\quad z = \frac{k}{k+\gamma} \quad\Longrightarrow\quad \delta^* = \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{k}\right) - \Delta,$$

which is (23.7) once Δ is written in terms of v. Substituting back, G at its optimum is $Ae^{-k\delta^*}\gamma/(k+\gamma)$, and

$$e^{-k\delta^*} = \left(1+\frac{\gamma}{k}\right)^{-k/\gamma}\frac{v(t,q\mp1)}{v(t,q)},$$

so both suprema in (23.8) become multiples of $v(t,q\mp1)/v(t,q)$. Finally $\partial_t\theta = \partial_t v/(kv)$; multiplying (23.8) through by kv/γ clears every denominator and gives (23.5) with the constants (23.6). □

Solving (23.5) produces a function satisfying (23.3). That it is *the* value function (23.2) defines — and not merely some solution of the same equation — needs one more step.

Theorem 23.4 (Verification)**.** *Let $\tilde{u}$ be any sufficiently regular solution of* (23.3) *with terminal condition $\tilde{u}(T,x,q,s) = -\exp(-\gamma(x+qs))$ — the function* (23.4) *constructs, for instance. Then $\tilde{u}$ equals the value function* (23.2)*, and the distances* (23.7) *are the optimal quotes.*

Proof. Fix any admissible control (δ_t^a, δ_t^b) and apply Itô's formula to $\tilde{u}(t, X_t, q_t, S_t)$ along the path it generates. S contributes the usual diffusion terms; X and q jump when a fill arrives, at the controlled intensities $\lambda(\delta_t^a)$ and $\lambda(\delta_t^b)$, and compensating those jumps leaves

$$\mathbb{E}\big[\tilde{u}(T, X_T, q_T, S_T)\big] - \tilde{u}(0,x,q,s) \;=\; \mathbb{E}\left[\int_0^T \Big(\partial_t\tilde{u} + \tfrac{1}{2}\sigma^2\partial_{ss}\tilde{u} + \lambda(\delta_t^a)\,\mathcal{J}_t^a + \lambda(\delta_t^b)\,\mathcal{J}_t^b\Big)\,dt\right],$$

where $\mathcal{J}_t^a, \mathcal{J}_t^b$ are (23.3)'s two jump brackets evaluated at the current state (t, X_t, q_t, S_t) and the current control, and the martingale parts of the diffusion and the compensated jumps vanish in expectation. Because $\tilde{u}$ solves (23.3), the two suprema together sum to zero, so for *any particular* δ_t^a, δ_t^b — not necessarily the maximiser — the integrand is at most zero: one term of a supremum that totals to it. The integral is therefore non-positive, and since $\tilde{u}(T,\cdot)$ is exactly the objective's terminal payoff,

$$\mathbb{E}\big[-\exp\big(-\gamma(X_T + q_T S_T)\big)\big] \;\le\; \tilde{u}(0,x,q,s)$$

for every admissible control: $\tilde{u}(0,\cdot)$ dominates every achievable payoff, hence dominates the supremum (23.2) defines.

Now take the specific control $\delta_t^{a*} = \delta^a(q_{t^-})$, $\delta_t^{b*} = \delta^b(q_{t^-})$ of (23.7) — the one attaining the sup in (23.3) at every state. Along the path it generates the integrand above is identically zero by construction, so the inequality is an equality: this one control already achieves the payoff $\tilde{u}(0,x,q,s)$. A control that achieves a value cannot be beaten by a supremum that dominates it, so (23.2)'s supremum is at least $\tilde{u}(0,x,q,s)$ too. The two inequalities together force $u = \tilde{u}$, and force δ^* to be optimal. □

Structure (The same trick as the affine models)**.** A control problem and a pricing problem have just turned out to be the same kind of object.

Chapter 14 showed that a nonlinear equation becomes linear when the unknown is an exponential whose parameters the equation can absorb — there the family was $e^{\phi+\psi x}$, the parameters moved by a Riccati equation, and the nonlinearity came from the second derivative bringing a parameter down twice. Here the family is $e^{-\gamma(x+qs+\theta)}$, and γ, the risk aversion, plays the role ψ played there. Dividing (23.3) by $|u|$ is the step that removes the wealth and the price from the problem entirely, exactly as dividing by u removed the state there.

The difference is instructive. In the affine case the reduction left a *Riccati* equation, quadratic in the parameter. Here it leaves a *linear* one, and the reason is visible in the proof: the price enters the utility linearly, through qs, so the second derivative $\partial_{ss}u$ contributes a term in q^2 that is a coefficient rather than an unknown. The nonlinearity that would have made it Riccati was spent on the optimisation instead, and the optimisation closed in one line.

So exponential utility is not a convenience here in the sense of making the algebra shorter. It is the choice that makes an invariant family exist, and Definition 23.1 would be a numerical problem

with any other utility. That is the same trade chapter 14 describes throughout: a modelling restriction adopted because it is what makes the mathematics close.

What it says

Solving (23.5) numerically and reading off (23.7) gives the quotes.[1] At $\sigma = 30\%$, $\gamma = 0.1$, $A = 140$, $k = 1.5$ and a one-year horizon:

Inventory q	bid distance	ask distance	total spread	skew
-10	0.574	0.725	1.300	$+0.075$
-5	0.613	0.685	1.298	$+0.036$
0	0.649	0.649	1.298	0
5	0.685	0.613	1.298	-0.036
10	0.725	0.574	1.300	-0.075

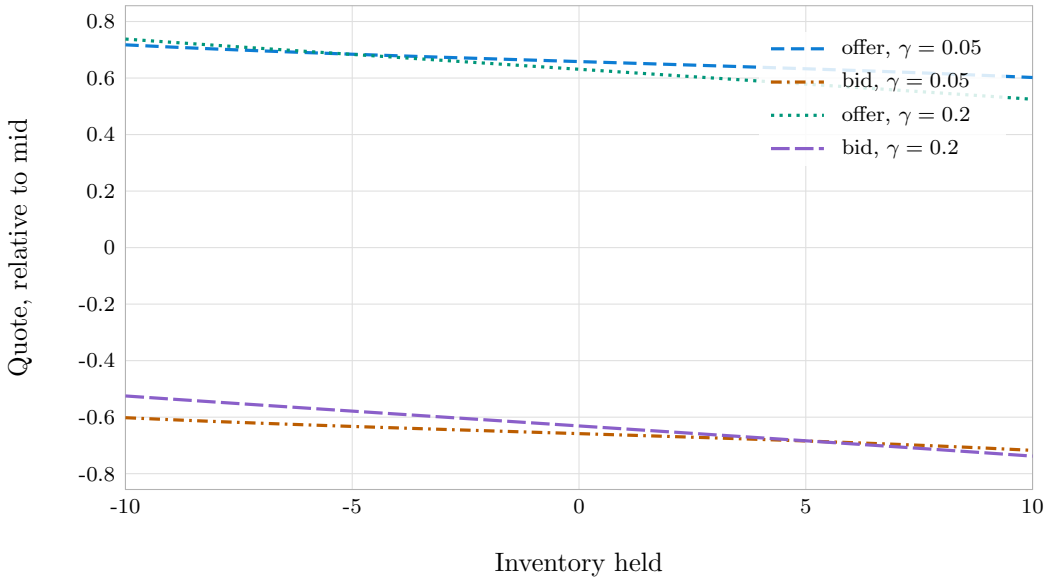

Figure 23.1: The solved quotes against the inventory held, at two levels of risk aversion. Read the pair rather than either line: as the position grows the bid and the offer slide *down together*, keeping their width, so a maker who is long shows a better offer and a worse bid rather than a wider market. Raising γ steepens the slide without changing the width. The width is set by the arrival decay k and the risk aversion; the slide is set by the position, and the two are independent, which is the model's one genuinely practical statement.

Drawn by `Quoting::linear_solution` and `Quoting::quotes`.

[1] `marketmaking::Quoting`, which also solves (23.3) directly without the substitution. The two agree to the discretisation — the gap halves each time the time step does and extrapolates to nothing — which is how one checks that a substitution linearises a problem rather than approximating it.

Remark (Inventory skews the quote, it does not widen it)**.** The fourth column is nearly constant and the fifth is not, and that is the model's central practical statement. A maker who is long does not quote a wider market; it quotes a *lower* one — a better ask and a worse bid — and keeps the width. The position is managed by moving both quotes down together, so that the flow it attracts is the flow that reduces the position.

This is the analogue of an indifference price. A maker holding q values the asset below the mid, by

$$\text{reservation offset} = -q\gamma\sigma^2(T-t), \tag{23.9}$$

and quotes symmetrically about *that* rather than about the mid. The offset is linear in the inventory, in the risk aversion and in the variance still to be borne, all three of which are as one would hope.

The width itself is the other term of (23.7), and it has a floor. As $T-t \to 0$ the inventory term vanishes and the spread does not: it tends to $(2/\gamma)\ln(1+\gamma/k)$, which for these parameters is 1.291 of the 1.298 quoted at a year. Almost all of the spread here is not compensation for risk at all. It is the margin implied by (23.1) — the maker quotes wide because it can, and k is what limits how wide.

Remark (Which parameters a desk can actually observe)**.** Of the four, σ is estimable from the market and γ is a policy choice, however uncomfortably. The other two are the difficulty. A and k describe how the arrival of client business responds to the price shown, they are properties of a franchise rather than of a market, and estimating them requires the desk's own historical quotes and fills — data no external party has, and which is heavily censored, since a quote that lost tells you the trade happened elsewhere but not at what level.

This is chapter 20's identification problem in a new place. The parameter that matters most for the answer, k, is the one estimated from the least reliable data, and (23.7) depends on it directly rather than through a second-order term.

23.3 What the Model Cannot See

Return to (23.1). The intensity depends on the distance quoted and on nothing else. In particular the arrival of a fill carries no information about the subsequent path of S: a buy and a sell are equally likely at symmetric distances, whatever is about to happen.

This is not a small idealisation. It removes adverse selection from the problem entirely, and adverse selection is the thing a real desk spends most of its attention on. The model's inventory is a risk — a symmetric bet that happens to be unwanted. Real inventory is worse than that, because of *how* it was acquired: the maker is long precisely because somebody chose to sell.

Example 23.1 (The minimal model of the other cost)**.** Glosten and Milgrom's model isolates what Avellaneda-Stoikov omits exactly as cleanly as that model isolates inventory.

An asset is worth either V^+ or V^-, each with probability one half. A fraction π of arriving traders know which; the rest trade for reasons of their own, buying or selling with equal probability. The maker holds no inventory — it can offload instantly — and is competitive, so it quotes at zero expected profit.

The informed always buy at the ask if the value is V^+ and sell at the bid if it is V^-. So a buy order is evidence for V^+, and the ask must be the conditional expectation of the value *given* that a buy arrived:

$$\text{ask} = \mathbb{E}\big[V \,\big|\, \text{buy}\big], \qquad \text{bid} = \mathbb{E}\big[V \,\big|\, \text{sell}\big],$$

and the spread between them is entirely the information content of the order. Working the conditional expectations out,

$$\text{spread} = \pi\big(V^+ - V^-\big), \tag{23.10}$$

linear in the fraction of informed flow and in how much there is to know.

Not that there is no inventory anywhere in (23.10) and no risk aversion, so this spread is not compensation for bearing anything — a risk-neutral maker with instant hedging quotes it too. And the maker earns nothing on average: it loses to the informed exactly what it makes from the rest. The spread is not profit, it is the price of not being able to tell the two apart.

Kyle's continuous-auction model makes the same point with a strategic informed trader choosing an order size rather than a discrete buy-or-sell decision, and gets a linear-price-impact rule out of the equilibrium rather than a fixed spread — the mechanism is the same adverse selection, priced by a different market structure.

Structure (Two costs, two models, no model with both)**.** The two models are complementary in a way that is more than tidy.

Avellaneda-Stoikov has inventory and no information: its flow is exogenous, its cost is variance, and its answer is a skew. Glosten-Milgrom has information and no inventory: its flow is endogenous, its cost is being picked off, and its answer is a width. Each is solvable, and each is solvable *because* it omits the other's mechanism — allowing the arrival intensity of (23.1) to depend on the future price would destroy the invariance theorem 23.3 needs, and giving Glosten-Milgrom's maker an inventory it must hold removes the zero-profit condition that pins its quotes.

Models with both exist and none is tractable. A desk therefore does not use a single model; it uses one model for the skew, another argument for the width, and measurement for everything else. The rest of this chapter is about the measurement, because that is where the actual work is.

23.4 Measuring Toxicity

If a fill is followed by the price moving against the maker, the spread it earned was an illusion. The standard measurement makes exactly that comparison.

Definition 23.5 (Markout)**.** For a fill of signed size Q (positive when the maker bought) at price p against a mid m_0, and the mid m_h observed h later, the *markout* at horizon h is

$$\text{markout}(h) = Q\big(m_h - p\big). \tag{23.11}$$

At $h = 0$ it is the captured spread, $Q(m_0 - p)$, which is positive by construction.

The object of interest is the whole curve of markout against h, and its shape is the diagnosis. A flat curve means the maker keeps what it earned. A curve decaying towards zero means the price drifts against the fill and the spread is being given back. A curve going negative means the flow is worth less than nothing: the maker would have done better to decline.

Remark (The decomposition this implies)**.** Writing $s_{\text{eff}} = Q(m_0 - p)$ for the spread captured and $I(h) = Q(m_0 - m_h)$ for the price impact over the horizon, (23.11) rearranges to

$$\underbrace{\text{markout}(h)}_{\text{what is kept}} = \underbrace{s_{\text{eff}}}_{\text{what is charged}} - \underbrace{I(h)}_{\text{what moves against}} . \tag{23.12}$$

This is microstructure's effective, realised and impact triple, and only the middle term is a decision. The desk sets s_{eff}; the market determines I; what the desk keeps is the residual. Widening the quote raises the first term and, through (23.1), reduces the number of fills — but it also changes *which* fills arrive, and generally not in the desk's favour, since the clients most willing to pay a wide spread are the ones most confident about direction.

Remark (Horizons, and what each one is measuring)**.** The choice of h is not a detail. A few seconds measures whether the desk is being arbitraged by faster participants. A few minutes measures whether the client had a short-lived signal. A day or more measures whether the client was simply right — which, for a pension fund hedging a liability, it may well be without anything improper having occurred.

Flow that is directional but slow is not toxic in the sense that matters; it is a hedging cost that the desk can price and manage by holding the position, and refusing it damages a franchise for no gain. Toxic flow is flow whose information decays faster than the desk can hedge. The relevant comparison is not markout against zero but markout against the desk's own time to flatten — the same distinction Easley, López de Prado and O'Hara's flow-toxicity measure is built to make precise, from order-flow imbalance rather than markout.

23.5 The Race the Short Markout Measures

A markout at a few seconds was described above as measuring whether the desk is being arbitraged by faster participants. The taxonomy of chapter 22 turns on it: the mispricings that vanish immediately are the ones somebody is racing for, and this is what the racing consists of.

The section is description rather than derivation, and is marked as such. Nothing in it is measured here — doing so honestly would need tick and venue data this book does not carry — so it is offered as an account of a mechanism, with the argument at the end being the part that is not merely reportage.

What the race is physically made of

The winning trade is decided by the time between a message arriving and an order leaving, and the industry has spent two decades compressing it.

The path is shortened first. Machines sit in the exchange's own data centre, and links between venues follow the straightest available line — microwave and millimetre wave rather than fibre, because light moves about half as fast again through air as through glass and the towers can run closer to a great circle than cable trenches do. Then the machine is shortened: the operating system is taken out of the path, and the decision logic is put into programmable hardware so that a packet can be parsed and an order emitted without a general purpose processor being involved. Then the

message itself is shortened, by decoding the exchange's binary protocol directly, acting on a partial message before the remainder has arrived, and taking whichever of two redundant feeds happens to arrive first.

And on a price-time priority book there is a further asset that is not speed but is bought with it: queue position. Being early at a price level determines whether a resting order is filled at all, it accrues only by resting, and it is forfeited by cancelling — which is what makes the decision to pull a stale quote expensive rather than free.

Two roles, one race

The strategies divide into two.

A taker watches a fast instrument and trades against a stale quote in a slower related one — a future against its underlying, one venue against another. Its race is against the maker's cancel.

A maker posts quotes and tries not to be the stale one. Its race is against the taker's take, and every fill it loses shows up as a negative short-horizon markout: exactly the $I(h)$ of (23.12) at small h. So Definition 23.5 is not merely diagnostic of this activity, it is the scoreboard for it.

Why the spending does not stop, and who pays for it

Here is the part that is an argument rather than a description.

A mispricing is a bounded quantity available at a stale price, so the fastest participant takes essentially all of it and the second fastest takes nothing. A payoff of that shape has a known consequence: competitors bid for the advantage until the cost of holding it consumes the rent. The profit pool is fixed by how many stale quotes the market produces, and it does not grow when everyone gets faster — so the spending converts a flow of trading profit into a stock of fixed cost, and the participants end up no better off than before while the equipment gets more expensive.

That equilibrium is not a fact about human nature. It is a consequence of the market's *design*. A continuous limit order book resolves simultaneous orders by arrival time, so it makes an infinitesimal speed advantage decisive, and Budish, Cramton and Shim showed that this is what manufactures the race. Batching orders and crossing them at a single price every fraction of a second removes it, because orders arriving within a batch are simultaneous by construction and speed stops being the tiebreak. The arms race is a choice of matching rule, and a different rule would end it.

Meanwhile the cost is not borne by the participants who lose the race. A maker that is picked off must widen until the spread covers the expected loss, so the expenditure reappears in s_{eff} and is paid by everyone who trades, including those who have never heard of any of this.

Remark (How much of this reaches a rates desk)**.** Less than one might think, and the next section says why: most rates risk changes hands by request for quote rather than in a lit book, and a request that arrives at several dealers at once is not won by microseconds.

Where it does reach the desk is in the hedge. The benchmark futures and on-the-run bonds a swaps book hedges itself with *are* traded in central limit order books, so the desk meets the race not when it prices the client but when it lays the risk off — which puts the cost in the hedging leg of §23.8 rather than in the quote, and makes it a reason the hedge is partial rather than a reason the spread is wide.

23.6 Quoting on a Rates Desk

The general theory above is agnostic about the market. Rates market making has a structure that changes the problem materially, and in four ways.

Requests, not a book

Most rates risk does not change hands in a limit order book. A client sends a request for quote to several dealers at once and trades with the best price. Benchmark futures and on-the-run government bonds have genuine order books; swaps, swaptions, and everything structured are quoted on request.

That inverts the adverse selection problem in a way, because it produces the cost *without any informed trader at all*.

Calculation 23.6 (The winner's curse)**.** Suppose n dealers each value an instrument at its fair value plus an independent error of standard deviation s, reflecting different curves, different models and different axes. Nobody is informed; nobody has an edge; the errors have mean zero. The client trades with the best price.

The dealer who wins is the one whose error was most aggressive, so conditioning on having won selects the minimum of n draws. Measuring it:[2]

Dealers	2	3	5	10
Expected error of the winner, in units of s	-0.56	-0.84	-1.16	-1.54

So against two competitors the winner has underpriced by more than half its own pricing error, and against ten by more than one and a half times it. Eventually the growth is like $\sqrt{2\ln n}$, which is slow: ten competitors cost less than three times what two do, not five times.

Structure (Winning is information)**.** This is separate from the informed-trader story, and the remedy is different.

In Glosten-Milgrom the maker loses because the counterparty knows something. Here nobody knows anything, and the maker still loses, because *the act of winning* is evidence about its own price. The fill is a draw from a conditional distribution — conditional on having been the most aggressive quote — and that distribution has a worse mean than the unconditional one.

The consequence is that a dealer's own historical fills are a biased sample of its own pricing, and biased in a direction that makes the desk look better than it is: the trades it did are the ones it priced most aggressively, and the ones it declined or lost are absent.

Practically it means the correction is a function of how many dealers are in competition, which the client knows and the dealer often does not. A desk that quotes the same margin into a two-dealer and a ten-dealer request is systematically mispricing one of them.

[2] `marketmaking::winners_curse`, with the two-dealer case checked against its closed form $-1/\sqrt{\pi}$.

Who is on the other side

The client's identity is the single most informative variable available, and unlike almost everything else in this book it is known before the quote is made. A desk's flow divides roughly as follows.

- *Real money* — pension funds, insurers, liability-driven investors. Large, directional, slow, and usually one-way for months at a time. Markout at a day is often against the desk, because these clients are right about direction on a horizon far longer than the desk's. This is a hedging cost to be priced, not toxicity to be refused.

- *Corporates.* Hedging an issuance or a floating exposure. Benign, infrequent, and price-insensitive in the sense of (23.1) — the low-k flow that pays for the franchise.

- *Fast funds.* Small in size, high in frequency, and selecting on exactly the horizons the desk cannot hedge within. Markout decays within minutes. This is the flow the word toxic was invented for.

- *Other dealers.* Interdealer trades are usually somebody laying off risk they just took, so a request is evidence about the market's aggregate position, and one dealer's hedge is another's inventory.

A quote is on a portfolio of factors

This is the structural feature with no equity analogue, and it is a direct consequence of chapter 8's picture of the state as a curve.

A request for a ten-year swap is not a request on a single asset. The instrument's value depends on the whole curve, and its risk decomposes onto the factors of chapter 12 — level, slope, and whatever else the model carries. So a desk that has bought ten-year risk has not acquired a position in the ten-year point; it has acquired a particular combination of level and slope, and whether that combination offsets or compounds the existing book depends on what the book already holds.

Two consequences follow immediately.

The first is that the inventory variable q of Definition 23.1 is the wrong object. The relevant state is a vector of factor exposures, and the reservation offset (23.9) generalises to one in which $\gamma\sigma^2$ becomes a covariance matrix and q a vector. The skew on any one instrument then depends on the whole book through that matrix, which is why a desk long the front end may quote a tight offer in the long end: the trade that reduces its net slope exposure is not the trade that reduces its position in the instrument requested.

The second is that some risk is far cheaper to shed than the rest. The benchmark points are liquid; the points between them are not. A position in a ten-year swap can be hedged in size within seconds; the same risk expressed at seven years and three months cannot, and must be hedged with a combination of benchmarks that leaves a residual. The residual is a basis risk, it is small per trade, and it accumulates in one direction because clients ask for the dates their liabilities fall on rather than the dates the market is liquid at.

Volatility that cannot be hedged

The swaption surface makes the same point more sharply. A desk quoting a seven-year option on a five-year swap has taken vega in a bucket where nothing liquid trades. It can hedge the delta immediately and the vega approximately, using whatever combination of quoted expiries and tenors comes closest, and the remainder is warehoused — carried deliberately, because the alternative does not exist.

Chapter 12's parameters are exactly what determine whether that remainder is dangerous. If the model's mean reversion is wrong, the mapping from the liquid buckets to the illiquid one is wrong, and the desk's belief about what it has hedged is wrong in a way no amount of care with the delta repairs. This is the practical cost of the identification problem that chapter describes: a parameter nobody can see from European prices is what decides whether a warehoused vega position is offset or doubled.

23.7 A Signal, and Who It Is Worth More To

Nothing so far has let the maker have a view. Definition 23.1 makes the mid a martingale precisely so that quoting and forecasting stay separable. What happens when they are not?

Suppose the maker believes fair value is μ rather than the observed mid S_t. The machinery already derived handles it without modification: the quotes were never centred on the mid but on the mid plus the reservation offset (23.9), and a signal enters exactly where inventory does — as a displacement of the centre, so it is the *deviation* $\mu - S_t$ that appears rather than μ itself.

$$\text{centre of the quoted market} \;=\; \underbrace{S_t}_{\text{mid}} \;\underbrace{-\,q\gamma\sigma^2(T-t)}_{\text{inventory}} \;\underbrace{+\,(\mu - S_t)}_{\text{signal}} \;=\; \mu - q\gamma\sigma^2(T-t). \tag{23.13}$$

Remark (A signal skews the quote, it does not tighten it)**.** Equation (23.13) moves the *centre* and leaves the width alone, which is the same conclusion the inventory term reached and for the same reason. The width answers a different question — how price sensitive the flow is, and how much risk the position carries — and a forecast bears on neither.

So a maker who becomes bullish shows a better bid and a worse offer. Read from one side that looks like tightening, but it is not a symmetric tightening which would be the wrong response. Tightening both sides buys more flow of both kinds, which is a decision to trade more rather than a decision to express a view; the maker would acquire the position it wants and the position it does not in equal measure, and the signal would earn nothing. The whole content of acting on a forecast is the asymmetry.

Structure (The same signal is worth more to a maker than to a taker)**.** There is an asymmetry here that is easy to state and easy to miss, and it explains why the same research is used differently on the two sides of a market.

A hedge fund or proprietary trader with a signal of size e must *cross the spread* to express it. The signal is worth e less the round trip, and chapter 22's calculation applies: below a threshold set by the cost there is no trade at all, however good the forecast.

A market maker with the same signal does not cross anything. It shifts (23.13) and waits, and the position arrives through flow it was going to quote on anyway — while being *paid* the spread for taking it. The signal's value is therefore *e plus* the spread earned on the trades it selects, and the breakeven signal strength is near zero rather than a round trip away.

Two consequences follow. A signal too weak to trade on can still be worth acting on if one is quoting, which is why market making desks fund research that would not clear a proprietary trader's threshold. And the maker's version is capacity constrained in a way the taker's is not: the position accumulates only as fast as clients happen to ask, so a strong short-lived signal cannot be expressed in size, and a maker with a genuinely urgent view has to cross the spread like everybody else — at which point it is a taker, and pays.

The asymmetry is not an inefficiency. It is payment for the service of §23.3: the maker is exposed to adverse selection on every quote it shows, and the right to monetise a signal cheaply is part of what compensates that exposure.

23.8 The Partial Hedge

Everything above converges on one decision. The desk cannot hedge completely — sometimes because the instrument does not exist, always because trading costs money — and the question is how much to leave. Almgren and Chriss ask the same cost-against-risk question in its dynamic form, over a trajectory of trades that liquidates a position by a deadline; here the version needed is a single, static fraction, and it has a closed form.

Calculation 23.7 (How much to leave)**.** Take a unit exposure with variance Σ over the horizon it would be held. Hedging a fraction h costs $c\,h$ in spread and fees and leaves residual variance $\Sigma(1-h)^2$, penalised at $\gamma/2$ as in (23.2). Minimising

$$c\,h + \frac{\gamma}{2}\Sigma\,(1-h)^2$$

over h gives $c = \gamma\Sigma(1-h)$, so

$$1 - h^* = \frac{c}{\gamma\Sigma}, \tag{23.14}$$

truncated at zero and one.[3]

Remark (What (23.14) says)**.** The fraction left unhedged is the transaction cost divided by the risk penalty.

It is zero only if hedging is free. Any positive cost leaves a residual, so a completely hedged book is not a well-run one — it is one that has paid too much. At $\Sigma = 4\%$ and $\gamma = 2$, a cost of one per cent of notional leaves an eighth of the risk standing; four per cent leaves half of it.

It is total when the cost is large enough. Past $c = \gamma\Sigma$ the optimum is to hedge nothing at all, which is the illiquid vega case above: the bid-ask on the only available hedge exceeds the risk it removes, and the position is warehoused as a matter of arithmetic rather than of nerve.

The residual is proportional to the cost and inversely proportional to the variance, so the positions a desk is most tempted to leave unhedged — illiquid, wide, and therefore expensive to trade — are exactly the ones (23.14) says to leave the most of.

[3] `marketmaking::optimal_hedge_fraction`, checked against a direct minimisation.

Structure (Hedging is a projection, and the residual is what is left over)**.** Equation (23.14) is one-dimensional and the real problem is not, but the shape of the answer survives.

With several hedging instruments, the exposure is a vector and the hedgeable directions span a subspace. The best hedge ignoring costs is the orthogonal projection onto that subspace, and the unhedgeable part is the residual — which is exactly chapter 6's structure One price per risk, and no-arbitrage is what sets it, where the market price of risk was a projection of excess returns onto the span of the volatility matrix and the residual was an arbitrage. Here the residual is not an arbitrage, because these instruments are not available at any price. It is a position.

Adding costs tilts the projection, shrinking it towards zero by the factor (23.14) in each direction according to what that direction costs to trade. So a desk's residual risk has two quite different components: the part that is unhedgeable because no instrument spans it, and the part that is unhedged because spanning it was not worth the fee. The first is a modelling constraint and cannot be reduced by spending more. The second is a choice, and can.

23.9 What Is Handed On

A day of market making ends with a book that nobody designed. Its composition is the sum of what clients asked for, which is not a portfolio anybody would choose; its hedges are approximate by decision, per (23.14); and its residual concentrates in the illiquid corners, because those are the corners where hedging was least worth the cost.

That object is what chapter 24 is about, and the chain should be stated in one place since it runs through the whole book. The quote came from a model, so it inherits the model's assumptions — chapter 20's question of whether the model has a view on what the trade is a bet on. The hedge came from the model's sensitivities, so it inherits the errors chapter 10 measures, where a wrong backbone produces a delta wrong by five per cent of notional in a consistent direction. And the residual is real, unavoidable and one-signed.

So the risk that arrives at the end of the chain is not the risk of a chosen position. It is the risk of a position accumulated by a process, hedged by a model, and left partly open on purpose. Measuring it requires knowing which of those three contributed what, which is why chapter 24 begins with the structure A number without a decomposition is not a risk number.

References

- Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, 8(3), 217–224.

- Gueant, O., Lehalle, C.-A., & Fernandez-Tapia, J. (2013). Dealing with the inventory risk: a solution to the market making problem. *Mathematics and Financial Economics*, 7(4), 477–507.

- Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1), 71–100.

- Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6), 1315–1335.

- Easley, D., Lopez de Prado, M., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. *Review of Financial Studies*, 25(5), 1457–1493.

- Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: frequent batch auctions as a market design response. *Quarterly Journal of Economics*, 130(4), 1547–1621.

- Almgren, R., & Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of Risk*, 3(2), 5–39.

Chapter 24

Risk Management

We stop asking what a trade is worth and start asking how wrong we might be about it. We set out the axioms a risk measure ought to satisfy, show that the industry's standard measure fails one of them, derive the measure that repairs it, and then work through how such numbers are actually computed, tested, and set aside as capital — including for the model risk the previous four chapters kept uncovering.

24.1 A Different Question

Every chapter so far has answered the same question: what is this worth? The answer was always an expectation under a measure chosen so that no arbitrage is possible, and the work was in finding the measure and computing the expectation.

Risk management asks something else. Given that we have marked the book at those prices, how much could we lose, how likely is that, and how much capital should be held against it? The question is asked in the real world measure — the $\mathbb{P}^{\text{historic}}$ of chapter 4, the one that dropped out of every pricing formula — because we are now asking what will actually happen rather than what a hedged position costs.

This is a genuine reversal. The whole apparatus of chapters 4 through 12 was built to eliminate the historic measure from pricing. Risk management puts it back, and the two answers are not comparable: a position can be worth zero and be enormously risky, which is precisely what a hedged book is.

Three kinds of risk turn up:

- *Market risk*: the book loses money because prices move. The subject of most of this chapter.
- *Model risk*: the book was marked with a model, and the model was wrong. Chapters 9, 10 and 11 each ended by identifying a parameter that the calibration could not determine and the product depended on. This chapter says what to do with them.

- *Counterparty risk*: the trade was right and the other side did not pay.

Structure (A number without a decomposition is not a risk number)**.** A risk measure returns one number. That number is nearly useless on its own, because the position it describes was assembled by three quite different processes and only one of them is a live decision.

- *Risk the desk chose.* A view, taken deliberately, in a size somebody signed off. This is what a risk report is usually imagined to be about, and in a hedged book it is the smallest of the three.

- *Risk the flow chose.* Chapter 23 ends with a book nobody designed: its composition is the sum of what clients asked for. Nothing about it was selected, and the only decisions available are how much to quote and what to hedge.

- *Risk the model chose.* The hedge was computed from a model. If the model is wrong, the hedge is wrong, and the residual is a position the desk holds without knowing it holds. Chapter 10 measures one instance — a delta wrong by five per cent of notional, in a consistent direction — and §24.8 measures how large this term is in general.

Only the first can be changed by deciding to change it. The second is changed by quoting differently, which is a business decision with revenue on the other side. The third cannot be changed at all without changing the model, and no amount of capital held against it makes it smaller.

So the useful output is not $\rho(L)$ but the split, and a risk function whose answer cannot be attributed is not doing the job it is there for. That claim is the reason this chapter comes last.

24.2 What a Risk Measure Should Satisfy

Write L for the loss on a portfolio over some horizon — a random variable, positive when money is lost. A risk measure is a function ρ turning that random variable into a single number, to be read as the capital that should stand behind it.

Rather than propose a formula and defend it, the productive move is to write down what any such number ought to satisfy and see what survives. This is the approach of Artzner and coauthors.

Definition 24.1 (Coherent risk measure)**.** ρ is coherent if for all losses L, M and constants $c > 0$:

$$\begin{aligned}
&\text{(monotonicity)} && L \le M \implies \rho(L) \le \rho(M),\\
&\text{(translation invariance)} && \rho(L + c) = \rho(L) + c,\\
&\text{(positive homogeneity)} && \rho(cL) = c\,\rho(L),\\
&\text{(subadditivity)} && \rho(L + M) \le \rho(L) + \rho(M).
\end{aligned}$$

Monotonicity: a position that loses more in every state requires more capital. Hard to argue with.

Translation invariance: adding a certain loss of $\$c$ increases the required capital by exactly $\$c$. This is what makes the number readable as an amount of money — it says that holding $\rho(L)$ in cash against the position brings its risk to zero, since $\rho(L - \rho(L)) = 0$.

Positive homogeneity: doubling the position doubles the risk. This is the most arguable of the four, because in a crisis a position twice as large is more than twice as hard to sell, and a measure honouring that would be superadditive in size. It is retained for tractability.

Subadditivity: a merged portfolio is no riskier than the two apart. This is the important one. It is the statement that diversification cannot hurt, and it is what makes risk numbers add up sensibly across a firm — if it fails, the sum of the desks' risk can be less than the firm's, and a trader can reduce measured risk by moving a position to another book without changing anything real.

24.3 Value at Risk, and What Is Wrong With It

Definition 24.2 (Value at Risk)**.** At confidence α, the Value at Risk is the α-quantile of the loss:

$$\mathrm{VaR}_\alpha(L) = \inf\{x : \mathbb{P}(L \le x) \ge \alpha\}.$$

Read it as: on all but the worst $1 - \alpha$ of days, the loss is no greater than this.

Three of the four axioms hold, and quickly. Monotonicity: if $L \le M$ everywhere then the distribution function of M is everywhere below that of L, so every quantile of M is at least that of L. Translation invariance and positive homogeneity are the corresponding properties of quantiles, which shift and scale with the variable.

Subadditivity fails. The standard reassurance — that it only fails for strange distributions — is not true, and the case where it fails is one banks hold a great deal of.

Example 24.1 (Diversification punished)**.** Take a corporate bond, notional \$100, which defaults over the horizon with probability 4% and is then worth nothing. The loss is \$100 with probability 0.04 and zero otherwise.

At $\alpha = 95\%$: the bond survives with probability 0.96, which is above 95%, so the 95% quantile of the loss is

$$\mathrm{VaR}_{0.95} = \$0.$$

The measure says this position needs no capital at all.

Now diversify. Hold two independent bonds of \$50 each, same issuer quality, same default probability. Then

$$\mathbb{P}(\text{neither defaults}) = 0.96^2 = 0.9216,$$
$$\mathbb{P}(\text{exactly one defaults}) = 2 \times 0.04 \times 0.96 = 0.0768.$$

Now $0.9216 < 0.95$, so the 95% quantile is no longer at zero loss but at the one-default outcome:

$$\mathrm{VaR}_{0.95} = \$50.$$

Splitting one position into two independent halves has taken the measured risk from nothing to half the notional. Writing A and B for the two halves, $\rho(A) = \rho(B) = 0$ but $\rho(A + B) = 50$, so subadditivity fails as badly as it can.

The failure is not a curiosity. It says that a firm managed against Value at Risk is being told that concentration is free and diversification is expensive, and the mechanism is completely general: the measure looks at a single quantile and is blind to everything beyond it. A position whose losses are rare and enormous — selling deep out-of-the-money options, writing insurance against a crisis, holding senior tranches — can be given a Value at Risk of zero, because the loss lives entirely in the tail the measure does not look at.

Remark. Notice which axiom is doing the work in the argument, and which is not. Nothing here depends on the losses being large or the probabilities being unrealistic; it depends only on the loss distribution having a jump near the quantile. Credit portfolios are made of jumps. So are books of digital and barrier options, whose payoffs are discontinuous by construction — and chapter 9 showed that a butterfly is the market's way of pricing exactly such a jump.

24.4 Expected Shortfall

The repair is to stop looking at one quantile and average over all of them beyond it.

Definition 24.3 (Expected shortfall)**.**

$$\mathrm{ES}_\alpha(L) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(L)\,du. \tag{24.1}$$

In words: the average loss over the worst $1-\alpha$ of outcomes. Where Value at Risk asks how bad things get before the tail starts, expected shortfall asks how bad the tail is.

Remark (Not the conditional expectation)**.** It is often written as $\mathbb{E}[L \mid L \geq \mathrm{VaR}_\alpha]$, and for a loss with a continuous distribution the two agree. When the distribution has atoms they do not, and the difference is large.

Take the single bond of Example 24.1. Its 95% Value at Risk is zero, so the event $\{L \geq \mathrm{VaR}_{0.95}\}$ is the whole sample space, and the conditional expectation is the unconditional mean of \$4. Definition (24.1) instead averages the quantiles above 95%, of which the top 4% sit at a loss of \$100 and the rest at zero, giving

$$\mathrm{ES}_{0.95} = \frac{1}{0.05}\big(0.04 \times \$100\big) = \$80.$$

Eighty against four. The conditional form is not merely imprecise here, it is useless — and it is the form that fails to be coherent. When a distribution has jumps, and credit distributions are made of them, (24.1) is the definition.

Theorem 24.4 (Expected shortfall is a worst-case expectation)**.** *Let $\mathcal{Q}$ be the set of measures $\mathbb{Q} \ll \mathbb{P}$ whose density $Z = d\mathbb{Q}/d\mathbb{P}$ satisfies $0 \leq Z \leq \frac{1}{1-\alpha}$. Then*

$$ES_\alpha(L) = \sup_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[L]. \tag{24.2}$$

Proof. Write $q = \mathrm{VaR}_\alpha(L)$, and let Z^* put the maximum allowed weight on the worst outcomes: $Z^* = \frac{1}{1-\alpha}$ where $L > q$, $Z^* = 0$ where $L < q$, and on $\{L = q\}$ — if that event has positive probability, exactly the atom case the remark above works out — whatever value in $[0, \frac{1}{1-\alpha}]$ makes $\mathbb{E}[Z^*] = 1$. By construction Z^* is a valid density and $\mathbb{E}^{Z^*}[L]$ averages L over precisely its worst $1-\alpha$ tail, which is (24.1): $\mathbb{E}^{Z^*}[L] = \mathrm{ES}_\alpha(L)$.

For any other Z with $0 \le Z \le \frac{1}{1-\alpha}$ and $\mathbb{E}[Z] = 1$, the pointwise product $(Z^* - Z)(L - q)$ is non-negative everywhere: where $L > q$, $Z^* = \frac{1}{1-\alpha} \ge Z$, so both factors are ≥ 0; where $L < q$, $Z^* = 0 \le Z$, so both factors are ≤ 0; and at $L = q$ the second factor vanishes. Taking expectations,

$$0 \;\le\; \mathbb{E}\big[(Z^* - Z)(L - q)\big] \;=\; \mathbb{E}[Z^* L] - \mathbb{E}[ZL] - q\big(\mathbb{E}[Z^*] - \mathbb{E}[Z]\big) \;=\; \mathbb{E}[Z^* L] - \mathbb{E}[ZL],$$

the last term vanishing because both are densities of probability measures. So $\mathbb{E}^{\mathbb{Q}}[L] \le \mathbb{E}^{Z^*}[L] = \mathrm{ES}_\alpha(L)$ for every $\mathbb{Q} \in \mathcal{Q}$, with equality at Z^* itself, which is (24.2). □

Theorem 24.5. *Expected shortfall is coherent.*

Sketch. Monotonicity, translation invariance and homogeneity are inherited from the same properties of the quantiles being averaged in (24.1).

Subadditivity is the substantial one, and follows from theorem 24.4: for any fixed $\mathbb{Q}$ expectation is additive, so

$$\mathbb{E}^{\mathbb{Q}}[L + M] = \mathbb{E}^{\mathbb{Q}}[L] + \mathbb{E}^{\mathbb{Q}}[M] \le \sup_{\mathbb{Q}} \mathbb{E}^{\mathbb{Q}}[L] + \sup_{\mathbb{Q}} \mathbb{E}^{\mathbb{Q}}[M],$$

and taking the supremum on the left preserves the inequality. A supremum of linear functionals is convex; that is the whole argument. □

This representation says expected shortfall is the worst average loss across a set of stress scenarios, where the set is everything that does not distort the real world probabilities by more than a factor of $1/(1-\alpha)$. A coherent risk measure and a stress test are the same object, described two ways.

Exercise (Where the jump comes from)**.** In Example 24.1, find the confidence level α at which the single bond's Value at Risk jumps from \$0 to \$100, and the level at which the two-bond portfolio's jumps from \$0 to \$50. Deduce that the counterexample works for every α strictly between 0.9216 and 0.96, and that it disappears outside that window. This is why a measure defined by one quantile can be made to say almost anything by moving the quantile.

24.5 Computing It

Three methods, in increasing order of cost and fidelity. All three need the same input — a distribution of portfolio value changes — and differ in how they get it.

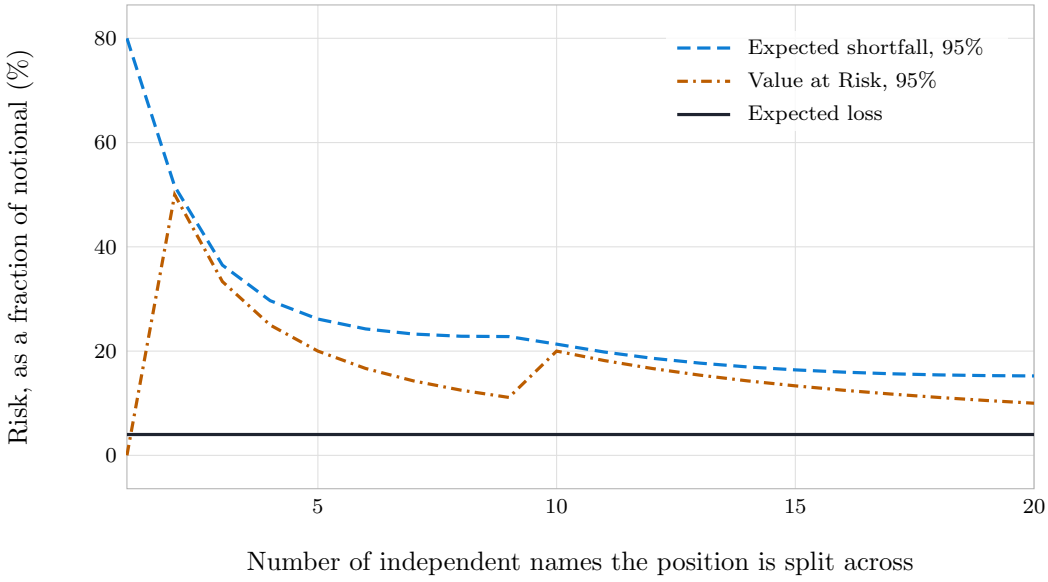

Figure 24.1: The two measures as a fixed notional is split across more independent names, each defaulting with probability 4%. Expected shortfall falls steadily towards the expected loss, which is what diversification ought to look like. Value at Risk starts at zero, jumps to half the notional at the first split, and then declines — and the jump is the failure of subadditivity, visible as the one place where diversifying makes the reported number worse.

Drawn by `Losses::value_at_risk` and `Losses::expected_shortfall`.

Variance-covariance

Assume the portfolio's loss is linear in a vector of risk factors X, and that those factors are jointly normal:

$$L = -\delta^\top \Delta X, \qquad \Delta X \sim N(0, \Sigma).$$

Then L is normal with variance $\delta^\top \Sigma \delta$, and both measures are available in closed form:

$$\mathrm{VaR}_\alpha = z_\alpha \sqrt{\delta^\top \Sigma \delta}, \qquad \mathrm{ES}_\alpha = \frac{n(z_\alpha)}{1-\alpha} \sqrt{\delta^\top \Sigma \delta},$$

with z_α the standard normal quantile and n its density.

The δ here is exactly the vector of sensitivities the pricing chapters produce: the DV01 buckets of chapter 7, the delta and vega of chapters 5 and 10. So the risk system consumes the same Greeks the trading desk hedges with, which is both convenient and the source of its main weakness.

The weakness is that the assumption of linearity is false for anything with optionality, and false in a direction that matters. A hedged option book has $\delta \approx 0$ and is not riskless at all — its risk is second order, in the gamma, and a measure linear in the risk factors reports zero. This is the same blindness as Example 24.1 wearing different clothes: the method cannot see a risk that is quadratic because it only asked about the first derivative.

Historical simulation

Take the last few hundred days of actual factor moves, apply each to today's portfolio, and read the quantile off the resulting distribution of losses. No distributional assumption at all, and the correlations are whatever they actually were.

Its weakness is its sample. A few hundred days contain very few tail events, so the estimate of a 99% quantile rests on a handful of observations, and the estimate of anything beyond it rests on none. And the window has to end somewhere: a crisis drops out of the sample on a particular day, and the reported risk of an unchanged portfolio falls that morning for no reason connected to the world.

Monte Carlo with full revaluation

Simulate factor paths, reprice the whole book on each, build the distribution. This is the only one of the three that is right for a book with optionality, because it reprices rather than approximating, and it is expensive for the same reason — a book of Bermudans repriced under fifty thousand scenarios is fifty thousand backward inductions.

This is where the Markovian state of chapter 12 stops being an aesthetic preference. A model whose state is two numbers can be revalued in a nested simulation; one whose state is the whole forward curve cannot.

24.6 The Coordinates a Risk Report Uses

A book of ten thousand trades is not revalued from scratch against every scenario. It is reduced to a short list of sensitivities, and every number downstream — the value at risk, the limits, the capital — is computed from those. So the sensitivities are the coordinates in which the book exists as far as risk is concerned, and anything they do not capture is invisible to everything built on them.

Definition 24.6 (The first-order risk report)**.** For a book of value V depending on an underlying S, its volatility σ, time t and a rate r, the standard sensitivities are

$$\Delta = \frac{\partial V}{\partial S}, \qquad \Gamma = \frac{\partial^2 V}{\partial S^2}, \qquad \mathcal{V} = \frac{\partial V}{\partial \sigma}, \qquad \Theta = \frac{\partial V}{\partial t}, \qquad \rho_r = \frac{\partial V}{\partial r},$$

known as delta, gamma, vega, theta and rho. In a rates book the first is a vector — one sensitivity per curve point or per factor — and vega is a matrix, indexed by expiry and tenor.

Three observations, each of which the earlier chapters have already established and none of which is a detail.

The list is a *Taylor expansion*, so it is an approximation whose error is the terms omitted. Chapter 5 shows that Θ and Γ are not independent for a hedged book but two readings of one quantity, which is why a gamma limit and a theta budget constrain the same thing.

Every entry is *model-dependent*, including delta. Chapter 10 is a chapter about exactly this: two models agreeing on every price today disagree on Δ, because the total derivative includes vega multiplied by how the smile moves, and that is a modelling statement rather than an observable. So a risk report is not a measurement of the book. It is a measurement of the book according to a model.

And the list is *complete only if the factors are.* A one-factor rates model has one delta, so a curve twist is not represented in its coordinates at all. Chapter 12 measures the pricing cost of that; the next section measures the risk cost, which is larger and easier to miss.

24.7 Explaining the Profit and Loss

The sensitivities make a prediction, and it can be checked. At the end of a day the book has made or lost a definite amount, and the risk report says what it should have been:

$$\Delta P\&L_{\text{predicted}} = \Delta \cdot \delta S + \tfrac{1}{2}\Gamma\,(\delta S)^2 + \mathcal{V} \cdot \delta\sigma + \Theta\,\delta t + \dots \tag{24.3}$$

Comparing that with what happened is *profit and loss explain*, it is run daily on every derivatives desk, and under current bank capital rules a desk whose explain is poor loses the right to use its own model. What is left over after (24.3) is subtracted is the unexplained residual.

The residual deserves care, because it is easy to read as noise and it is not. Three things produce it, and only the first is innocent.

- *Higher-order terms.* Genuine and small, and shrinking as the moves shrink. The cure is more terms.

- *Sensitivities computed badly.* A bumped delta from an unstable numerical scheme, or a stale surface. An operational problem with an operational fix.

- *Risk factors that do not span the moves.* The book was exposed to something the report has no coordinate for. No number of extra Taylor terms helps, because the missing quantity is not a derivative with respect to anything in the list.

The third is model risk appearing as a measurement, and it is why the explain is run at all.

Calculation 24.7 (What a one-factor report cannot see)**.** Take a curve driven by two independent factors, a level and a slope, with sizes measured from a year of Treasury curves rather than assumed.[1] Annualised, the level moves about 184 basis points a year and carries 79% of the variance of daily changes, the slope about 69 and carries 11%, and the curvature about 38 and carries 3%. Let the risk system carry a sensitivity to the level alone.

For a book with exposures a to the level and b to the slope, the unexplained share of profit and loss variance is

$$\frac{b^2\sigma^2_{\text{slope}}}{a^2\sigma^2_{\text{level}} + b^2\sigma^2_{\text{slope}}}, \tag{24.4}$$

which a regression of realised against predicted profit and loss recovers exactly.[2] Evaluating it:

Book	**level exposure**	**unexplained**
outright	1.0	1.2%
duration hedged to a tenth	0.1	93%
exactly level neutral	0	100%

with the slope exposure held at 0.3 for the first row and 1.0 for the others. The middle row is the one to sit with: a book whose duration has been hedged down to a tenth has ninety-three per cent of its profit and loss outside what its risk report can describe.

Structure (Hedging is what breaks the explain)**.** Equation (24.4) says the unexplained share depends on the book's *exposures* and not only on how much of the curve's variance each factor carries, and that has an uncomfortable consequence.

Hedging removes exposure to the factor the model understands. It does not touch the others. So a duration hedge takes the numerator of (24.4) as it finds it and shrinks the denominator, and the explain deteriorates for the same reason the book got safer. A level-neutral curve trade — which is what most relative value positions are — has all of its profit and loss in factors a one-factor report cannot see, however small a share of curve variance those factors carry.

The usual defence fails for exactly this reason. It is true that the slope factor carries an eighth of the level's variance, and tempting to conclude it can be neglected. What matters is the product of a factor's size with the book's exposure to it, and a hedged book has arranged for the other product to be small.

[1] `risk::curve_factors`, run on the history in `public/marketdata`. The first three components are the textbook ones and the code checks that they are: the first loads positively at every tenor, the second changes sign once across the curve and the third twice.

[2] `risk::ExplainTest`, both routes.

This is the same result as §24.8's, reached from the other end. There, hedging more often left model error as a larger share of the risk. Here, hedging more completely leaves unspanned factors as a larger share of the profit and loss. Both say that a better hedged book is not a book with less model risk but a book whose remaining risk is more purely model risk — and both are invisible to a risk measure applied to the model's own coordinates.

Remark (What the explain is therefore for)**.** Not for confirming that the greeks are right, which it does only weakly. If a book is marked and its sensitivities computed with the same model, the explain will look good even when the model is wrong, because both sides of the comparison share the error.

Its use is the opposite: the residual is an estimate of the risk the report has no coordinate for. A desk with a persistent unexplained term has measured a factor it is not carrying, and the useful response is to add the factor — which usually means a richer model — rather than to widen a limit. That is why the regulatory version withdraws model approval instead of demanding more capital: the finding is not that the book is risky but that the risk system is not describing it.

24.8 Attributing the Number

Everything so far computes $\rho(L)$. This section is about the claim made at the start, that the number is not useful until it is split.

Take a short one-year at-the-money call, delta hedged. It is the most thoroughly hedged position in this book: one factor, a liquid hedge, a known model. Whatever risk survives here is a floor for anything more complicated.

Calculation 24.8 (Three sources, on the same paths)**.** Hedge the option twice on *identical* paths — once with the volatility that is actually driving them, once with a volatility four points too high — so that the difference between the two is the mis-specification and nothing else. A four point error on a twenty vol is well inside the range a surface is marked to.[3]

Rebalances	correct model hedging error	wrong model hedging error	model part	ratio
50	0.98	1.12	0.55	1.15
250	0.44	0.71	0.55	1.60
1000	0.22	0.59	0.55	2.64

Read the columns rather than the rows. The first falls as the square root of the rebalancing frequency, which is chapter 4's hedging error behaving as it should. The third does not move at all. And the last therefore grows.

Structure (Hedging harder converts one risk into another)**.** The flat column is the content, and its consequence is not the obvious one.

Discretisation error is a sampling error: it comes from applying a correct hedge ratio at finitely many times, and trading more often reduces it without limit. Model error is a *bias*: it comes from

[3] `risk::hedge_risk_sources`.

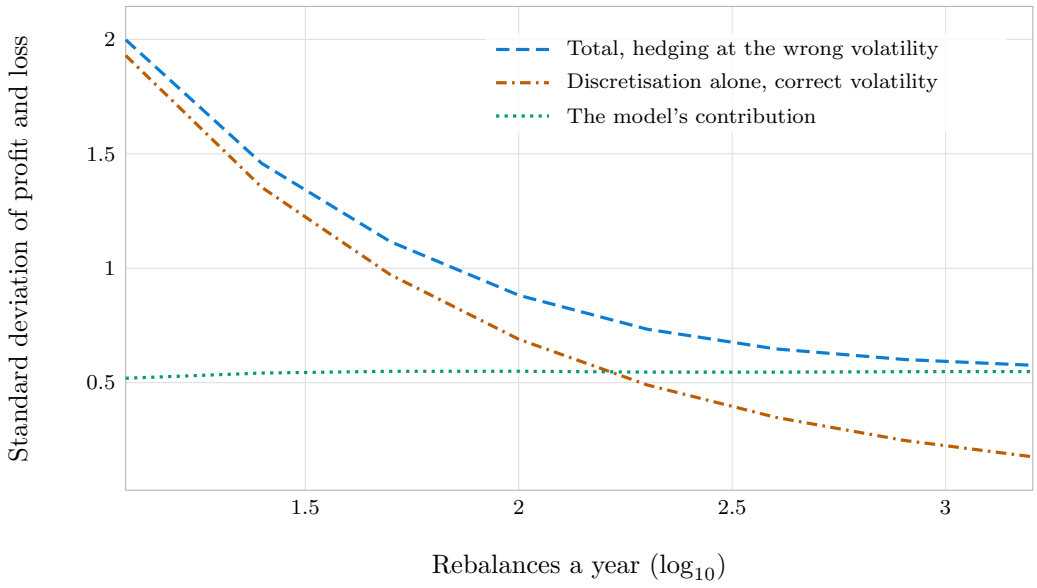

Figure 24.2: The same experiment across a range of rebalancing frequencies, on a logarithmic scale. The discretisation curve falls without limit, as the square root of the frequency; the model's contribution is flat, because it is a bias and not a sampling error. The two cross at about a hundred and sixty rebalances a year — roughly two a week — and past the crossing the total is set by the model rather than by the trading. A desk can move the falling curve by working harder and cannot move the flat one at all.

Drawn by `hedge_risk_sources`.

applying a wrong hedge ratio, and trading more often applies the wrong ratio more often. Nothing about frequency touches it.

So a desk that improves its hedging operation does not reduce its risk proportionally. It converts discretisation risk into a floor, and the floor is set by the model. At fifty rebalances a year the mis-specification is a minor addition to the total; at a thousand it is nearly all of it. The better the operation, the larger the share of what remains that is nobody's decision.

This gives a second and less familiar reason to stop rebalancing. The usual one is transaction costs, which trade off against discretisation error and produce an optimal frequency in the manner of Leland, who showed that replicating with a volatility inflated in proportion to the cost and inversely to the square root of the rebalancing interval reproduces the expected transaction cost exactly, leaving that interval free to be chosen to minimise the total. The reason here is different and applies even with free trading: past the point where model error dominates, additional rebalancing buys a reduction in a term that is no longer the binding one. A desk hedging continuously against a surface it has marked to the nearest vol point is being precise about the smaller half of its problem.

Remark (What a risk report on this book would say)**.** The number a risk system produces for the position above is the first column. It is computed from the book's stated sensitivities, which are the model's sensitivities, so it measures the risk the model believes the book has. The truth is the second column, and at a realistic frequency the report understates it by a factor of 1.6.

The understatement is not a modelling refinement. It is the difference between the risk of the position the desk thinks it holds and the risk of the position it holds, and no choice of ρ repairs it — expected shortfall computed from a wrong delta is a coherent measure of the wrong thing. That is why §24.10 treats model risk as a separate line rather than as an adjustment to this one.

Remark (The three sources across the book)**.** Calculation 24.8 isolates the model term on a position with only one. In a real book each of the three has an identifiable origin, and the point of naming them is that they are reduced by different actions.

Source	**Where it came from**	**How it is reduced**
chosen	a view, sized deliberately	by changing the view
flow	chapter 23's client requests	by quoting differently
unhedged by decision	the partial hedge, cost against variance	by paying the cost
discretisation	finite rebalancing	by trading more often
model	a wrong hedge ratio	not by any of the above

The last row is the one with no remedy in its own column, and there are two ways it arrives. One is a mis-specified parameter, as measured above. The other is chapter 19's: a valuation whose numerical error cannot be bounded, which has to be carried as model risk because a reserve requires a number and none is available. Both end in the same place, which is a provision made by judgement rather than by calculation.

24.9 Backtesting

A risk number is a prediction, and predictions can be scored. If the model is right, a loss exceeding VaR_α should happen on a fraction $1-\alpha$ of days, independently from day to day.

So count. Over n days, the number of exceptions N should be Binomial$(n, 1-\alpha)$, and one can test that directly: at $\alpha = 99\%$ over 250 days the expected count is 2.5, and seeing nine is a one-in-a-thousand event under the null. This is the basis of the traffic-light system regulators use, in which too many exceptions raises a bank's capital multiplier.

Two subtleties matter here.

The first is that independence matters as much as the count. Exceptions arriving in a cluster — four in one week and none for a year — is consistent with the right total and inconsistent with the model, because it says the model is missing the volatility clustering that produced them.

The second is a genuine awkwardness with expected shortfall, and it takes a definition to state. A risk forecast is only useful if it can be graded against what actually happened once the outcome is known, and that property has a name.

Definition 24.9 (Elicitable)**.** A statistic T of a distribution is *elicitable* if there is a scoring function $S(x, y)$ — a penalty for having forecast x and seen y — whose expected value is minimised by the truth:

$$T(F) \;=\; \operatorname*{argmin}_{x} \; \mathbb{E}_{Y\sim F}\big[S(x, Y)\big].$$

The idea is more familiar than the name. Squared error $S(x, y) = (x - y)^2$ is minimised in expectation by the mean, which is why least squares estimates a mean; absolute error is minimised by the median. For a quantile the corresponding penalty is the asymmetric one

$$S_\alpha(x, y) \;=\; \big(\mathbf{1}\{y \le x\} - \alpha\big)\,(x - y), \tag{24.5}$$

which charges α per unit of shortfall when the outcome lands above the forecast and $1-\alpha$ per unit when below, and is minimised exactly at the α quantile. So Value at Risk is elicitable, with (24.5) the score.

Remark (Why that matters)**.** Two things follow from possessing such a score, and both are practical.

Forecasts become comparable. Each day a desk states its number, the outcome arrives, and (24.5) converts the pair into a penalty; averaging over a year ranks two models, or two banks, on a single figure. Without a scoring function "their risk numbers were better than ours" is not a well posed comparison.

And honesty becomes optimal. Because the truth minimises the expected score, a forecaster who believes one number and reports another expects to be penalised for it. A regulator scoring submissions this way does not have to ask anyone to be candid.

Expected shortfall has no such function, and not for lack of trying: the obstruction is structural. An elicitable statistic must satisfy a consistency condition on mixtures — if two distributions agree on the statistic's value, so must every weighted average of them — because a forecast minimising the expected penalty against each distribution separately automatically minimises it against any mixture of the two. Value at Risk satisfies this: if F and G both put mass α below x, so does every mixture of them. Expected shortfall does not: it depends on *where* the quantile sits as well as on the mass beyond it, and mixing moves the quantile, so two distributions can share a shortfall value while their mixture does not.

What is available instead is that the *pair* is jointly elicitable: there is a scoring function of a Value at Risk forecast, an expected shortfall forecast and the outcome, minimised at the true pair together. So the measure can be backtested, and Basel's move to expected shortfall is not unbacktestable in practice. What is unavailable is attributing the score: a joint penalty says the pair was good or bad and does not decompose into a verdict on the tail average alone.

Structure (Two desiderata, pulling opposite ways)**.** This could not have been fixed by a cleverer choice between the two, because the two properties are about different things.

Coherence, in §24.2, is a property of the measure *as a summary of risk*: it asks whether the number behaves sensibly when portfolios are combined, and subadditivity is the part Value at Risk fails. Elicitability is a property of the measure *as a forecast*: it asks whether a prediction of the number can be scored against what happened, and it is the part expected shortfall fails.

Nothing connects them, and the two measures land on opposite sides of both. Value at Risk can be backtested and cannot be aggregated; expected shortfall can be aggregated and cannot be backtested on its own. The regulatory move from one to the other bought the diversification property at the cost of the single-measure backtest, and calling it a strict improvement misdescribes it. It was a choice about which of two failures a capital regime could better live with — and, given that the alternative was a measure telling banks that splitting a position reduces its risk, a defensible one.

Remark (A measure exists with both properties, and it is not used)**.** The tradeoff above is a fact about Value at Risk and expected shortfall specifically, not a theorem that no measure can have both properties. The τ-expectile — defined by an asymmetric squared-error score in place of (24.5)'s asymmetric absolute one — is both coherent and elicitable for $\tau \geq \frac{1}{2}$, and is in fact the only law-invariant risk measure with both properties.[4]

It has not displaced either in practice, for a reason that is about interpretation rather than mathematics. Value at Risk names a probability and expected shortfall names a tail average; both are statements about a specific set of bad scenarios, which is what a risk committee wants pointed at. An expectile is defined by a fixed-point balance between gains and losses across the *whole* distribution, has no such scenario reading, and cannot be described as "the loss if things go badly" to anyone who has not already seen the definition. Basel's 2012 consultation considered it for exactly this reason and rejected it, choosing the measure examiners could explain over the one with the better mathematics.

24.10 Model Risk, and Turning It Into a Number

Four of the modelling chapters each ended in the same place. Chapter 10 found that the smile does not determine β, and β determines the hedge. Chapter 11 found that the vanilla surface does not determine the mixing weight, and the mixing weight determines the exotics. Chapter 12 found that European swaptions do not determine the mean reversion, and the mean reversion determines what a Bermudan is worth. Chapter 13 found that a hundred or so liquid swaptions do not determine a

[4] Bellini, F., Klar, B., Müller, A., & Gianin, E. R. (2014). Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, 54, 41–48.

correlation matrix with hundreds of entries, and the correlation determines every product written on more than one rate.

In each case the calibration succeeded perfectly and left the answer undetermined. That is model risk in its purest form, and it is not addressed by calibrating harder.

What is done instead is to price the book across the range of the undetermined parameter and reserve the difference.

1. Fix a plausible range for the parameter — from history, from a related market, or by policy.
2. Revalue the portfolio at each end.
3. Take the difference as a reserve, held against the marked value.

This is what prudent valuation adjustments are, and the discipline it imposes is worth more than the number: it forces someone to write down what the range is and to defend it, which turns an invisible assumption into a line item that a committee can argue about.

Remark. The reserve should be computed at the level of the book, not the trade. Model risk does not add across trades — it can cancel: two trades whose exposure to β offsets have no joint model risk even though each has plenty on its own — so computing it trade by trade and summing overstates the truth, incoherent in exactly the way Value at Risk is (Definition 24.1).

24.11 Counterparty Risk

Finally, the risk that the trade was right and the money did not arrive.

The exposure at a future date is what is owed to us, which is the value of the trade if positive and nothing if negative — so it is a claim with an option's payoff on our own portfolio, $\max(V_t, 0)$, though nobody chooses when it is exercised: default is a stopping time set by the counterparty, not by us. Averaging over scenarios gives the expected exposure $\mathrm{EE}(t)$, and the credit valuation adjustment is that exposure weighted by the probability of default and the loss given it:

$$\mathrm{CVA} = \mathrm{LGD} \int_0^T \mathrm{EE}(t)\, d\,\mathbb{P}(\tau \le t).$$

Two observations connect this back.

The first is that CVA has volatility risk. A desk that hedges it holds vega against its own counterparties' exposure profiles, which is why counterparty risk ended up staffed by derivatives quants rather than by credit analysts.

The second is that this is where chapter 7's discussion of collateral pays off. A fully collateralised trade has an exposure that is reset to zero continuously, so the integral above nearly vanishes and the adjustment is small. That is the sense in which the collateral agreement was doing real work: it converts a credit problem into a funding problem, and chapter 7 showed that the funding problem is solved by discounting on the collateral rate. Uncollateralised, the credit problem stays, and its price is this integral.

24.12 One Model or Many

A real book is priced by several models at once. Chapter 12's Bermudans sit in one, caps and floors in another, a CMS in the static replication of chapter 15 with no dynamic model at all, a callable bond behind the option-adjusted spread of chapter 7. Each was chosen because it fits the instruments that product is hedged in. The question this raises is whether the book as a whole is then incoherent, and whether pricing everything in one rich model would fix it.

The answer has two halves, and they point in opposite directions.

What multiple models genuinely break

Three things, and only the third is usually anticipated.

Internal arbitrage. Two models calibrated to the same vanillas disagree about an exotic — that is the whole content of §24.10. If two desks price adjacent products with different models, there is a trade between them that both books mark as profitable. Nothing has been created; the firm has written down the difference between two models twice, with opposite sign, and called it revenue. This is not hypothetical and it is the strongest practical argument for a single valuation model within any set of products that trade against each other.

Joint distributions rather than marginals. Chapter 18 is the general statement: agreeing on every marginal constrains the joint distribution not at all. Two models that price every instrument in the book correctly on their own can still disagree completely about what happens to the book, because a portfolio's value depends on the joint law of its inputs. A netted position that one model says is flat is not flat under the other.

The explain report absorbs the inconsistency. §24.7 decomposes the day's profit into factor moves and a residual. With one model the residual is model error plus hedging error. With several, it also contains the disagreement between them, and those three are no longer separable — which removes the diagnostic that the report exists to provide.

What does not break, and why

First-order risk aggregation.

The reason is Definition 24.6's observation read forwards. A risk system does not store models or parameters; it stores sensitivities to *traded instruments* — bucketed curve deltas against the calibrating swaps, a vega grid indexed by expiry and tenor, sensitivities to the basis quotes of chapter 7. Those are derivatives of value with respect to observable prices, and derivatives with respect to the same quantity add up regardless of what computed them. A Bermudan's vega to the ten year ten year swaption and a cap's vega to the same instrument net exactly, whatever models produced them, because both are answers to the same question about the same price.

So the common language is the market, not the model. That is the same principle as key rate durations being reported against the instruments rather than against the curve's internal parametrisation, and it is why a multi-model book is not the operational disaster it sounds like.

Where a single model is genuinely required

Two calculations cannot be assembled this way, and they are the reason global models were built.

The first is portfolio-level scenario analysis. Revaluing a book under a joint move of the curve and the volatility surface requires a consistent way of moving every input together, which is a joint distribution over inputs — and a joint distribution over inputs is a model. Sensitivities cannot supply it, because the whole point of a stress is that it is large enough for the Taylor expansion to fail.

The second is counterparty exposure. The expected exposure of the previous section is

$$\mathrm{EE}(t) = \mathbb{E}\big[\max(V_t, 0)\big],$$

where V_t is the value of the *whole netting set* at a future date. The maximum does not distribute across trades, so this cannot be computed trade by trade and summed; every trade in the set must be simulated forward under one measure, on one set of paths. That is a hard requirement rather than a preference, and it is what forced the industry to build global models after 2008 — not a desire for elegance, but the fact that a netted exposure is not a sum of exposures.

Structure (The two requirements want different models)**.** Put the two halves together and the apparent conflict dissolves, because the calculations that need local accuracy and the calculations that need global consistency are not the same calculations.

Pricing and marking need to be exactly right on the instruments the product is hedged in, and need to be right about that product's own second-order behaviour. They tolerate having nothing to say about the rest of the book, because the hedge is executed in that product's own market.

Exposure, capital and stress need one consistent measure across everything and tolerate being coarse. An expected exposure profile is an average of an average; a stress is a scenario chosen by judgement to within a wide band. Neither is improved much by a model that prices each trade to the last basis point, and chapter 19's argument applies with force — a global model rich enough to price everything well has more parameters than the market identifies, so its extra richness buys consistency of an arbitrary choice rather than accuracy.

So the architecture that desks actually run — a model per product for pricing, and one coarser model for the whole netting set — is not a compromise between two ideals. It is the correct answer, because the two jobs have different requirements and no single model is best at both. The failure mode is not having two systems; it is using each for the other's job: marking a book off the exposure model, or computing a netted exposure by adding per-trade sensitivities.

Remark (Consistency is not correctness)**.** One argument for the single global model deserves to be answered directly, because it is the one usually made and it is weaker than it sounds.

A single model does give consistent hedges. What it does not give is correct ones. If the model is wrong — and §24.10 is a list of parameters the market does not determine — then every hedge in the book is wrong in the same direction, and the book has no way of finding out. Two models disagreeing is unpleasant, but the disagreement is a *measurement*: it is the model reserve of §24.10, arrived at for free by having built the second model. Unifying on one model does not remove that risk. It removes the instrument that was detecting it.

There is a second, quieter cost. A model calibrated to every market at once fits none of them exactly, so the misfit is spread across all products rather than concentrated where it can be reserved against. The Bermudan desk's model that fits swaptions exactly has a known, bounded error — it prices its hedges right and its exotic wrong by an amount §24.10 can bound. The global model prices everything slightly wrong, including the hedging instruments, and the error in the hedge is the one that costs money continuously rather than once.

The reasonable position is therefore narrower than either slogan. Unify within a set of products that trade against each other, because internal arbitrage is real. Do not unify across sets that do not, because the consistency bought is between quantities nobody trades, and it is paid for with accuracy on quantities everybody does.

References

- Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3), 203–228.

- Acerbi, C., & Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7), 1487–1503.

- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.

- Fissler, T., & Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4), 1680–1707.

- Bellini, F., Klar, B., Müller, A., & Gianin, E. R. (2014). Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, 54, 41–48.

- Gregory, J. (2020). *The xVA Challenge*. Wiley.

- Leland, H. E. (1985). Option pricing and replication with transactions costs. *The Journal of Finance*, 40(5), 1283–1301.

Chapter 25

Solvability, Realism and What Is Left

The preceding chapters were written to be read in order and the argument they make is distributed across them, so this one assembles it. The claim is that the models in this book were chosen under a constraint that no longer binds, that the constraint which replaced it is not compute but identification, and that what solvability still buys is not speed but a bounded error — which makes the choice between a tractable model and a realistic one a trade of realism for auditability rather than a matter of taste. Along the way the same result appears from three directions: hedging a book better does not reduce its model risk, it concentrates it.

25.1 What the Book Did

The spine of these notes is two reductions, and chapter 14 names them.

The first turns a functional integral into a differential equation, by insisting that the state be finite. The second turns that equation into a system of ordinary ones, by insisting that the generator preserve a small family of functions. Neither is a fact about markets. Both are conditions imposed so that a calculation can be finished, and almost every model in this book is what one of them left standing.

Be blunt about how much was given up, because the chapters measured it rather than asserting it.

Restriction	**Imposed for**	**What it cost, measured**
one factor	a small Markov state	a 2s10s spread option at a quarter of its value
exponential volatility	a finite realisation at all	every non-quasi-exponential maturity profile
local volatility	an exact fit to vanillas	a forward skew a fifth of today's after a year
affine dynamics	a transform in closed form	a smile shape the market does not have at short expiry

Each row was established where it belongs — the first in chapter 12, the second there too, the third in chapter 10, the fourth in chapter 10's closing section — and each cost is large enough that nobody would accept it if the alternative were available.

25.2 The Constraint That Replaced Compute

The alternative is available. Chapter 19 sets out what a machine will now do, and the honest summary is that the first reduction is no longer needed for pricing and the second is no longer needed at all. A fifty-factor model can be simulated; a pricing equation in fifty dimensions can be solved to useful accuracy without a grid.

So the question is what binds instead, and the book answers it in three places without ever collecting the answer.

- Chapter 10 shows that a smile does not determine a model's dynamics: two models agreeing on every quoted price disagree on the hedge.
- Chapter 20 measures a SABR calibration in which the quotes fix the level of the smile to one per cent and leave the correlation free across most of its range, with the model exactly right and the data noiseless.
- Chapter 12 finds the same for mean reversion, and chapter 11 for the mixing weight.
- Chapter 21 shows the same holds even where compute is not the constraint at all: a neural network trained to approximate a pricing map reproduces the flatness of an unidentified parameter faithfully, at speed — machine learning changes what is computationally feasible, not what is identifiable.

These are not separate difficulties. They are one, and it has a name.

Structure (Identification is the binding constraint)**.** A model is a map from parameters to prices. Calibration inverts it. The question of whether that inverse exists and is stable is not about optimisers or compute; it is about whether the instruments quoted carry enough information to distinguish the model from its neighbours, and generally they do not.

The situation is therefore the opposite of the one the classical models were built for. Then, a model that could not be solved could not be used, so tractability bound and data was ample relative to what a tractable model needed. Now a model can be solved at almost any complexity, and the data available to fit it has not grown at all — the same forty swaption quotes, the same handful of liquid expiries. A model with a million parameters fitted to forty quotes is the identification problem in its extreme form: a richer model does not extract information the market did not supply, it distributes the same information over more unknowns.

Chapter 20 makes the sharpest version of this quantitative, and it repays repeating because it is easy to file as a technicality. On a spread that genuinely mean reverts with a one-year half-life, five years of daily data gives a ninety per cent interval for the half-life spanning a factor of eight, and sampling ten times as often does not narrow it. The constraint is not the number of observations. It is the length of the record, and a desk cannot buy more history.

Chapter 22's convergence trade is this constraint held as a position rather than measured in the abstract: the risk is not the direction, which may be certain, but the holding period, priced off exactly the half-life uncertainty just measured.

25.3 What Solvability Still Buys

If tractability is no longer necessary, the question is whether it is still worth anything, and the answer is yes for two reasons that have nothing to do with speed. Chapter 19 establishes both.

Prices that are consistent structurally rather than statistically. A closed form cannot produce a negative butterfly, because it is an expectation under a measure and an expectation of a convex function is convex. Once the expectation is estimated the guarantee must be earned: priced on common paths a butterfly inherits the payoff's convexity term by term and cannot be negative even at ten thousand paths, and priced on independent paths over a quarter of narrow triples invert. Arbitrage-freeness moves from a property of the model to a property of the estimator, and survives on three disciplines rather than on a theorem.

An error one can write down. A grid's error is $O(h^p)$ and two runs estimate it; a Monte Carlo standard error is a genuine confidence statement. A network fitted to a residual offers neither, because the map from residual to error runs through a stability constant unavailable in exactly the dimensions the method exists to handle. Over four dimensions the network is the more accurate method. It is not the more *measurable* one.

Structure (The trade is realism for auditability)**.** Putting those together gives the decision this book is ultimately about.

A solvable model is less realistic. That is not incidental — chapter 14 shows it was chosen for tractability, and the table above measures what the choice cost. A numerical model is more realistic and its error is not a quantity. So the choice is not between a good model and a bad one but between two different things one might want to be able to say about a number.

The reason this matters and is not merely tidy is chapter 24's: a valuation uncertainty has to be reserved against, and a reserve requires a number. An unbounded error cannot be reserved for. It can only be provisioned against by judgement, which is to say carried as model risk — so a more realistic model with an unmeasurable error has not removed a risk, it has moved it from the valuation to the capital, where it is less visible and harder to argue about.

Stated that way the trade becomes a decision with an owner rather than a habit with a history. A desk pricing a vanilla book wants the auditable answer. A desk pricing a fifty-factor structured note has no auditable answer available and should say so, rather than reporting a number with the same confidence as the vanilla one.

25.4 One Result, From Three Directions

The practical chapters were written independently and arrive at the same conclusion, which is the strongest evidence in the book that it is not an artefact of any one setting.

- Chapter 23 ends on the partial hedge, whose unhedged fraction is the transaction cost over the risk penalty. It is zero only if hedging is free, so a completely hedged book is not a well-run one.

- Chapter 24's attribution: rebalancing more often drives discretisation error down as the square root of the frequency and leaves model error untouched, so the share of remaining risk that is model error rises — from 1.15 times the correct-model risk at fifty rebalances a year to 2.64 at a thousand.

- Chapter 24's explain: hedging removes exposure to the factor the risk model understands and leaves the others alone, so the unexplained share of profit and loss rises with the quality of the hedge — under two per cent for an outright book, over ninety for the same book duration hedged, and all of it for a level-neutral curve trade.

Structure (Hedging concentrates model risk)**.** The three are one statement. Hedging removes the risks the model can see. It cannot remove the risks the model cannot see, because a hedge is computed from the model. So what survives hedging is, disproportionately, whatever the model is wrong about — and the better the hedging, the purer that residue.

This has a consequence for how a risk number should be read that no amount of care with ρ supplies. A well hedged book reports a small risk, and reports it in the model's own coordinates, so the report is small for the same reason it is unreliable. The unhedged book's number is larger and more nearly true.

It also gives a limit to how much hedging is worth doing, distinct from the transaction cost argument. Past the point where model error dominates, additional rebalancing reduces a term that is no longer binding: a desk hedging continuously against a surface marked to the nearest volatility point is being precise about the smaller half of its problem. The binding term is reduced by a better model, not by more trading, and knowing which of the two one is short of is the practical content of chapter 24's insistence on decomposition.

25.5 What To Do About It

Three habits, each of which the chapters arrived at by measuring something and finding it was not what was expected.

Decompose before reporting. A risk number without an attribution cannot be acted on, because the three sources — chosen, flowed, and imposed by the model — are reduced by three different actions and only one of them is free. Chapter 24 makes this its thesis and chapter 23 supplies the book that requires it.

Distinguish what cannot be hedged from what is not worth hedging. Both appear in a risk report as a residual and they are not the same thing: the first is a modelling constraint and no amount of spending reduces it, the second is a decision with a price and can be revisited. Chapter 23's projection makes the distinction precise.

Prefer a structural reason to a fitted parameter. Chapter 22 needs both a reason and a horizon, and chapter 20 shows that the horizon is the part the data cannot supply. A claim about a

regulation or an index rule is a claim the evidence can support; a half-life from five years of history is a point estimate from a sample that could not distinguish six months from two years.

Remark (What will date and what will not)**.** The methods in chapter 19 will be superseded, probably soon, and the models in the middle chapters are already less realistic than what a machine can now handle. What will not date is the structure of the argument: that every tractable model is a restriction adopted for a reason, that the reason matters because it tells one where the model will fail, and that the failure is measurable before it is expensive.

That is why these notes spend as much effort on what the models get wrong as on what they compute. A reader who leaves knowing which compromise sits under each formula, and how to measure what it cost, is equipped for models that have not been invented yet. A reader who leaves knowing the formulae is equipped for the ones in here.

Acknowledgements

My interest in and understanding of quantitative finance has been shaped over the years by brilliant colleagues and mentors: Antoine Savine, Hamza Hoummady, Anas Benabid, Valeri Rousskikh, Jérôme Lebuchoux, Arvind Giridhar, Dunstan Marris, Elie Franko, Tingfai Ng, Jiri Hoogland, George Lowther, and many more.

Further Reading

- Hull, John C. *Options, Futures, and Other Derivatives*, EBook, Global Edition. Pearson, 2021.

- Shreve, Steven E. *Stochastic Calculus for Finance II: Continuous-Time Models*. Vol. 11. New York: Springer, 2004.

- Karatzas, Ioannis, & Shreve, Steven E. *Brownian Motion and Stochastic Calculus*. 2nd ed. Springer, 1991.

- Bouchaud, Jean-Philippe, & Potters, Marc. *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*. 2nd ed. Cambridge University Press, 2003.

- Henry-Labordère, Pierre. *Analysis, Geometry, and Modeling in Finance: Advanced Methods in Option Pricing*. CRC Press, 2008.