LUNA LABS

# Financial Markets

A First-Principles Course

THE TERRAIN VOLUME



Prepared for Riddhish • the SST research programme

MMXXVI

# About This Volume

The Course Book teaches the mathematics of prediction; the Operational Playbook scripts the procedures. This volume teaches the terrain both of them stand on: what a market actually is, who is on the other side of a trade, where prices, spreads, and financing charges come from, and how the machinery fails. It exists because a model of prices that does not understand its marketplace is a map drawn by someone who has never been outdoors.

The book serves four pillars in the order the platform's versioned roadmap attacks them: foreign exchange CFDs first and deepest (v1, the live system), options (v2), equities and index products (v3's habitat), and commodities (v4) — with crypto covered for the two 24/7 instruments the platform already carries.

**How facts are handled.** Market STRUCTURE — who trades, how clearing works, what a rollover is — is stated as fact. Market FIGURES — spreads, rate levels, leverage caps, correlations — drift, and every such number in this book is labelled indicative, with the standing instruction to verify current specifications before relying on one. Worked examples use hypothetical scenarios, every number computed by a verification script kept beside the manuscript. Where a practice is controversial (last look, B-book warehousing), the book reports the documented concern and no more.

**Reading machinery.** Definition, Result, Key idea, Trap, and Your project boxes work exactly as in the Course Book. Gold chips such as [Course Book §22.4] point to the chapter of the Course Book where the corresponding mathematics is built — this volume states formulas once and never re-derives them. Exercises come with ruled working space and page-broken solutions: attempt first, turn after.

**The route.** Part I builds the anatomy every market shares — venues, participants, the order book, market making, and the cost of immediacy. Part II is the FX pillar: the dealer network, macro drivers, the retail CFD dealing model, and positioning. Part III covers equities and indices; Part IV, futures, listed options, and commodities. Part V assembles the system view: rates and macro transmission, intermarket dynamics and regimes, six case studies in market failure, and the crypto appendix-chapter. The back matter carries the conventions tables, a glossary, annotated references, and the platform's instrument-by-instrument market map.

# Contents


**I What a Market Is** **1**

**1 Why Markets Exist, and Who Is In Them** **2**
1.1 What a Market Does . . . 2
1.2 The Participant Map . . . 4
1.3 Why Anyone Trades With You . . . 9
1.4 Order Types and the Trading Day . . . 13
1.5 Price Formation in One Page . . . 17

**2 The Order Book** **21**
2.1 Anatomy of a Limit Order Book . . . 21
2.2 Walking the Book . . . 24
2.3 Where the Spread Comes From . . . 27
2.4 Market Making as a Business . . . 32
2.5 Modern Structure: Fragmentation, Internalisation, and Dealer Streams . . . 35
2.6 Auctions: Opening, Closing, and Uncrossing . . . 37
2.7 Exercises . . . 40

**3 Transaction Costs, Measured Honestly** **52**
3.1 The Cost Taxonomy . . . 52
3.2 Quoted, Effective, and Realised Spread . . . 55
3.3 Price Impact and Implementation Shortfall . . . 58
3.4 Costs at Retail CFD Scale . . . 62
3.5 Financing and Carry Costs . . . 64
3.6 Measuring Your Own Costs . . . 67
3.7 Exercises . . . 69

**II The Foreign Exchange Market** **81**

**4 The Foreign Exchange Market** **82**
4.1 The shape of the market . . . 82
4.2 Instruments . . . 83
4.3 Quoting conventions . . . 85
4.4 The trading week and the liquidity clock . . . 88
4.5 Who moves the price . . . 92
4.6 Retail FX and where OANDA sits . . . 93
4.7 Exercises . . . 94

**5 The Economics of Exchange Rates** **104**
5.1 Covered interest parity . . . 104
5.2 Uncovered interest parity and the carry trade . . . 108
5.3 Long-run anchors: PPP, real exchange rates, and the balance of payments . . . 112

5.4 Central banks: the driver that dominates the platform's horizons . . . 114
5.5 What moves what, at which horizon . . . 117
5.6 Exercises . . . 120

**6 FX CFDs and the Retail Dealing Model 133**
6.1 The counterparty reality . . . 133
6.2 A-book, B-book, and the hybrid . . . 135
6.3 The dealing engine . . . 138
6.4 Where the spread comes from . . . 141
6.5 Financing: swap points and the rollover seam . . . 145
6.6 Margin, leverage, and dying well . . . 148
6.7 What this means for the platform . . . 152
6.8 Exercises . . . 153

**7 Positioning, Flow, and Sentiment Data 166**
7.1 The Commitments of Traders report . . . 166
7.1.1 Which markets have a COT report, and the futures-as-proxy problem . . . 168
7.2 Reading positioning: net, z-scores, and the crowding narrative . . . 169
7.2.1 Extremes, and two ways to read them . . . 170
7.2.2 Managed money against commercial: the divergence reading . . . 170
7.3 Order flow: what actually moves price . . . 171
7.4 Risk-on/risk-off as a market regime . . . 172
7.5 Using positioning data honestly . . . 174
7.6 Exercises . . . 175

**III Equities and Indices 184**

**8 Stock Markets 185**
8.1 What a Share Is and How It Lists . . . 185
8.2 The Venues . . . 186
8.3 Indices . . . 187
8.4 Corporate Actions . . . 191
8.5 The Information Cycle . . . 193
8.6 Sectors and Factors . . . 195
8.7 Exercises . . . 196

**9 Index CFDs and Trading Single Names 207**
9.1 The Index Exposure Chain . . . 207
9.2 Dividends in Index CFDs . . . 209
9.3 Sessions and Gaps . . . 211
9.4 Single Names for Pairs . . . 212
9.5 Costs at the Index/Equity Layer . . . 214
9.6 Exercises . . . 216

**IV Derivatives and Commodities 225**

**10 Futures Markets 226**
10.1 The Contract . . . 226
10.2 Margin and Mark-to-Market . . . 229
10.3 Volume, Open Interest, and Who Is in the Market . . . 231
10.4 The Curve . . . 232

10.5 The Roll . . . 234
10.6 Basis and Convergence . . . 237
10.7 Exercises . . . 239

**11 Options Markets in Practice 250**
11.1 The Listed Landscape . . . 250
11.2 How Options Are Quoted, and What Actually Trades . . . 252
11.3 The Market Makers and Their Hedging Flows . . . 255
11.4 The Surface as a Market Object . . . 258
11.5 Costs and Liquidity . . . 262
11.6 The Volatility Trader's Workflow . . . 264
11.7 Exercises . . . 267

**12 Commodities 281**
12.1 The Physical Layer . . . 281
12.2 Storage and the Curve . . . 283
12.3 Seasonality, Where It Is Real . . . 284
12.4 Gold, Silver, and the Platform's Metals . . . 286
12.5 Crude Oil Markets . . . 288
12.6 Commodity Returns, Honestly . . . 290
12.7 Exercises . . . 292

**V The System View 302**

**13 Rates, Money, and Macro Transmission 303**
13.1 Policy rates and the short end . . . 303
13.2 The yield curve . . . 304
13.3 Duration in one section . . . 307
13.4 The macro calendar . . . 309
13.5 Transmission to the four pillars . . . 312
13.6 Inflation and real rates . . . 314
13.7 Exercises . . . 315

**14 Intermarket Dynamics and Regimes 325**
14.1 The correlation map . . . 325
14.2 The dollar at the centre . . . 330
14.3 Risk-on/risk-off, mechanically . . . 331
14.4 Correlation in stress . . . 335
14.5 Carry unwinds as regime events . . . 339
14.6 Leads, lags, and honest testing . . . 340
14.7 Why regimes exist . . . 342
14.8 Exercises . . . 345

**15 Crises: Six Case Studies in Market Failure 358**
15.1 October 1987: the feedback machine . . . 358
15.2 LTCM 1998: leverage against convergence . . . 360
15.3 2007–2008: the funding run . . . 363
15.4 6 May 2010: the liquidity vacuum . . . 365
15.5 15 January 2015: the peg that broke . . . 367
15.6 March 2020 and the dash for cash . . . 372
15.7 The failure-mode taxonomy . . . 374
15.8 Exercises . . . 376

**16 Crypto Markets** **387**
16.1 What Trades . . . 387
16.2 The 24/7 Clock . . . 389
16.3 Spot, Perpetuals, and the Funding Rate . . . 390
16.4 Volatility and Correlation Character . . . 392
16.5 Market Quality, Honestly . . . 393
16.6 Exercises . . . 394

**A Market Conventions: A Quick Reference** **402**
A.1 Pip sizes and pip values . . . 402
A.2 The session clock . . . 403
A.3 FX value dates and rollover . . . 404
A.4 The COT calendar . . . 404
A.5 Contract structure: futures, options, and CFDs . . . 405
A.6 Index-CFD dividend adjustment . . . 406
A.7 Crypto funding-rate mechanics . . . 406

**B Glossary** **408**

**C Annotated References** **413**

**D The Platform's Market Map** **417**

PART I

# What a Market Is



LUNA LABS

# 1

# Why Markets Exist, and Who Is In Them

*This chapter asks the question every later chapter takes for granted: why does a market exist at all, and who, precisely, is standing on the other side of every order your system sends? It derives a market's three economic jobs — discovering a price, moving risk to whoever will bear it most cheaply, and supplying the readiness to trade that a lone buyer and a lone seller could never arrange for themselves — and shows how those jobs are split, structurally, between exchanges and dealer networks, with this book's four pillars falling cleanly on one side or the other. It then walks the cast of participants who actually populate that structure, ending precisely where this book's reader sits: a systematic CFD account whose every fill is a decision by OANDA, not by "the market", that the reader's price was a good one to accept. That last fact reframes the entire trading problem. Whether a strategy has an edge is, underneath the statistics, a question about who is paying for its winning trades and why they keep doing it — the question this chapter converts from a slogan into arithmetic, and the one Chapter 2's order book and later chapters' cost surfaces spend the rest of this book answering properly.*

## 1.1 What a Market Does

Two people who want to trade with each other, and who already know it, do not need a market: they need a phone call. A market exists because, in general, neither condition holds. The buyer of a currency, a share, or a barrel of oil does not know in advance who wants to sell it, at what price, or whether that seller can be trusted to deliver; a would-be seller has the mirror problem. A market is the institutional answer to that double search problem, and it does three distinct jobs at once — jobs worth separating now, because later chapters diagnose both failures and profits by asking which one is being paid for.

**Price discovery.** Every participant holds some private read on value — a corporate treasurer's view of a currency's fair value drawn from trade flows nobody else sees, a macro fund's view drawn from a rate forecast, a retail trader's view drawn from a chart pattern. None of these views is directly observable to anyone else. What a market does is force every participant who wants to act on a view to express it as an order, and then aggregate those orders, mechanically, into a single public number: the price. That number does not report any one participant's belief; it reports the price at which the marginal willing buyer and the marginal willing seller currently agree, given everything anyone has been willing to put money behind. Section 1.5 makes this aggregation mechanism precise; here it is enough to note that the resulting number is a public good — once discovered, it is visible for free to every onlooker, including participants who paid nothing towards discovering it.

**Risk transfer.** A wheat farmer who has planted a crop is exposed, involuntarily, to the price of wheat at harvest; a bakery that has committed to a fixed price for bread is exposed, involuntarily, to the same price in the other direction. Neither wants the exposure a market lets them lay off — at a cost. A hedger who sells that risk to a market and a speculator who buys it are not doing the same trade for the same reason: the hedger is paying, in expectation, to convert an uncertain outcome into a known one, and the party absorbing the risk is charging a premium for

bearing what the hedger would not. This is the oldest justification for organised futures trading (Keynes, 1930), developed in full in this book's later derivatives and commodities chapters; the general principle — a market lets a risk seek out whoever is best placed, or best paid, to hold it — applies equally to an airline's fuel hedge, a fund's currency overlay, and, in a more contested form (Section 1.3), to a retail CFD position.

**Liquidity provision.** Even with a price freely visible and a willing risk-bearer somewhere in the economy, a specific buyer and a specific seller must still arrive *at the same venue at the same moment* for a trade to occur. Almost nobody's buying and selling needs actually coincide in time. A market solves this timing mismatch by concentrating trading interest — either in a shared order book that patient participants leave resting, or in a dealer who quotes continuously and stands ready to trade with whoever turns up next — so that a buyer arriving on a Tuesday afternoon does not need to find a seller who happens to want to sell at that exact hour. The service of bridging that timing gap is not free: it is compensated by the bid-ask spread, and it is undertaken, deliberately, only by participants equipped to manage the risk of holding one side of a trade before the other side turns up (Section 1.2).

### How prices actually get made: auction, dealer, or both

Markets differ enormously in how they perform these three jobs, along two largely independent dimensions. The first is the *matching mechanism*:

- In an **auction (order-driven) market**, participants submit orders into a shared book, and a matching algorithm — almost always price, then time, priority — pairs buy and sell orders directly against each other. No party stands between the two sides of a trade; each is the other's direct counterparty.
- In a **dealer (quote-driven) market**, one or more dealers post a bid and an ask at which they will trade; a client transacts against the dealer's quote, not against another client's order. The dealer is the direct counterparty to every trade and carries the resulting position until it can be laid off.
- Most real venues are **hybrids**: an exchange with a central limit order book may also license designated market makers obliged to keep quoting through thin periods; an OTC dealer market may run an electronic request-for-quote layer in which several dealers compete, auction-like, for a single client order (Harris, 2003).

The second dimension is *governance*: is the venue a centralised, rule-bound exchange, or a decentralised network of bilateral relationships?

> **Definition — Exchange versus over-the-counter (OTC)**
>
> An **exchange** is a centralised venue trading standardised contracts under a common rulebook, with publicly disseminated prices and, typically, a central counterparty that interposes itself after a trade is agreed so that neither original party bears the other's default risk. An **OTC market** has no such centre: contracts are negotiated bilaterally — directly, or through a dealer — terms can be customised, there is no single public order book, and, unless the trade is subsequently novated to a central counterparty (increasingly common for standardised OTC instruments, but not universal), each party bears its direct counterparty's credit risk for the life of the contract (Duffie, 2012).

Governance and matching mechanism are correlated but not identical: exchanges are usually, but not always, auction markets, and OTC markets are usually, but not always, dealer markets, which is why the two dimensions are kept separate here rather than treated as a single dichotomy.

### Where this book's four pillars live

Foreign exchange spot trading is the cleanest OTC dealer market among this book's four pillars: there is no FX exchange, no single order book, and no central counterparty for a spot trade between a bank and its client. Instead, a tiered network of dealers — mostly banks — quotes prices to each other and to clients, with the tightest, deepest pricing concentrated among a small number of the largest dealers (Lyons, 2001; Rime, Schrimpf, 2013). Measured by the Bank for International Settlements' triennial survey, global FX turnover averaged on the order of $7.5 trillion a day in April 2022 (BIS, 2022; indicative, subject to revision at the next survey) — and essentially none of that volume crosses a public order book of the kind a listed share does.

Listed equities, listed equity and index options, and listed futures, by contrast, trade on exchanges: standardised contracts, a public order book (Section 1.4), and a central counterparty that clears every trade. This is true even though, as later chapters detail, a single listed stock can now trade across several competing venues at once — some publicly transparent exchanges, others operating with reduced pre-trade transparency — which complicates what "the price" means (Section 1.5) without changing the basic governance point: the contract a listed-equity or listed-derivative trader holds is issued and cleared by a centralised, regulated venue, not negotiated bilaterally with a single counterparty.

CFDs sit in a third position, and it is the one that matters most directly for this book's reader: a CFD is, in every case, a bilateral OTC contract between the client and the broker, regardless of what it references. An equity CFD references an exchange-formed price (the underlying share's exchange quote); an FX CFD references an OTC-formed price (the underlying spot rate, itself dealer-quoted); a commodity CFD typically references an exchange-traded futures price. The CFD contract itself never touches the venue that formed the reference price — it is a side bet, settled in cash against that reference, between exactly two parties. Section 1.2's project box makes this concrete for the reader's own instruments.

## 1.2 The Participant Map

Section 1.1's three jobs are not performed by an abstraction; they are performed by identifiable classes of participant, each showing up in a market for a different, usually declared, reason. The oldest classification splits every trade by economic motive into three groups (Hull, 2021):

- **Hedgers** hold, or are about to hold, an exposure that arises from their underlying business — a currency receivable, an inventory of a physical commodity, a portfolio's existing market exposure — and trade to reduce it, generally accepting a worse expected price in exchange for lower variance.
- **Speculators** take on a position *because* of a view about future prices, with no offsetting exposure elsewhere in their business; they exist to bear risk hedgers do not want, in exchange for the compensation the market offers for bearing it.
- **Arbitrageurs** trade to capture a price discrepancy between related instruments that should, by a no-arbitrage argument, be priced consistently, taking on as little net market risk as the discrepancy allows.

This trichotomy classifies a *trade's motive*, not an institution: the same desk can hedge one exposure, speculate on another, and arbitrage a third within the same hour. It also does not obviously cover everyone who shows up in a modern market, a gap the next list closes.

### The modern cast

**Dealers and market makers.**

**Definition — Market maker**

A **market maker** continuously quotes both a bid and an ask for an instrument and stands ready to trade in either direction at those prices. Its business is not forecasting the instrument's direction; it is managing a fluctuating inventory — built up automatically as clients trade against its quotes — profitably enough that the spread earned across many trades exceeds the cost of inventory risk and the losses taken on the informed minority of that flow (Section 1.3). A **dealer** in an OTC market performs the identical function without necessarily carrying a formal market-making obligation to any exchange; the OANDA desk pricing this book's reader's CFD account is a dealer in exactly this sense.

This is not one of the three motives above: a market maker is not hedging an outside exposure, is not purely speculating (its quotes are two-sided and its intended holding period is short, by design), and is not arbitraging a mispricing between two instruments. It is selling a service — continuous availability to trade — and it is compensated for that service, not for a view.

**Electronic market makers and high-frequency trading (HFT).** The market-making function above is increasingly performed not by a human trader working a book but by automated systems posting and cancelling quotes across many venues and instruments at very high frequency, typically holding close to zero inventory overnight and relying on high turnover and low per-trade risk rather than a large position (Menkveld, 2013). The underlying economics are the market maker's: capture the spread, manage the inventory, survive the adverse-selection minority of the flow. What changes is speed and scale, not the function.

**Asset managers.** Pension funds, mutual funds, and sovereign wealth funds trade primarily as a by-product of an asset-allocation mandate — rebalancing a portfolio back to target weights after prices move, accommodating client inflows and redemptions — rather than because of a trade-specific view. Their trading is real order flow that a market maker must absorb and price, but it does not fit the hedger/speculator/arbitrageur trichotomy cleanly either: the mandate, not a market view, drives the order.

**Hedge funds.** A loose umbrella for pooled vehicles pursuing an absolute-return mandate, spanning several distinct styles: *global macro* (directional bets on rates, currencies, and indices driven by a top-down economic view); *trend-following / managed futures* (systematic positions in the direction of an instrument's recent trend, across many markets at once); *relative-value / statistical arbitrage* (positions sized to be close to market-neutral, profiting as a spread between related instruments converges); *long/short equity* (paired long and short stock positions expressing a view on one name relative to another, or relative to the market); and *event-driven*, including merger arbitrage (positions taken around a specific corporate event, priced with real numbers in Exercise 1.1). Most of these are speculators or arbitrageurs by motive; which one depends on the style.

**Corporates.** Non-financial firms trade to manage exposures generated by their actual business — an exporter selling foreign-currency receivables forward, an airline hedging fuel, a multinational's treasury managing working capital across currencies. This is the hedger role in its purest form.

**Central banks.** Central banks transact in FX and government debt markets for reasons that are not economic-profit motives at all: managing reserves, implementing monetary policy, and, occasionally, intervening directly in the FX market to influence the exchange rate itself when it is judged to be moving in a way that harms the domestic economy (Sarno, Taylor, 2001). A central bank's trade can look, mechanically, like a very large speculative or hedging trade, but its objective is a policy outcome, not a return on capital.

**Retail.** Individual traders, participating either directly on an exchange for listed instruments (via a broker), or, for this book's reader's own instruments, entirely off-exchange through a CFD provider.

**Your project — Where the SST platform's orders actually go**

Every order the platform sends is a CFD order, and a CFD is — by Section 1.1's definition — a bilateral OTC contract. Concretely: an order placed against any of the platform's 21 OANDA instruments (14 FX pairs, XAU/XAG, BTC/ETH, two indices, WTI) is filled by OANDA, not by an exchange, an ECN, or another trader. OANDA is the direct legal counterparty to every position the platform holds, and it prices every fill from its own dealing engine, itself most likely priced off a basket of upstream liquidity providers it does not disclose trade by trade. No order ticket the platform generates ever reaches an exchange's or an ECN's matching engine directly. This is not a peculiarity of one broker; it is what "CFD" structurally means (Section 1.1). Its consequence is the entire subject of Section 1.3: OANDA's willingness to keep quoting the platform a price is itself an economic decision, made fill by fill, about whether the platform's flow is worth taking the other side of.

**Exercise 1.1 — Classifying the cast**

For each of the eight hypothetical participants described below, state (i) which of hedger/speculator/arbitrageur/market-maker best describes the *trade* described — noting where none fits cleanly, and why — and (ii) the single best-matching category from this section's "modern cast" list.

(1) An airline buys US-dollar forward contracts to lock in the price, in its home currency, of jet fuel it will purchase in six months.

(2) A fund buys the shares of a takeover target at $38.80 immediately after a rival announces a firm, fully-financed all-cash offer of $40.00 a share, planning to hold until the deal closes or collapses.

(3) A bank desk continuously quotes two-way EUR/USD prices to corporate clients throughout the trading day, ending most days close to flat but never holding a position it did not first acquire from a client trade.

(4) A proprietary trading firm posts and cancels two-sided quotes on a listed futures contract thousands of times a second across several venues, and is rarely still holding a position an hour later.

(5) A pension fund sells part of its equity holdings and buys bonds each quarter, purely to bring its portfolio back to a 60/40 target weighting after equities have outperformed.

(6) A fund is long every G10 currency that has appreciated against the US dollar over the past twelve months and short every one that has depreciated, rebalanced monthly.

(7) A central bank sells its own currency and buys US dollars to slow an appreciation judged to be hurting export competitiveness.

(8) An individual trades a systematic FX strategy through a CFD account, with every order filled by the broker itself.

Then compute the following, using hypothetical but exactly specified figures. (a) The bank desk in (3) completes 500 round trips a day, averaging 200,000 EUR/USD notional per round trip, and captures an average of 0.4 pips of spread, net of its own hedging, on each one. Compute its gross daily spread revenue. (b) The fund in (2) buys 10,000 shares at $38.80. If the deal closes at the announced $40.00, compute its dollar profit, its percentage return over the holding period, and that return expressed as a simple annual rate assuming the deal takes four months to close.

▷ working space

**Solution — Exercise 1.1**

**Classification.**

| # | Trade fits | Modern-cast category | Why |
|---|---|---|---|
| 1 | Hedger | Corporate | exposure reduction |
| 2 | Arbitrageur (loosely) | Hedge fund — event-driven | real deal risk |
| 3 | None cleanly | Dealer / market maker | sells liquidity, not a view |
| 4 | None cleanly | HFT / electronic MM | same function, faster |
| 5 | None cleanly | Asset manager | mandate-driven |
| 6 | Speculator | Hedge fund — trend-following | momentum bet |
| 7 | None cleanly | Central bank | policy objective |
| 8 | Speculator | Retail | directional view, own capital |

Three of the eight (3, 4, 5) and one more (7) are exactly the gap this section flagged when it introduced the modern cast: the hedger/speculator/arbitrageur trichotomy classifies a trade's economic motive, and a market maker's motive is neither hedging an outside exposure, nor speculating on direction, nor arbitraging a mispricing between two instruments — it is selling continuous availability to trade, compensated by the spread, which is why (3) and (4) are the same economic animal at different speeds (Menkveld, 2013). The pension fund in (5) is trading for a third reason again: not a market view of any kind, but a mandate that fires mechanically when weights drift, which is exactly why its flow is comparatively easy for a market maker to identify and price as (relatively) uninformed — a point Section 1.3 returns to. The central bank in (7) is not economically motivated in the sense the trichotomy assumes at all: its objective is a policy outcome (Sarno, Taylor, 2001), and its trade should not be read as encoding a forecasting edge the way a hedge fund's would.

Item (2) deserves a specific correction to the label "arbitrageur." Textbook arbitrage is close to riskless: a covered-interest-parity violation (this book's foreign exchange chapters derive the no-arbitrage condition) or a put-call-parity violation (derived in this book's later derivatives chapters) can be locked in with simultaneous trades that leave no open exposure. Merger arbitrage is not that: the fund in (2) is exposed, for the life of the deal, to the real probability that it breaks — on financing, competition clearance, or a shareholder vote — and the \$1.20 gap between the \$38.80 market price and the \$40.00 offer is compensation for exactly that risk, not a free lunch. This is why the strategy is more accurately called *risk* arbitrage; "arbitrageur" is the closest of the three trichotomy labels, but only loosely.

**(a) Dealer spread revenue.** At 200,000 notional, one pip is worth $200{,}000 \times 0.0001 = \$20.00$; capturing 0.4000 pips per round trip earns $0.4000 \times \$20.00 = \$8.00$ per round trip, and 500 round trips a day give gross daily spread revenue of

$$500 \times \$8.00 = \$4{,}000.00.$$

This is gross of the desk's own hedging costs, any residual inventory losses, and the adverse-selection losses Section 1.3 quantifies — it is revenue, not profit.

**(b) Merger-arbitrage return.** The deal spread is $\$40.00 - \$38.80 = \$1.20$ a share; on 10,000 shares, the profit if the deal closes as announced is $10{,}000 \times \$1.20 = \$12{,}000.00$. As a fraction of the \$38.80 committed per share, the return is

$$\$1.20/\$38.80 = 3.0928\%$$

over the holding period; expressed as a simple annual rate over a four-month holding period, $3.0928\% \times (12/4) = 9.2784\%$. That annualised figure is the fund's compensation for four months of genuine deal risk, not a riskless rate — precisely the correction just made to the classification.

## 1.3 Why Anyone Trades With You

Every participant just catalogued who quotes a price — a dealer, an HFT market maker, a resting limit order left by anyone else — is, in effect, offering to trade with you. Before asking why a market exists in the abstract, it is worth asking the sharper, more uncomfortable question this book's reader actually needs answered: why would any of them want to trade with *you*, specifically, at the price they are quoting? The honest answer starts with what that willingness costs you, in money, on every single trade, illustrated next with a hypothetical but exactly computed scenario, not a reported market quote.

**Worked Example 1.1 — A round trip through a dealer market**

A dealer quotes EUR/USD at a bid of 1.09538 and an ask of 1.09554. A trader buys 100,000 units at the ask and, with no time elapsed and no change in the quote, immediately sells the same 100,000 units back at the bid.

The spread is $1.09554 - 1.09538 = 0.00016$, or — dividing by the pip size of 0.0001 for a four-decimal pair ([Course Book §1.1]) — 1.6000 pips. Buying costs $100{,}000 \times 1.09554 = \$109{,}554.00$; selling returns $100{,}000 \times 1.09538 = \$109{,}538.00$. The round trip loses

$$\$109{,}554.00 - \$109{,}538.00 = \$16.00.$$

Checking the same number the other way: one pip on 100,000 units of a USD-quoted pair is worth $100{,}000 \times 0.0001 = \$10.00$, so 1.6000 pips cost $1.6000 \times \$10.00 = \$16.00$ — the two routes to the answer agree exactly, as they must, since the pip-value shortcut is nothing but the same multiplication factored differently. Relative to the \$109,546.00 average of the two cash amounts, that is

$$\$16.00 / \$109{,}546.00 \times 10{,}000 = 1.4606 \text{ basis points of notional.}$$

For a pair trading close to 1.0955, a pip and a basis point of price very nearly coincide ([Course Book §1.1]), which is why FX cost figures are quoted in pips throughout this book while equity costs are more often quoted in basis points.

Nothing about this trade involved a price move, a forecast, or a mistake: the loss is the fee for having found a counterparty instantly, in full size, with certainty of execution. That fee is the subject of the rest of this section.

Why is that fee not zero? A dealer who quoted a bid and ask exactly equal to its own best estimate of fair value would, on the arithmetic just shown, earn nothing from a round trip — and would, in expectation, lose money on every trade, for a reason that has nothing to do with operating costs.

Model the instrument's near-term fair value as a random variable $V$ with unconditional mean $\mathbb{E}[V] = m$ — the dealer's best guess absent any order. A fraction $\alpha > 0$ of the traders who approach the dealer are *informed*: each already knows the realisation of $V$, and buys only if $V$ exceeds the dealer's ask $A$, sells only if $V$ is below the dealer's bid $B$, and otherwise stays away. The remaining $1 - \alpha$ are *uninformed*, trading for reasons — rebalancing, hedging, noise — unrelated to $V$, and split roughly evenly between buying and selling regardless of where $V$ actually sits.

Now ask what the dealer should infer from the mere fact that a buy order has arrived. Some buy orders come from informed traders who know $V > A$; the rest come from uninformed traders who would have bought regardless of $V$. Conditioning on either type of buy leaves the uninformed contribution centred on the unconditional mean $m$, but the informed contribution is centred on $\mathbb{E}[V \mid V > A]$, which is strictly above $m$ whenever there is any probability that $V \le A$.

A buy order is therefore always at least weakly, and with $\alpha > 0$ strictly, *good news* about $V$:

$$\mathbb{E}[V \mid \text{buy order arrives}] > m = \mathbb{E}[V].$$

For the dealer to avoid losing money on its ask on average, the ask must be set no lower than what it expects to give up once it accounts for exactly who chooses to buy at that price, i.e. $A \geq \mathbb{E}[V \mid \text{buy at } A]$; combined with the inequality above, this forces $A > m$. The symmetric argument on the sell side forces $B < m$. A spread $A - B > 0$ therefore exists *even for a dealer with no operating costs and no risk aversion whatsoever*, purely because order flow is informative: the very act of being willing to trade with the dealer, at the dealer's own quoted price, is evidence about which side of fair value the price sits on (Glosten, Milgrom, 1985). Kyle's (1985) $\lambda$, developed fully in Chapter 2, is the same idea recast as a continuous price-impact-per-unit-of-order-flow coefficient rather than a two-sided quote.

**Definition — Adverse selection**

**Adverse selection**, in a trading context, is the risk a liquidity provider bears that a counterparty who chooses to trade against its quote possesses superior information about the instrument's near-term value, so that — averaged only over the trades that are actually accepted, not over all possible trades — the price moves against the liquidity provider after the fact. It is distinct from *inventory risk* (the risk of holding an unwanted position while waiting to lay it off) and from *order-processing cost* (the pure operating cost of running the desk); Chapter 2 separates all three formally. The spread derived above compensates for adverse selection alone, and is the minimum spread a rational, riskless, costless dealer would still have to charge.

**Result — The breakeven move equals the spread**

Suppose a position is opened by buying at the ask $A_0 = m_0 + s/2$ and later closed by selling at the bid $B_1 = m_1 - s/2$, where $m_0, m_1$ are the mid price at entry and exit and the quoted spread $s$ is unchanged between the two. The trade's profit is

$$B_1 - A_0 = (m_1 - m_0) - s,$$

which is zero exactly when the mid has moved in the trader's favour by $m_1 - m_0 = s$: the full spread, not half of it. A directional trade therefore only breaks even once the underlying has moved far enough to repay *both* the entry cost and the exit cost the round trip incurs ([Course Book §1.6]), which is why a wider spread does not merely shave a little off every trade's profit — it raises the minimum move a position needs before it was worth having taken at all.

None of this says a dealer, or anyone else quoting you a price, is being unfair. The spread just derived is compensation for a real risk, and every fill you receive has one of exactly three honest explanations for why your counterparty was willing to give it to you:

1. **Inventory rebalancing.** Your counterparty needed to move a position it already had, for reasons unrelated to a view on you — a market maker unwinding an accumulated long, an index fund forced to trade by a rebalancing date. Your order happened to be convenient for them.
2. **Risk transfer at a premium.** Your counterparty is a hedger, or a natural risk-bearer, paying (or receiving) a premium to convert an uncertain exposure into a known one, exactly as Section 1.1 described. They are not trying to be right about the price; they are trying to be certain about it.
3. **You are slower, or less informed.** The uncomfortable one: your counterparty has better information, a faster reaction to the same information, or a structural read on order flow

(Section 1.2's dealers and HFT market makers) that lets them price your order correctly and you incorrectly, and they are capturing exactly the edge Glosten and Milgrom's argument assumed some fraction of arriving flow would supply.

For any single fill, you cannot generally tell which of the three applies — but a systematic strategy generating thousands of fills does not have that excuse. Its aggregate P&L, net of the spread just derived, *is* the answer to which reason predominantly describes it.

**Trap — "Someone must be on the other side" is not a truism you can skip past**

Apply this to a systematic FX book rather than a single trade. Every one of its fills was accepted by OANDA, or by whichever liquidity provider OANDA itself trades with, for one of exactly the three reasons above. Absent specific evidence otherwise, a dealer facing a stream of retail systematic orders has no reliable way to distinguish reasons (1) and (2) from reason (3) except by watching what happens to the price *after* each fill — which is precisely how a dealer would go about measuring, and then pricing in, how adversely selected a given client's flow is. The platform's own *cost surface* (the Playbook's Chapter 2) — the measured spread, slippage, and financing actually paid, across instrument, time of day, and volatility state ([Course Book §14.2]) — is not a neutral tax charged equally to everybody. It is, to a first approximation, the market's running estimate of how much of reason (3) your flow looks like, and it will not fall by finding a cleverer execution algorithm alone: it falls only if the strategy's own fills stop looking, statistically, like the informed side of this section's derivation. That is exactly the question Stage 0 ([Course Book §14.7]) is designed to answer honestly, net of this very cost, before any capital is risked on the belief that the answer is already known.

**Exercise 1.2 — A different round trip, and the move it would take to profit**

A dealer quotes GBP/USD at a bid of 1.26514 and an ask of 1.26538 — a second hypothetical scenario, computed exactly.

(a) A trader buys 250,000 units at the ask and immediately sells them back at the bid, with the quote unchanged. Compute the round-trip cost in pips and in dollars, by both routes shown in Worked Example 1.1, and confirm they agree.

(b) Instead, the trader holds the position. Using Result 1.3, compute the new mid price at which closing the position — at the then-current bid, with the spread unchanged — exactly breaks even.

(c) Compute the mid price at which closing the position would instead leave a net profit of exactly $100.

▷ working space

**Solution — Exercise 1.2**

**(a)** The spread is $1.26538 - 1.26514 = 0.00024$, i.e. 2.4000 pips. Buying costs $250{,}000 \times 1.26538 = \$316{,}345.00$; selling returns $250{,}000 \times 1.26514 = \$316{,}285.00$; the round trip loses $\$316{,}345.00 - \$316{,}285.00 = \$60.00$. By the pip-value route: one pip on 250,000 units is worth $250{,}000 \times 0.0001 = \$25.00$, and $2.4000 \times \$25.00 = \$60.00$ — the two agree exactly, as Worked Example 1.1 showed they must.
**(b)** The entry mid is $m_0 = (1.26538 + 1.26514)/2 = 1.26526$. By Result 1.3, the mid must rise by the full spread, $s = 0.00024$, i.e. to

$$m_1 = 1.26526 + 0.00024 = 1.26550,$$

a rise of exactly 2.4000 pips — the same number as the round-trip cost in (a), as it must be: breaking even means recovering exactly what the round trip cost.
**(c)** A net profit of \$100 requires $\$100/\$25.00 = 4.0000$ additional pips beyond breakeven, i.e. a total mid move of $2.4000 + 4.0000 = 6.4000$ pips, taking the mid to $1.26526 + 0.00064 = 1.26590$. Check: at that mid, the exit bid is $1.26590 - 0.00012 = 1.26578$, and the trade nets $250{,}000 \times (1.26578 - 1.26538) = 250{,}000 \times 0.00040 = \$100.00$, confirming the figure.

## 1.4 Order Types and the Trading Day

However a price is formed, a participant reaches it by submitting one of a small number of standard order types (Harris, 2003).

- A **market order** demands immediate execution at the best price currently available, guaranteeing execution but not price; against a thin book (Chapter 2) it can walk through several price levels and fill at a materially worse average price than the best quote.
- A **limit order** specifies the worst price the submitter will accept, guaranteeing price (or better) but not execution — it may sit unfilled, in whole or in part, if the market never trades through it.
- A **stop order** sits inactive until a specified trigger price trades, at which point it converts into a market order (or, for a *stop-limit*, a limit order); it is used both defensively, to cap a loss on an open position, and offensively, to enter on a breakout.

An order also carries a **time-in-force** instruction governing how long it remains live: common conventions include *day* (expires unfilled at the day's close), *good-till-cancelled* (GTC, stays live until explicitly withdrawn), *immediate-or-cancel* (IOC, fills whatever quantity is immediately available and cancels the rest), and *fill-or-kill* (FOK, executes in full immediately or not at all).

### How an auction actually clears

**Result — Call-auction clearing**

Collect a book of buy orders $\{(L_b^i, q_b^i)\}$ and sell orders $\{(L_s^j, q_s^j)\}$, where $L$ is a limit price and $q$ a quantity. At any candidate price $p$, define cumulative demand and supply

$$D(p) = \sum_i q_b^i \, \mathbf{1}\{L_b^i \geq p\}, \qquad S(p) = \sum_j q_s^j \, \mathbf{1}\{L_s^j \leq p\},$$

the total quantity willing to buy, respectively sell, at $p$ or better. The volume that can actually trade at $p$ is $V(p) = \min\{D(p), S(p)\}$: no more than the shorter side of the book can be filled. An opening or closing **call auction** sets its clearing price $p^*$ at the price maximising $V(p)$, since any price with $D(p) \neq S(p)$ leaves a side of the book with unmatched, but

crossable, interest that a small price adjustment could still capture. If several prices tie for the maximum — which happens whenever no order sits strictly between them, so that neither $D$ nor $S$ changes across the tied range — venues resolve the tie with a secondary rule, typically minimising the remaining imbalance $|D(p) - S(p)|$ and, if that is still tied, moving to the price closest to a reference (commonly the previous close) (Harris, 2003).

**Worked Example 1.2 — Clearing a toy call auction**

A hypothetical opening auction receives exactly six orders:

| Order | Side | Limit price and quantity |
|---|---|---|
| B1 | Buy | 500 units at \$50.20 |
| B2 | Buy | 300 units at \$50.05 |
| B3 | Buy | 400 units at \$49.90 |
| S1 | Sell | 200 units at \$49.80 |
| S2 | Sell | 350 units at \$50.05 |
| S3 | Sell | 250 units at \$50.15 |

Evaluating $D(p)$ and $S(p)$ at each of the five distinct submitted prices:

| Price $p$ | $D(p)$ | $S(p)$ | $V(p) = \min(D, S)$ |
|---|---|---|---|
| \$49.80 | 1,200 | 200 | 200 |
| \$49.90 | 1,200 | 200 | 200 |
| \$50.05 | 800 | 550 | **550** |
| \$50.15 | 500 | 800 | 500 |
| \$50.20 | 500 | 800 | 500 |

At $p =$ \$50.05: $D(50.05) =$ B1+B2 $= 500 + 300 = 800$ (B3's \$49.90 limit no longer clears a \$50.05 price), and $S(50.05) =$ S1+S2 $= 200 + 350 = 550$ (S3's \$50.15 limit is not yet reached). The matched volume is $\min(800, 550) = 550$, strictly the largest of the five candidates, so by Result 1.4 the auction clears *uniquely* at $p^* =$ \$50.05 for 550 units. No tie-break is needed here because a buy order (B2) and a sell order (S2) happen to sit at the identical price \$50.05: that is exactly what breaks the tie mechanism described in Result 1.4 — without an order sitting at exactly this price on both sides, the maximum would instead have been shared across a small range of prices, and the secondary rule would have been needed to pick one. At the clearing price, demand (800) exceeds supply (550): B1 and B2 together want 800 units, but only 550 are on offer, so the buy side is rationed — 250 units of buy interest go unfilled, with the specific allocation between B1 and B2 set by the venue's own priority rule (commonly price, then time), not by this calculation.

### Sessions and the liquidity clock

Equity, listed-option, and futures exchanges bound each trading day with exactly the mechanism just derived: an **opening auction** establishes the first trade of the day from overnight-accumulated orders, continuous trading runs through the session, and a **closing auction** concentrates a large share of the day's volume into a single clearing price used for end-of-day valuation and index tracking (Harris, 2003). Exact session times are set by each exchange and shift with local daylight-saving transitions; treat any specific open or close time quoted anywhere as indicative and confirm the current schedule before relying on it.

Foreign exchange has no such single session at all. By long-standing market convention — not statute — FX is treated as trading continuously from the Monday morning open in Wellington

through to the New York afternoon close on Friday, closed over the weekend; the trading day itself is conventionally deemed to roll over at 5 p.m. New York time, the point at which value dates change and financing is assessed on open positions. Both the weekly boundaries and the rollover convention are broker- and venue-dependent in their precise administrative treatment and shift with US, UK, and Australian daylight-saving transitions at different times of year, so treat the times above as indicative and confirm current specifications for any live system; Chapter 4 returns to this session structure in full, including the rollover mechanics and the platform's own session-bucketed cost surface.

Liquidity is not uniform across that week. It concentrates where the largest number of active participants overlap — most obviously the London/New York overlap in the afternoon UK time, morning US time — and thins out where few major centres are open at once, most notably around the 5 p.m. New York rollover and through the early Asian session. The platform's own cost-surface work (the Playbook, Chapter 2) buckets spreads exactly this way and finds its tightest measured spreads in the London-New York overlap bucket, precisely because that is where $D(p)$ and $S(p)$ in Result 1.4's language are both largest at once, and its widest spreads in the thin buckets around rollover and the Asian session — consistent with the platform canon that spreads widen at rollover and in thin hours.

**Exercise 1.3 — A different auction book**

A second hypothetical opening auction receives:

| Order | Side | Limit price and quantity |
|---|---|---|
| B1 | Buy | 300 units at \$75.40 |
| B2 | Buy | 200 units at \$75.10 |
| B3 | Buy | 450 units at \$74.85 |
| S1 | Sell | 150 units at \$74.70 |
| S2 | Sell | 400 units at \$75.10 |
| S3 | Sell | 350 units at \$75.30 |

Tabulate $D(p)$, $S(p)$, and $V(p)$ at each of the five distinct submitted prices, identify the clearing price and volume, and state which side of the book is rationed at that price and by how many units.

▷ working space

**Solution — Exercise 1.3**

| Price $p$ | $D(p)$ | $S(p)$ | $V(p) = \min(D, S)$ |
|---|---|---|---|
| \$74.70 | 950 | 150 | 150 |
| \$74.85 | 950 | 150 | 150 |
| \$75.10 | 500 | 550 | **500** |
| \$75.30 | 300 | 900 | 300 |
| \$75.40 | 300 | 900 | 300 |

At $p = \$75.10$: $D(75.10) = \text{B1+B2} = 300 + 200 = 500$ (B3's \$74.85 limit does not clear a \$75.10 price), and $S(75.10) = \text{S1+S2} = 150 + 400 = 550$ (S3's \$75.30 limit is not yet reached). The matched volume is $\min(500, 550) = 500$, strictly the largest of the five candidates (against 150 below and 300 above), so the auction clears uniquely at $p^* = \$75.10$ for 500 units, again because a buy order (B2) and a sell order (S2) coincide exactly at \$75.10. Here it is supply that exceeds demand at the clearing price (550 against 500): the sell side is rationed, with 50 units of sell interest — split between S1 and S2 by the venue's own priority rule — left unfilled, the mirror image of Worked Example 1.2's buy-side rationing.

## 1.5 Price Formation in One Page

Assemble the last four sections into a single causal chain. **Information** arrives — a data release, a corporate announcement, a shift in aggregate risk appetite, or simply the accumulated private views of Section 1.2's participants. Each participant translates whatever information it holds into an **order**, filtered through its own motive (hedge, speculate, arbitrage, or make a market) and its own capacity to bear the resulting risk. Those orders meet the prevailing **quotes** — a dealer's bid and ask, or the resting limit orders in an auction book — and where an incoming order is marketable against those quotes, a **trade** results; where it is not, it simply updates the best quote for the next arrival. The sequence of trade prices and updated quotes *is* what is meant by "the price" (Hasbrouck, 2007), and it immediately becomes new information for the next participant in line, closing the loop. Empirically, in FX, this loop is dominated by *order flow*: the net signed volume of buyer- versus seller-initiated trades forecasts subsequent price changes precisely because that flow reveals dispersed private information the price has not yet incorporated (Evans, Lyons, 2002) — order flow is not noise to be averaged away, it is Section 1.3's adverse-selection logic operating continuously rather than trade by trade.

**Efficiency, recapped.** [Course Book §1.2] derived the price at time $t$ as the mean-squared-error-optimal forecast of the price at $t + 1$ given everything observable up to $t$, and argued that any function of that information set which still predicted the sign or size of the next move would be traded on until the trading itself removed the predictability (Fama, 1970). This section's chain is the microstructure mechanism by which that argument is actually enforced, moment to moment, rather than merely assumed: it is precisely because information gets converted into orders, and orders get executed against quotes, that a predictable price change cannot simply sit unexploited — someone in Section 1.2's cast, most likely a speculator or an arbitrageur, has a direct financial incentive to trade on it, and their trading is the channel through which the price moves to remove it. Market efficiency, in other words, is not a property markets have instead of microstructure; it is a property microstructure produces, to the extent that the participants and mechanisms of this chapter actually do their jobs.

**What "the price" means when there are many venues.** Section 1.1 already noted that a listed share can trade on several competing venues at once; the same fragmentation, in a more extreme form, is FX's permanent condition, since there was never a single venue to fragment away from. Two distinct conventions have grown up to answer "what is the price" under

fragmentation. In FX, because no exchange exists to publish an official price, data vendors, brokers, and benchmark administrators each compute their own *composite* or *consolidated mid* — typically a volume- or contributor-weighted blend across a panel of dealer quotes, refreshed continuously or, for benchmark purposes, sampled at fixed reference times. In US-listed equities, by contrast, regulation requires the continuous computation of a **National Best Bid and Offer (NBBO)**: the single best bid and single best ask currently available across every competing lit venue trading that stock, aggregated by rule rather than by any one vendor's discretion. NBBO is a US regulatory label specifically (a Regulation NMS concept), described here structurally rather than asserted for any other jurisdiction; other regimes solve the same fragmentation problem — publish a usable reference price despite multiple competing venues — with their own rules and institutions, without necessarily using this name or this exact mechanism. Either way, the lesson is the same: once more than one venue can trade the same instrument, "the price" stops being a single fact waiting to be observed and becomes, itself, a constructed statistic, built the same way the clearing price of Result 1.4 was — from the same underlying orders, quotes, and trades, just aggregated by a different rule.

**Exercise 1.4 — Whose money are you taking?**

A candidate systematic FX rule is backtested and reports the following illustrative statistics, not the platform's actual measured figures: a hit rate of $p = 58\%$, an average winning trade of $\bar{W} = 6.0$ pips, an average losing trade of $\bar{L} = 9.5$ pips, and an average round-trip cost of $\bar{c} = 1.2$ pips.

(a) Using the expectancy identity ([Course Book §14.1]), $\mathbb{E}[X] = p\bar{W} - (1-p)\bar{L} - \bar{c}$, compute the rule's expectancy per round trip.

(b) Solve for the hit rate $p^*$ that would make the expectancy exactly zero, holding $\bar{W}$, $\bar{L}$, and $\bar{c}$ fixed, and state the shortfall between $p^*$ and the rule's actual 58% in percentage points.

(c) In two or three sentences, explain why this chapter's answer to "does this system have an edge" reduces to the question "whose money am I taking, and why do they keep giving it to me" rather than to any hit-rate or accuracy statistic on its own, and name the honest reason (of Section 1.3's three) that would have to describe this rule's flow for the gap in (b) to be closeable without the rule simply becoming more often right about direction. What test does this book's platform run ([Course Book §14.7]) before accepting that the gap has, in fact, closed?

▷ working space

**Solution — Exercise 1.4**

**(a)**

$$\begin{aligned}\mathbb{E}[X] &= 0.58 \times 6.0 - 0.42 \times 9.5 - 1.2 \\ &= 3.4800 - 3.9900 - 1.2000 = -1.7100 \text{ pips per round trip,}\end{aligned}$$

negative, despite a hit rate eight points above a coin flip.
**(b)** Setting $p^*\bar{W} - (1-p^*)\bar{L} - \bar{c} = 0$ and solving for $p^*$:

$$p^*(\bar{W}+\bar{L}) = \bar{L}+\bar{c} \quad \Longrightarrow \quad p^* = \frac{\bar{L}+\bar{c}}{\bar{W}+\bar{L}} = \frac{9.5+1.2}{6.0+9.5} = \frac{10.7}{15.5} = 69.0323\%.$$

The rule needs a hit rate $69.0323\% - 58.00\% = 11.0323$ percentage points higher than its measured value before this particular payoff asymmetry and cost pays for itself.
**(c)** A hit rate, alone, answers "how often was the direction right," which Section 1.3 already showed is the wrong functional: it is silent on $\bar{W}$, $\bar{L}$, and $\bar{c}$, all three of which move the 11.0323-point gap in (b) just as much as $p$ itself would. Whether that gap can close is really a question about which of Section 1.3's three honest reasons describes the counterparty on the other side of these fills. Reasons (1) and (2) — a counterparty rebalancing inventory, or paying a genuine risk premium to transfer risk — can fund a persistent edge without the rule ever needing to become more accurate, because the counterparty is compensating the rule for a service rather than losing a forecasting contest. Reason (3) — the rule itself being the slower or less-informed side — cannot: it predicts exactly the arithmetic in (a) and (b), a persistent shortfall no amount of backtesting will fix, because the market is already pricing this flow correctly. This is precisely why the platform does not accept a favourable backtest at face value: Stage 0 ([Course Book §14.7]) is the pre-committed, cost-net, out-of-sample test of whether the envelope's own edge clears exactly this hurdle, which is the only way to find out which of the three reasons is actually true of this rule's flow, rather than assumed.

**Key idea — What to remember**

A market exists to discover a price, move risk to whoever will bear it most cheaply, and supply the readiness to trade that two isolated counterparties could never arrange for themselves, and it performs those jobs through some mix of auction matching and dealer quoting, inside either a centralised exchange or a bilateral OTC network — a distinction that sorts this book's four pillars cleanly, with the reader's own CFD account sitting entirely on the OTC side regardless of what it references. Every fill a systematic strategy receives has an honest explanation — inventory rebalancing, a genuine risk premium, or the strategy itself being the slower or less-informed side of the trade — and the spread and cost surface it pays are not a neutral toll but the market's running estimate of which explanation applies, an estimate that only the strategy's own net-of-cost, out-of-sample performance can overturn. Price formation itself is nothing more than this chapter's cast converting information into orders against quotes, repeatedly, which is the concrete mechanism behind the abstract martingale benchmark the Course Book derived; and even "the price" is a constructed, venue-dependent statistic once more than one place can trade the same instrument, exactly as contingent a construction as the toy auction's clearing price. Chapter 2 now opens the order book this chapter has been describing from the outside, teaching the reader to walk it by hand, level by level, and to decompose the spread just derived into the three structural pieces — order-processing cost, inventory risk, and the adverse selection introduced here — that together set its width.

# 2

# The Order Book

*Chapter 1 described a market from the outside: what it does, who is in it, and why anyone trades with anyone else. This chapter opens the machine. The limit order book is the object that every auction market's price is made of — a queue of promises, sorted by price and then by time, against which every incoming market order is matched mechanically and without discretion. A reader who can walk that book by hand can compute, before sending an order, exactly what the order will cost: not the quoted spread, which is a price available only to an infinitesimal trade, but the average fill price a specific quantity will actually achieve, the slippage that quantity pays for its own size, and the mark on the book it leaves behind. The chapter then asks the harder question: where does the spread come from in the first place? It decomposes the answer into three structural pieces — the cost of processing an order, the risk of holding inventory, and the losses a liquidity supplier makes to counterparties who know more — states the Glosten–Milgrom argument that turns the third piece into a breakeven quote, computes it numerically, and shows how Roll's estimator recovers an effective spread from nothing but a sequence of transaction prices. It ends where the reader's own account sits: in a dealer-streamed FX market with no consolidated book at all, where the quote on screen is one firm's opinion rather than the market's, and where a mid-only price history silently deletes exactly the object this chapter spends its length teaching you to read.*

## 2.1 Anatomy of a Limit Order Book

A limit order is a promise: *I will buy up to this quantity at this price or better, and my promise stands until I withdraw it*. An exchange that accepts such promises must store them, and the data structure it stores them in — one queue per price level, on each side of the market — is the limit order book. Everything else about a continuous auction market follows from how that structure is maintained and matched.

**Definition — Limit order book, level, depth, touch**

The **limit order book** (LOB) of an instrument is the complete set of unexecuted limit orders resting at a venue, organised into two sides. The **bid side** holds buy orders, sorted from the highest price downwards; the **ask** (or offer) **side** holds sell orders, sorted from the lowest price upwards. Orders sharing a price form a **level**, and the total quantity resting at a level is that level's **depth**; *depth* used without qualification means the cumulative quantity available within some stated distance of the best prices. The **best bid** $P_b$ is the highest price any buyer will pay and the **best ask** $P_a$ the lowest price any seller will accept; the pair $(P_b, P_a)$ is the **touch**, or best bid and offer (BBO), and the two prices must satisfy $P_b < P_a$, since any buy order priced at or above the best ask would have executed instead of resting. The **quoted spread** is $s = P_a - P_b$ and the **mid price** is $m = (P_a + P_b)/2$ (Harris, 2003).

The book is a two-sided ledger of *intentions*, not of positions or of value. Nothing in it is a forecast; every line is a standing offer that its owner may cancel, and in modern electronic markets

the overwhelming majority of orders are cancelled rather than executed. What the book gives a trader is not a prediction but a schedule: the price of immediacy, quantity by quantity, at this instant.

### A book, in full

Table 2.1 shows five levels each side of a hypothetical order book for a listed share — an illustrative scenario computed exactly, not market data — trading around $27.50 with a one-cent minimum price increment. Every calculation in Sections 2.1 to 2.4 refers back to it.

**Table 2.1.** A hypothetical five-level limit order book. Sizes in units; tick $0.01. Cumulative columns run outward from the touch.

| Bid side | | | Ask side | | |
|---|---|---|---|---|---|
| Price | Size | Cum. | Price | Size | Cum. |
| **27.52** | **60** | 60 | **27.53** | **140** | 140 |
| 27.51 | 220 | 280 | 27.54 | 60 | 200 |
| 27.50 | 400 | 680 | 27.55 | 200 | 400 |
| 27.49 | 260 | 940 | 27.56 | 150 | 550 |
| 27.48 | 500 | 1,440 | 27.57 | 350 | 900 |

Read the touch first. The best bid is $27.52 for 60 units; the best ask is $27.53 for 140 units. The quoted spread is

$$s = 27.53 - 27.52 = \$0.01,$$

exactly one tick, and the mid is $m = (27.53 + 27.52)/2 = \$27.525$. Those two numbers are what a price feed usually reports, and between them they describe only 200 units of the 2,340 units resting in the table. The rest of the book — 1,440 units bid, 900 units offered across five levels — is invisible to anyone who looks only at the touch, and it is precisely the part that determines what a trade of any consequential size will cost.

Two features of the touch already carry information. First, the sizes are lopsided: 140 units offered against 60 bid. Second, the notional amounts resting differ correspondingly — $60 \times \$27.52 = \$1{,}651.20$ on the bid against $140 \times \$27.53 = \$3{,}854.20$ on the ask. A trader wanting to buy faces nearly two and a half times as much displayed supply as a trader wanting to sell faces displayed demand. The standard way to compress that asymmetry into one number is the order book imbalance, and the standard way to convert it into a price is the microprice.

**Definition — Imbalance and the microprice**

With best bid and ask quantities $Q_b$ and $Q_a$, the **touch imbalance** is

$$I = \frac{Q_b - Q_a}{Q_b + Q_a} \in [-1, 1],$$

positive when more size rests on the bid. The **microprice** (or size-weighted mid) is the quantity-weighted average of the two best prices, weighted by the *opposite* side's quantity:

$$P_{\text{micro}} = \frac{P_b Q_a + P_a Q_b}{Q_a + Q_b}.$$

The crossed weighting is deliberate: a large resting ask is unexecuted supply, and pushes the weighted price *down* towards the bid, which is where a book heavy with sellers is more likely to trade next.

**Result — The microprice is the mid, shifted by imbalance**

Writing $P_b = m - s/2$ and $P_a = m + s/2$ and substituting into Definition 2.1,

$$P_{\text{micro}} = \frac{(m - \frac{s}{2})Q_a + (m + \frac{s}{2})Q_b}{Q_a + Q_b} = m + \frac{s}{2} \cdot \frac{Q_b - Q_a}{Q_a + Q_b} = m + \frac{s}{2} I.$$

The microprice is therefore the mid displaced towards the thin side of the touch by a fraction $I$ of the half-spread; it can never leave the interval $[P_b, P_a]$, and it equals the mid exactly when the touch is balanced.

For Table 2.1, $I = (60 - 140)/(60 + 140) = -0.40$, so

$$P_{\text{micro}} = 27.525 + 0.005 \times (-0.40) = \$27.5230,$$

which the direct formula confirms: $(27.52 \times 140 + 27.53 \times 60)/200 = 5{,}504.60/200 = \$27.5230$. The microprice sits 0.2 of a tick below the mid. Both numbers are legitimate summaries of "the price", and they disagree — a first, small instance of a theme this book returns to repeatedly: every scalar price is a chosen statistic of a richer object, and the choice matters.

### Tick size, and why a big tick lengthens the queue

The **tick size** is the minimum price increment a venue permits. It is a rule, not a market outcome, and venues set it by schedule — commonly finer for low-priced instruments and coarser for high-priced ones. Tick schedules differ by venue and asset class and are revised from time to time, so treat any specific increment as indicative and verify the current specification for any instrument you intend to trade.

The economics of the tick are worth stating because they are stable even when the numbers are not. If the tick is small relative to the spread that competition would otherwise produce, it does not bind: liquidity suppliers undercut each other in price until the spread reaches its economic floor, and the book near the touch is thin because there is little advantage to standing in a queue you can step in front of for a hundredth of a cent. If the tick is large relative to that floor — the case for many lower-priced shares — it binds: the spread cannot narrow further, competition among suppliers can no longer be expressed in price, and it moves instead to *time*. The result is a long queue at each of the few available prices, a wider-than-necessary spread, and a real economic value attached to queue position (Harris, 2003). A binding tick transfers surplus from liquidity demanders (who pay a spread wider than competition would set) to liquidity suppliers (who earn it), and rewards speed, because being early in the queue becomes the only way to compete.

### Price–time priority, and the arithmetic of the queue

Almost every continuous auction market matches on **price–time priority**: among resting orders, the one offering the best price executes first, and among orders at the same price, the one that arrived first executes first. The rule is mechanical and public, which is the point — it removes discretion from matching, so no participant can be favoured after the fact.

Price–time priority makes queue position a quantity you can compute with. Suppose you post a buy order for 100 units at \$27.52, joining the back of the 60 units already resting there in Table 2.1. Incoming sell orders execute against the resting 60 units first. Your order therefore begins to fill only after 60 units have traded at \$27.52, and fills completely only after 160 units have traded there — provided nobody ahead of you cancels, which would move you up, and provided the level is not skipped entirely by a price move away from \$27.52, which would leave you unfilled. If 110 units of selling arrive at that price, the first 60 fill the orders ahead of you and the remaining 50 fill exactly half of yours: $(110 - 60)/100 = 0.50$.

This is the fundamental asymmetry of the two order types. A market order buys certainty of execution and pays for it in price. A limit order buys price and pays for it in execution uncertainty

— an uncertainty that is not random but adversely correlated with what you want, because the states of the world in which your resting bid is filled are disproportionately the states in which the price was about to fall. Chapter 1 introduced the order types and their time-in-force conventions; this chapter supplies the arithmetic behind them, and Section 2.3 supplies the reason the correlation runs the wrong way.

### What the displayed book leaves out

Three caveats belong here, because a reader who treats the displayed book as the whole of available liquidity will systematically mis-model execution.

**Hidden and reserve quantity.** Many venues accept orders that display only part of their size (commonly called iceberg or reserve orders) or none of it. Displayed depth is therefore a lower bound on executable depth at a given price, and the difference is not observable ex ante. Which hidden-order types a venue supports, and on what terms, is venue-specific — verify current specifications rather than assuming.

**Latency.** The book you are looking at is the book as it was when the message describing it left the matching engine. By the time your order arrives, resting orders may have been cancelled and new ones added. A "walk" computed from a snapshot is therefore an estimate, not a guarantee, and the estimate degrades exactly when the book is moving fastest.

**Fleeting liquidity.** Displayed size can be withdrawn faster than a slow participant can reach it. This is not necessarily manipulation: a liquidity supplier who updates quotes on every change in a correlated instrument will cancel and repost continuously, and the resulting book looks deep in a snapshot and thin to anyone who tries to trade against it in size.

## 2.2 Walking the Book

A market order is an instruction to trade a stated quantity immediately, at whatever prices the book supplies. The matching engine executes it by a loop so simple it can be run by hand: take the best price level on the opposite side, fill against it up to its depth, decrement the quantity remaining, move outward one level, repeat until the order is filled or the book is exhausted. The trade therefore does not execute at one price. It executes at several, and the price a trader actually receives is the quantity-weighted average of them.

**Result — Walking the book**

Let the opposite side of the book consist of levels $(p_1, q_1), (p_2, q_2), \ldots$ ordered outward from the touch, and let $Q$ be the size of an incoming market order. Define the cumulative depth $C_k = \sum_{i \le k} q_i$ (with $C_0 = 0$) and let $K = \min\{k : C_k \ge Q\}$ be the last level touched. The executed quantities are $x_i = q_i$ for $i < K$ and $x_K = Q - C_{K-1}$, and the total consideration and average fill price are

$$V = \sum_{i=1}^{K} x_i\, p_i, \qquad \bar{p} = \frac{V}{Q} = \frac{1}{Q} \sum_{i=1}^{K} x_i\, p_i.$$

If $Q$ exceeds the total depth of the book, only $C_{\max}$ units execute and the remainder either rests as a limit order at the last price, or is cancelled, according to the order's instructions. For a buy, $\bar{p} \ge p_1$ always, with equality only when $Q \le q_1$; the excess $\bar{p} - p_1$ is the price the order pays for its own size.

**Worked Example 2.1 — Walking a five-level ask side**

A trader sends a market order to buy 250 units against the hypothetical book of Table 2.1. All figures below are computed exactly from that book; it is an illustrative scenario, not market data.

**(a) The fills.** Cumulative ask depth is 140, 200, 400 at the first three levels, so $K = 3$: the order clears the first two levels entirely and takes $250 - 200 = 50$ units from the third.

| Level | Price | Filled | Consideration | Cumulative filled |
|---|---|---|---|---|
| 1 | \$27.53 | 140 | \$3,854.20 | 140 |
| 2 | \$27.54 | 60 | \$1,652.40 | 200 |
| 3 | \$27.55 | 50 | \$1,377.50 | 250 |
| | | **250** | **\$6,884.10** | |

**(b) Average fill price.**

$$\bar{p} = \frac{6{,}884.10}{250} = \$27.5364.$$

**(c) Slippage against the quoted ask.** The screen showed \$27.53. The order paid $27.5364 - 27.53 = \$0.0064$ per unit more than that — 0.64 of a tick — or $250 \times 0.0064 = \$1.60$ in total. This is *slippage*: the gap between the touch and the achieved average, caused entirely by the order's own size walking up the schedule.

**(d) Cost against the mid, and the effective spread.** The economically correct benchmark is not the ask but the mid at the moment of the decision, $m = \$27.525$, since the mid is the price at which neither side is paying for immediacy. The *effective half-spread* paid is

$$\bar{p} - m = 27.5364 - 27.525 = \$0.0114 \text{ per unit} = 1.14 \text{ ticks},$$

a total cost of $250 \times 0.0114 = \$2.85$, which doubled gives an effective spread of \$0.0228, or 2.28 ticks. The quoted spread was one tick. **This order's effective spread was** 2.28 **times the quoted spread**, and a cost model built on the quoted spread would have predicted $(s/2)Q = 0.005 \times 250 = \$1.25$ against a true cost of \$2.85 — an understatement of \$1.60, or 56.1404% of the true figure. In relative terms the cost was $10^4 \times 0.0114/27.525 = 4.1417$ basis points of the mid one way, 8.2834 basis points round trip.

**(e) What the order leaves behind.** After the trade, levels one and two are gone and 150 units remain at \$27.55. The new touch is \$27.52/\$27.55 — a three-tick spread, three times its former width — and the new mid is \$27.535, one tick above the old one. The order has moved the mid by a full tick without a single new piece of information arriving. How long that displacement persists is a separate question, of *resilience*: liquidity suppliers who consider the move uninformative will repost, narrowing the book back towards its previous state within seconds; those who read the trade as informed will repost higher, or not at all.

**(f) The alternative: a marketable limit order.** Had the trader instead sent a limit order to buy 250 units *at \$27.54 or better*, it would have swept only the first two levels: 200 units at an average of $5{,}506.60/200 = \$27.5330$, an effective half-spread of 0.80 of a tick, with 50 units — 20% of the intended trade — left unfilled and resting at \$27.54. The protection is real: on the units that did trade, the average price improved by $27.5364 - 27.5330 = 0.34$ of a tick. What it costs is the fills you do not get, and those go missing precisely when the market is moving away from you — so the saving is booked in the good states and the shortfall in the bad ones.

**Definition — Quoted, effective, and realised spreads; price impact**

For a single execution of a buy at price $p$ when the prevailing mid was $m_0$, and with $m_\tau$ the mid a short interval $\tau$ later:

- the **quoted spread** $s = P_a - P_b$ is the width of the touch — a price available only for

quantities up to the displayed size;

- the **effective half-spread** is $p - m_0$, the cost actually paid relative to the decision mid (for a sell, $m_0 - p$); twice this is the **effective spread**;
- the **price impact** is $m_\tau - m_0$, the part of the trade's cost that is a permanent repricing rather than a payment to the counterparty;
- the **realised half-spread** is $p - m_\tau$, what the liquidity supplier actually keeps after the market has moved.

The three are linked by an identity that holds trade by trade, with no assumptions:

$$\underbrace{p - m_0}_{\text{effective}} = \underbrace{p - m_\tau}_{\text{realised}} + \underbrace{m_\tau - m_0}_{\text{impact}}.$$

Aggregated over many trades, this is the standard empirical decomposition of trading cost into the part that compensates the liquidity supplier and the part that reflects information (Hasbrouck, 2007). The corresponding portfolio-level measure — the total shortfall between the price that motivated a decision and the average price actually achieved, including the cost of orders never filled — is the **implementation shortfall** (Perold, 1988).

Apply the identity to Worked Example 2.1. Suppose that five minutes later the mid has settled at \$27.535 — the level the trade itself displaced it to, with no reversion. Then the price impact is $27.535 - 27.525 = \$0.0100$ per unit and the realised half-spread is $27.5364 - 27.535 = \$0.0014$ per unit, and indeed $0.0014 + 0.0100 = 0.0114$, the effective half-spread computed in part (d). Of the 1.14 ticks the buyer paid, the liquidity suppliers who sold to him keep 0.14 of a tick — 12.2807% — and the remaining 87.7193% is a repricing they did not capture. This is not an accounting curiosity. It is the entire economics of Section 2.4: the party quoting the spread earns far less than the spread, because the market moves against a proportion of their fills, and the size of that proportion is what sets the spread's width in the first place.

### Slippage is not the spread, and the spread is not the cost

Three distinct quantities are routinely conflated, and separating them is the practical payoff of this section.

**The quoted spread** is a property of the book at an instant. It is the cost of a round trip for a trade small enough to fit inside the displayed size at the touch, and nothing else. It is the only one of the three that is observable without trading.

**Slippage** is the difference between the price a trader expected — typically the touch, or the last trade — and the average price achieved. It is generated by two mechanisms that ought never to be confused: *size* (the order walks the book, as in Worked Example 2.1) and *delay* (the book moved between decision and arrival). The first is deterministic given a snapshot and can be computed in advance; the second is a random variable whose distribution widens sharply with volatility and with the latency of the path between trader and matching engine.

**The effective spread** is the realised, size-and-delay-inclusive cost, measured against the decision mid. It is the only one of the three that belongs in an expectancy calculation, and it is the one a backtest built from quoted spreads never sees. Every cost figure in this book's later chapters, and every cell of the reader's own cost surface, is an estimate of this quantity, not of the quoted spread.

### A linear book, and the cost of size

Exact walks require the whole book. A useful approximation requires only its slope. Suppose depth is roughly uniform per tick near the touch, so that executing $Q$ units pushes the price

outward linearly. Then the average fill price for a buy is approximately

$$\bar{p} \;\approx\; m + \frac{s}{2} + \frac{\lambda}{2} Q, \tag{2.1}$$

where $\lambda > 0$ is the book's local slope in price per unit of quantity, and the total cost against the mid is

$$\text{Cost}(Q) = Q(\bar{p} - m) \approx \frac{s}{2} Q + \frac{\lambda}{2} Q^2.$$

The first term is linear in size — the spread you would pay even for an infinitesimal trade — and the second is *quadratic*: doubling the order quadruples the impact component. This convexity is the whole reason large orders are worked in pieces rather than sent at once, and it is the cost term that execution-scheduling models trade off against the risk of delay (Almgren, Chriss, 2001).

Fitting (2.1) to Worked Example 2.1 recovers the slope of that particular book:

$$\lambda = \frac{2(\bar{p} - m - s/2)}{Q} = \frac{2(27.5364 - 27.525 - 0.005)}{250} = \frac{2 \times 0.0064}{250} = 5.12 \times 10^{-5} \text{ dollars per unit,}$$

so $1/\lambda = 19{,}531.25$ units of one-directional flow per dollar of price displacement. The decomposition then reads Cost $= 0.005 \times 250 + \frac{1}{2}(5.12 \times 10^{-5})(250)^2 = 1.25 + 1.60 = \$2.85$, reproducing part (d) exactly. Note what the split says: for this order size in this book, the size-dependent term is larger than the spread term. A trader who models cost as "half the spread" is not making a small error; on this trade he is capturing 44% of the cost and missing the majority of it.

## 2.3 Where the Spread Comes From

Section 2.2 treated the spread as given. It is not given: it is a price, set by the parties who supply liquidity, and like any price it must cover the costs of supplying the thing. There are exactly three such costs, and the classification is the single most useful piece of microstructure theory a systematic trader can hold, because each component responds to different conditions and each leaves a different signature in the data (O'Hara, 1995; Harris, 2003).

**Result — The three-component decomposition of the spread**

The equilibrium quoted spread of a competitive liquidity supplier decomposes as

$$s \;=\; \underbrace{2c}_{\text{order processing}} \;+\; \underbrace{s_{\text{inv}}}_{\text{inventory risk}} \;+\; \underbrace{s_{\text{adv}}}_{\text{adverse selection}},$$

where

- $c$ is the per-side, per-unit **order-processing cost**: exchange and clearing fees, connectivity, settlement, and the amortised cost of the technology and personnel required to quote continuously. It is largely fixed per trade, so it falls as a fraction of value as trade size and price rise, and it sets a hard floor under the spread of any small or infrequently traded instrument.
- $s_{\text{inv}}$ compensates **inventory risk**: having bought at the bid, the supplier holds a position whose value moves before it can be laid off. This component scales with the volatility of the instrument, with the size of the position relative to the supplier's risk capacity, and with the expected time to flatten — so it widens when volatility rises, when the supplier's inventory is already skewed, and when the opposite side of the flow is slow to arrive.
- $s_{\text{adv}}$ compensates **adverse selection**: the supplier's quotes are options written to the whole market, and the counterparties most eager to exercise them are those who know something. This component is the subject of Glosten and Milgrom (1985) and is derived below.

Only the third component would survive in a world of costless, risk-neutral, infinitely capitalised liquidity supply — and it would survive at full strength, which is why it dominates the theory.

The first two components are, in a precise sense, the easy ones: they are costs of *doing* the business, and they would exist even if every counterparty were uninformed. They also explain the cross-section of spreads that Amihud and Mendelson (1986) showed is priced — assets that are expensive to trade must offer higher gross returns to be held at all, because investors discount them for the round-trip cost they will eventually pay. The third component is different in kind: it is a cost of being *on the other side* of someone better informed, and it does not vanish with scale, technology, or capital.

### Adverse selection, made numerical: the Glosten–Milgrom argument

Consider the simplest market in which information can be asymmetric. An asset's value $V$ will be revealed shortly and takes one of two values, $V_H$ or $V_L$, with $V_H > V_L$; write $\Delta = V_H - V_L$ and let $\pi = \mathbb{P}(V = V_H)$ be the public prior. Traders arrive one at a time. With probability $\mu$ the arriving trader is *informed*: she already knows $V$, and buys if $V = V_H$, sells if $V = V_L$. With probability $1 - \mu$ the arriving trader is *uninformed* — trading for liquidity, hedging, or noise — and buys or sells with probability $\frac{1}{2}$ each, independently of $V$. A competitive, risk-neutral dealer with no inventory concerns and no processing costs quotes a bid and an ask before seeing who has arrived, and competition drives expected profit on each side to zero.

The zero-profit condition is not "the dealer breaks even on average across all trades". It is stronger and stranger: the dealer must break even *conditional on the trade that actually arrives*, because the arrival of a buy is itself evidence about $V$. Hence

$$P_a = \mathbb{E}[V \mid \text{buy}], \qquad P_b = \mathbb{E}[V \mid \text{sell}],$$

and the spread is the difference between two posterior means computed from the same prior with opposite evidence. This is the whole of the Glosten–Milgrom (1985) mechanism: *the spread is the market's Bayesian update, priced*.

**Result — The Glosten–Milgrom breakeven spread**

Under the assumptions above, Bayes' rule gives

$$\mathbb{P}(V_H \mid \text{buy}) = \frac{\pi\left[\mu + \frac{1-\mu}{2}\right]}{\pi\left[\mu + \frac{1-\mu}{2}\right] + (1-\pi)\frac{1-\mu}{2}}, \qquad \mathbb{P}(V_H \mid \text{sell}) = \frac{\pi\frac{1-\mu}{2}}{\pi\frac{1-\mu}{2} + (1-\pi)\left[\mu + \frac{1-\mu}{2}\right]},$$

with $P_a = V_L + \Delta\,\mathbb{P}(V_H \mid \text{buy})$ and $P_b = V_L + \Delta\,\mathbb{P}(V_H \mid \text{sell})$. In the symmetric case $\pi = \frac{1}{2}$ these collapse to $\mathbb{P}(V_H \mid \text{buy}) = \frac{1+\mu}{2}$ and $\mathbb{P}(V_H \mid \text{sell}) = \frac{1-\mu}{2}$, so that

$$\boxed{s_{\text{adv}} = P_a - P_b = \mu\,\Delta}$$

— the adverse-selection spread is exactly the probability of facing an informed trader times the size of the information at stake. Restoring an order-processing cost $c$ per unit per side widens the quoted spread to $s = \mu\Delta + 2c$; adding inventory aversion widens it further and, when the dealer's position is skewed, shifts both quotes in the direction that encourages the flow which would flatten it (Section 2.4).

Three consequences deserve stating before any arithmetic. First, the spread widens with $\mu$ and with $\Delta$ *multiplicatively*: a market with rare but enormous information events can carry the same

spread as one with frequent trivial ones. Second, the dealer loses money on every informed trade and makes money on every uninformed one — the spread is not a fee levied on everyone equally but a transfer from uninformed to informed traders, intermediated by a dealer who nets to zero. Third, the mechanism is self-correcting in price: because each trade updates the posterior, quotes walk towards the true value as informed traders transact, which is precisely the microstructure channel by which private information becomes public price (Chapter 1's price-formation chain, now with the arithmetic attached).

**Worked Example 2.2 — Glosten–Milgrom quotes before a binary announcement**

The hypothetical share of Table 2.1 faces a binary announcement that will move it to either $V_H = \$27.60$ or $V_L = \$27.40$, judged equally likely, so $\pi = 0.5$ and $\Delta = \$0.20$ (20 ticks). A fifth of arriving order flow is informed: $\mu = 0.20$. All figures are computed exactly for this illustrative scenario.

**(a) Trade probabilities.** An informed trader buys only in the high state, an uninformed trader buys half the time:

$$\mathbb{P}(\text{buy} \mid V_H) = 0.20 + 0.80 \times 0.5 = 0.60, \qquad \mathbb{P}(\text{buy} \mid V_L) = 0.80 \times 0.5 = 0.40,$$

so $\mathbb{P}(\text{buy}) = 0.5 \times 0.60 + 0.5 \times 0.40 = 0.50$, as symmetry requires.

**(b) Posteriors and quotes.**

$$\mathbb{P}(V_H \mid \text{buy}) = \frac{0.5 \times 0.60}{0.50} = 0.60, \qquad \mathbb{P}(V_H \mid \text{sell}) = \frac{0.5 \times 0.40}{0.50} = 0.40.$$

Hence

$$P_a = 0.60 \times 27.60 + 0.40 \times 27.40 = \$27.52, \qquad P_b = 0.40 \times 27.60 + 0.60 \times 27.40 = \$27.48,$$

a spread of \$0.04 — four ticks — with a midpoint of \$27.50, exactly the prior mean $\mathbb{E}[V] = \$27.50$. The closed form agrees: $\mu\Delta = 0.20 \times 0.20 = \$0.04$.

**(c) The zero-profit check, decomposed.** Conditional on a buy arriving, the probability it came from an informed trader is $\mathbb{P}(\text{informed} \mid \text{buy}) = \mu\pi/\mathbb{P}(\text{buy}) = 0.20 \times 0.5/0.50 = 0.20$. On such a trade the dealer sells at \$27.52 an asset worth \$27.60 and loses \$0.08 per unit. On an uninformed buy the dealer sells at \$27.52 an asset worth $\mathbb{E}[V] = \$27.50$ and gains \$0.02 per unit. Expected profit per buy:

$$0.20 \times (-0.08) + 0.80 \times (+0.02) = -0.016 + 0.016 = \$0.00.$$

The dealer's "profit" is exactly the losing trades' mirror image: four winning trades of two cents pay for one losing trade of eight cents. This ratio — many small wins funding rare large losses — is the shape of the market-making P&L distribution, and it is why a market maker's average day tells you almost nothing about the risk being run.

**(d) With order-processing costs.** At $c = \$0.005$ per unit per side, the quoted spread becomes $\mu\Delta + 2c = 0.04 + 0.01 = \$0.05$ (five ticks), quoted \$27.475 / \$27.525. Of the five ticks, four — 80% — are adverse-selection compensation and one is cost recovery.

**(e) Comparative statics.** Holding $c$ fixed, $\mu = 0.35$ gives a spread of \$0.08 (eight ticks) and $\mu = 0.05$ gives \$0.02 (two ticks); doubling the news at stake to $\Delta = \$0.40$ with $\mu = 0.20$ gives \$0.09 (nine ticks). Running the relation backwards is more useful still: an observed quoted spread of \$0.06 with $c = \$0.005$ and $\Delta = \$0.20$ implies $\mu = (0.06 - 0.01)/0.20 = 0.25$. *The width of the spread is the market's estimate of how likely you are to be informed* — an estimate you can read off the screen and invert.

**(f) Reconciling with the quiet book.** Table 2.1 showed a one-tick spread, not a five-tick one, because that book was quoted on an ordinary day with no announcement pending. Suppose

that on such a day the value uncertainty resolved by a single trade is only $\Delta = \$0.03$ (three ticks) and processing costs are $c = \$0.002$ per side, with $\mu$ unchanged at 0.20. Then

$$s = \mu\Delta + 2c = 0.20 \times 0.03 + 0.004 = 0.006 + 0.004 = \$0.010,$$

exactly the one-tick spread of Table 2.1: 60% adverse selection, 40% order processing. The same model, with the same informed fraction, produces a one-tick book on a quiet day and a five-tick book before an announcement, purely because $\Delta$ moved. This is the mechanism behind every empirical regularity about spreads widening around news, and behind the reader's own platform's finding that measured spreads blow out at scheduled data releases.

### Kyle's $\lambda$: the same idea, priced per unit of flow

Glosten and Milgrom put one trader at a time through the book and let quotes update trade by trade. Kyle (1985) asks the complementary question: if a single informed trader can choose *how much* to trade, and hides inside a crowd of uninformed order flow, how much does the price move per unit of net flow? The equilibrium in that model has a property that has organised empirical microstructure ever since — price impact is *linear* in aggregate order flow:

$$\Delta P = \lambda \,(\text{net signed order flow}), \tag{2.2}$$

with a single coefficient $\lambda > 0$ whose reciprocal $1/\lambda$ is the market's **depth**: the quantity of net buying required to lift the price by one unit. In Kyle's equilibrium $\lambda$ rises with the amount of private information and falls with the volume of uninformed flow available to hide in — exactly the two forces that set $\mu$ and $\Delta$ in the Glosten–Milgrom account, now expressed as a slope instead of a spread. The informed trader, knowing (2.2), does not trade her whole position at once; she spreads it out to avoid revealing herself, which is the theoretical origin of order splitting and of the execution-scheduling problem generally. This book states the result and its interpretation; the derivation belongs to a course in market microstructure theory (Kyle, 1985; O'Hara, 1995).

Equation (2.2) is the same $\lambda$ that appears as the book slope in (2.1), but the two are conceptually distinct and should not be silently identified. The book slope is a *mechanical* quantity: it describes what the currently resting orders will charge you for size, and it reverses when the book replenishes. Kyle's $\lambda$ is an *informational* quantity: it describes how much the market's belief about value moves per unit of flow, and it does not reverse. In the decomposition of Definition 2.2, the mechanical part shows up in the realised half-spread and the informational part in the price impact — which is why measuring both, separately, on your own fills is worth more than any single cost number. In Worked Example 2.1, 87.7193% of the buyer's cost was impact and only 12.2807% was payment to the liquidity supplier.

### Roll's estimator: recovering the spread from prices alone

Suppose you have no order book at all — only a sequence of transaction prices. Roll (1984) showed that the effective spread is still identified, because the bid-ask bounce leaves a distinctive fingerprint: negative first-order autocovariance in price changes.

The argument is three lines. Let $m_t$ be the unobserved efficient price, following a random walk with increments independent of trade direction, and let each observed transaction print at

$$p_t = m_t + \frac{s}{2} q_t, \qquad q_t = \pm 1,$$

where $q_t = +1$ for a buyer-initiated trade and $-1$ for a seller-initiated one, with the $q_t$ independent, equally likely, and independent of the efficient price. Then $\Delta p_t = \Delta m_t + \frac{s}{2}(q_t - q_{t-1})$, and since $\Delta m$ is serially uncorrelated and independent of the $q$'s,

$$\operatorname{Cov}(\Delta p_t, \Delta p_{t-1}) = \frac{s^2}{4} \operatorname{Cov}(q_t - q_{t-1},\, q_{t-1} - q_{t-2}) = \frac{s^2}{4}\left(-\mathbb{E}[q_{t-1}^2]\right) = -\frac{s^2}{4},$$

because $\mathbb{E}[q_i q_j] = 0$ for $i \neq j$ and $\mathbb{E}[q^2] = 1$. Inverting gives the estimator.

**Result — Roll's implicit spread estimator**

Under the assumptions above — an efficient price uncorrelated with trade direction, no drift, and trade signs that are serially independent and equally likely — the effective spread satisfies

$$s = 2\sqrt{-\operatorname{Cov}(\Delta p_t, \Delta p_{t-1})},$$

estimated from $N$ consecutive transaction prices by

$$\hat{\gamma}_1 = \frac{1}{N-2}\sum_{k=1}^{N-2} \Delta p_k \, \Delta p_{k+1}, \qquad \hat{s}_{\text{Roll}} = 2\sqrt{-\hat{\gamma}_1} \quad \text{if } \hat{\gamma}_1 < 0.$$

The estimator is undefined when $\hat{\gamma}_1 \geq 0$, which occurs frequently in practice — with trending prices, with serially correlated trade signs (order splitting makes runs of same-signed trades the norm, not the exception), or simply by sampling noise in short windows (Roll, 1984; Harris, 2003; Hasbrouck, 2007).

**Worked Example 2.3 — Roll's estimator on six trades**

Later in the same hypothetical session the book has thinned to a three-tick touch of \$27.52 / \$27.55, and six consecutive trades print at

$$27.52, \quad 27.54, \quad 27.52, \quad 27.53, \quad 27.55, \quad 27.53.$$

Every print lies within the quoted touch — at one of the two quotes or between them — as it must.

**(a) Price changes.** Five changes: $+0.02,\ -0.02,\ +0.01,\ +0.02,\ -0.02$.

**(b) Lagged products.** Four consecutive products:

$$(0.02)(-0.02) = -0.0004, \qquad (-0.02)(0.01) = -0.0002,$$
$$(0.01)(0.02) = +0.0002, \qquad (0.02)(-0.02) = -0.0004,$$

summing to $-0.0008$, so $\hat{\gamma}_1 = -0.0008/4 = -0.0002$. The estimate is defined because the sum is negative — three of the four products are negative, the signature of a price bouncing between two quotes rather than trending.

**(c) The estimate.**

$$\hat{s}_{\text{Roll}} = 2\sqrt{0.0002} = 2 \times 0.0141421356 = \$0.0282842712,$$

that is, 2.8284 ticks, or 10.2721 basis points of a \$27.535 reference price.

**(d) Reading it.** The quoted spread over this window was three ticks; Roll's estimate is 2.8284 ticks, 94.2809% of it. The shortfall is the correct sign: trades executing inside the quotes — at a price improvement, or against hidden liquidity — make the *effective* spread smaller than the quoted one, and Roll measures the effective spread, because it is built from what was actually paid. Had the estimate exceeded the quoted spread materially, the natural readings would be that the touch had moved during the window or that the trade signs were not independent.

**(e) The drift caveat.** These six prices drift upward slightly: the mean change is $+\$0.002$. Demeaning the changes before multiplying gives $\hat{\gamma}_1 = -0.000206$ and $\hat{s} = \$0.0287054$, or 2.8705 ticks — a difference of 0.0421 of a tick, immaterial here. Over longer windows with genuine drift the correction is not immaterial, and Roll's estimator should always be computed on demeaned changes when drift is plausible.

## 2.4 Market Making as a Business

Section 2.3 derived the spread from the liquidity supplier's constraints. This section runs the business, because a systematic trader who understands the market maker's profit and loss understands, by symmetry, his own costs: they are the same numbers with the signs reversed.

A market maker's job is to quote a two-sided price continuously and to end each day approximately flat, earning the spread on the round trips in between. Three facts make this harder than it sounds. First, the two sides of a round trip do not arrive together — the maker is long or short for the interval between them, and that interval is a risk position nobody chose. Second, the arrival of the two sides is not independent of the price: buyers cluster when the price is about to rise. Third, the maker cannot decline a trade. Quotes are firm; the only levers are *width* (how far apart to quote), *size* (how much to show), and *skew* (where to centre the quote relative to the maker's own estimate of value).

**Result — The market maker's P&L identity**

Over a session in which the maker completes $n_g$ round trips of $q$ units at a quoted spread $s$ without incident, suffers $n_a$ adverse fills on which the fair value jumps by $J_i$ against the position before it can be flattened, and executes $E$ times in total at a processing cost $\phi$ per execution, the realised trading profit is

$$\Pi \;=\; \underbrace{n_g\, q\, s}_{\text{spread capture}} \;-\; \underbrace{q \sum_{i=1}^{n_a} \left(J_i - \tfrac{s}{2}\right)}_{\text{adverse selection}} \;-\; \underbrace{E\,\phi}_{\text{order processing}} \;-\; \underbrace{\mathcal{R}(\text{inventory})}_{\text{risk carried}}.$$

The first three terms are realised cash. The fourth is not a cash item at all: it is the risk attached to whatever position the maker still holds, and it is the term that ends careers, because it is the only one that can exceed the day's gross capture by an order of magnitude. Note the algebraic structure — capture is linear in $s$ and *so is the offset to adverse selection*, since a wider quote means the maker was filled further from fair value before the jump; both push the optimal $s$ up when $J$ or $n_a$ rises.

**Worked Example 2.4 — A market maker's toy day**

A hypothetical maker quotes the share of Table 2.1 one tick wide ($27.52 bid, $27.53 ask) in clips of $q = 100$ units, with a processing cost of $\phi = \$0.10$ per execution — an assumed figure for this illustration; real fee schedules vary by venue and by participant status and must be verified against current specifications. Over the session the maker completes 22 round trips: 20 benign, and 2 on which fair value jumps $J = 5.5$ ticks against the position immediately after the fill, so that the maker flattens at a loss of 5.0 ticks per unit after crediting the half tick earned on entry.

| Component | Working | Amount |
|---|---|---|
| Gross spread capture | $20 \times 100 \times \$0.01$ | $+\$20.00$ |
| Adverse selection | $2 \times 100 \times \$0.05$ | $-\$10.00$ |
| Order processing | $44 \times \$0.10$ | $-\$4.40$ |
| **Net trading P&L** | | $+\mathbf{\$5.60}$ |

**(a) What survives.** The maker quoted a one-tick spread all day and kept $5.60 of the $20.00 gross — 28.00%. Spread across 22 round trips that is $0.2545 each, and spread across the 4,400 units of one-way volume traded it is $0.0012727 per unit: 0.1273 of a tick, about one-eighth

of the spread quoted. The maker's realised margin is an order of magnitude thinner than the spread on the screen. Note that this is the same arithmetic as Section 2.2's decomposition seen from the other side: there, the liquidity supplier kept 0.14 of a tick out of the 1.14 ticks the buyer paid.

**(b) The fragility of that margin.** Hold the day's 22 round trips fixed and let $n$ of them be adverse. Then

$$\Pi(n) = (22 - n) \times \$1.00 - n \times \$5.00 - \$4.40 = \$17.60 - \$6.00\,n,$$

falling by \$6.00 for each additional informed counterparty — \$5.00 of loss plus the \$1.00 of capture that trade did not deliver. The day breaks even at $n^* = 17.60/6.00 = 2.9333$: at $n = 3$ the maker loses \$0.40 and at $n = 4$ he loses \$6.40. *One additional informed fill per day converts this business from profitable to unprofitable.* This is why market makers invest so heavily in classifying flow: knowing which counterparties are informed is worth more than any improvement in speed or fee schedule.

**(c) The term that is not in the table.** Suppose the day's flow was one-sided enough that the maker ends holding 300 units long, marked at \$27.53 — a notional of \$8,259.00. At an assumed daily volatility of 0.9%, a one-standard-deviation overnight move is $8{,}259.00 \times 0.009 = \$74.33$, which is 13.2734 times the entire day's net trading profit. Put the other way: an adverse move of just $\$5.60/(300 \times \$0.01) = 1.8667$ ticks on the residual position erases everything the maker earned. This is why market makers flatten, hedge, or pay up to lay off inventory rather than carry it, and it is the reason the inventory term in Result 2.4 is written separately from the cash terms: it is not a cost the maker pays, it is a risk the maker must not take.

**Trap — The spread is not the market maker's profit**

It is tempting to read a one-tick spread as "the market maker earns one tick per round trip". He does not. In Worked Example 2.4 he earned an eighth of a tick per unit, because two fills in twenty-two were against someone who knew more. The same error, mirrored, is what makes a retail cost model dangerous: the spread you are quoted is not what your trading costs, because your own fills also arrive disproportionately at the wrong moments. Both sides of the market systematically overestimate what the quoted spread tells them — the maker about his revenue, the taker about his cost.

### Skewing: how inventory moves the quote

A maker who is long wants to sell. He cannot refuse to buy more — quotes are firm — but he can make buying less attractive and selling more attractive by moving both quotes down. If the maker's own estimate of fair value is $\hat{m}$ and he holds inventory $y$ (positive long), his quotes take the form

$$P_b = \hat{m} - \frac{s}{2} - \kappa y, \qquad P_a = \hat{m} + \frac{s}{2} - \kappa y, \qquad \kappa > 0,$$

so the entire two-sided quote shifts down by $\kappa y$ while its width is unchanged. The effect is to raise the probability of the trade that flattens him and lower the probability of the trade that extends him, at the cost of accepting a slightly worse average price. Two consequences matter for a systematic trader.

First, **skew makes the quoted mid a biased estimate of fair value**. When a dealer is heavily positioned, the mid on his stream is displaced by his inventory, not by the market's view. In a centralised book with many competing suppliers these displacements largely cancel; on a single dealer's stream — which is what an FX CFD account sees — they do not.

Second, **skew is the mechanism by which flow moves prices without information**. A large

uninformed order that fills against a dealer forces the dealer to skew, which moves the price he shows to the next client, who trades at the displaced level, and so on. This is the microstructure version of order flow having price impact even when it carries no information — the effect documented for FX by Evans and Lyons (2002), where signed interdealer order flow explains a substantial share of daily exchange-rate variation, far more than contemporaneous macroeconomic news does.

### Why quotes widen when volatility rises

Every liquidity supplier widens in volatility, and the toy day shows exactly why. Suppose the same maker faces a volatile session: the adverse jump doubles to $J = 11$ ticks and the number of adverse fills doubles to four, while benign flow stays at twenty round trips. At the unchanged one-tick quote,

$$\Pi = 20 \times 100 \times \$0.01 - 4 \times 100 \times (11 - 0.5) \times \$0.01 - 48 \times \$0.10 = 20.00 - 42.00 - 4.80 = -\$26.80,$$

a heavy loss on a spread that was comfortably profitable an hour earlier. Writing the quote as a half-width $h$ in ticks, the identity of Result 2.4 gives

$$\Pi(h) = 20 \times 100 \times 2h \times \$0.01 - 4 \times 100 \times (11 - h) \times \$0.01 - \$4.80 = \$44.00\,h - \$48.80,$$

which is zero at $h^* = 48.80/44.00 = 1.1091$ ticks — a quoted spread of 2.2182 ticks. The maker must therefore roughly *double* the quoted spread to restore breakeven, and at a three-tick quote the session returns $\$44.00 \times 1.5 - \$48.80 = +\$17.20$. Nothing about the maker's costs or technology changed; only the size of the adverse move did, and the spread tracked it, exactly as Result 2.3 predicts when $\Delta$ rises.

Three practical corollaries follow, and all three show up in the reader's own instruments.

**Spreads widen around scheduled events** because $\Delta$ — the size of the value revision a single release can cause — is known in advance to be large at those moments. This is not the dealer opportunistically taxing news traders; it is the same breakeven arithmetic evaluated at a larger $\Delta$.

**Spreads widen in thin hours** because the maker's expected time to flatten lengthens, raising the inventory term, and because the uninformed flow that funds the business is absent while informed flow is not.

**Spreads widen, and then liquidity disappears, in stress** because widening has a limit. Beyond some point, no spread compensates for the adverse selection and inventory risk on offer, and the rational response is to stop quoting altogether. That is what the withdrawal of liquidity in a disorderly market actually is — not a failure of the mechanism but its correct operation (CFTC-SEC, 2010; Kirilenko, Kyle, Samadi, Tuzun, 2017). It is also why liquidity and funding conditions are not independent: a supplier whose capital or financing is under pressure widens and withdraws for reasons that have nothing to do with the instrument itself, which is the feedback loop Brunnermeier and Pedersen (2009) formalise.

**Your project — Your spread is somebody's breakeven calculation**

The platform's own cost work has already established that OANDA's quoted spreads widen at scheduled news, at the daily rollover, and in thin hours — and treats those widenings as a nuisance to be bucketed. Result 2.3 and this section say something stronger: those widenings are the dealer's estimate of $\mu\Delta$ plus inventory risk, evaluated at that moment, for flow like yours. Two consequences for the cost surface. First, a widening is *information*, not just a cost: a cell of the surface where the spread systematically triples is a cell where the dealer expects large value revisions and adverse flow, which is exactly where a distributional forecast should also be widest — the two objects are estimating related quantities and can be cross-checked. Second, if your fills cluster in the wide cells (because your signal fires on news), your realised average cost is not the surface's unconditional mean but its *flow-weighted*

mean, which can be far higher. Weight every cost cell by the historical frequency with which your own rule actually traded in it before quoting a cost per round trip [Course Book §14.2].

## 2.5 Modern Structure: Fragmentation, Internalisation, and Dealer Streams

Everything so far has assumed one book, in one place, visible to everyone. No major market has looked like that for a long time. Three structural changes separate the textbook book from the one a modern trader actually faces, and a fourth — the one that matters most for this book's reader — removes the book entirely.

### Who supplies liquidity now

The designated specialist with an affirmative obligation has largely been replaced by unobligated electronic market makers competing for queue position. Menkveld (2013) studies one such firm in detail during its entry into a European equity market and documents the economics this chapter has been deriving: the firm traded enormously, held tiny positions for very short intervals, ended almost every day close to flat, and earned a gross spread that was largely consumed by adverse selection — leaving a thin net margin that depended on volume, on venue fee rebates, and on the firm's ability to avoid the informed side of the flow. That is Worked Example 2.4 with more decimal places and far more round trips.

Two features of this population matter for anyone modelling execution. First, the liquidity is *unobligated*: nothing compels these participants to quote, so depth is a function of their current willingness, not of a rulebook. Second, it is *fast*: quotes are updated on every observable change in correlated instruments, which is why a snapshot of the book overstates what a slow participant can actually reach. Both features make displayed depth an upper bound on reachable depth exactly when it matters most.

### Fragmentation

The same instrument now trades simultaneously at several venues. Fragmentation splits a single notional book into pieces, so that the depth available at a given price is the *sum* across venues, reachable only by an order router that can reach all of them within the lifetime of the quotes. The consequences are structural and stable, whatever the venue names in a given jurisdiction:

- **The consolidated picture is a construction.** As Chapter 1 noted, some jurisdictions mandate the computation and dissemination of a consolidated best bid and offer; others leave it to vendors. Either way the aggregate book you see is assembled by a rule, and the rule's latency and coverage determine whether the depth it shows is reachable.
- **Fee structures differ by venue and shape where liquidity rests.** Venues commonly charge takers and rebate makers, or occasionally invert that; the level and even the sign of these fees vary by venue, participant status and jurisdiction, and change over time. Treat any specific schedule as indicative and verify the current specification.
- **Execution quality becomes a routing question.** With depth split across venues, the average price achieved depends on the router's decisions as much as on the aggregate book, which is why measured effective spread — Definition 2.2 — and not quoted spread is the only defensible input to a cost model.

### Internalisation

A large share of retail order flow in some markets never reaches a public book at all. A dealer (variously called a wholesaler or internaliser depending on jurisdiction) executes the order against its own capital at or inside the prevailing public quote, and manages the resulting position itself. The economics follow directly from Section 2.3: retail flow is, on average, far less informed than

the aggregate flow arriving at a public book, so its adverse-selection component $\mu\Delta$ is smaller, so a dealer can profitably quote it a *better* price than the public book shows and still earn more per unit of risk than by quoting the public book. The client receives genuine price improvement; the public book loses the uninformed flow that was funding its spreads, and must widen for everyone else. Whether the net effect is beneficial is a genuine policy question, argued on evidence, and this book takes no position on it; the mechanism, however, is not in dispute, and it follows from arithmetic already derived here. The specific arrangements — what such intermediaries may pay for order flow, what disclosure is required, what execution obligations apply — are jurisdiction-specific, change frequently, and must be verified against current rules rather than assumed from any description, including this one.

### What survives when there is no book at all: dealer streams and FX

Spot foreign exchange has no central limit order book, no consolidated tape, and no closing auction. It is a tiered network: a small number of large dealers quote each other on interdealer platforms (some of which do run genuine order books), and quote clients bilaterally on streams tailored to the client relationship (Lyons, 2001; Rime, Schrimpf, 2013; BIS, 2022, figures indicative and subject to revision at the next survey). What arrives at a retail or small-institutional client is therefore not a market; it is one firm's continuously updated opinion about where it is willing to trade with that particular client, sized for that client's typical clip.

Several properties of this arrangement follow, and every one of them applies to the reader's account.

**There is no queue and no priority.** A limit order placed with a CFD broker does not rest in any book. It is an instruction held by the broker to trade with you when its own quote reaches your price. You have no queue position, cannot be jumped, cannot jump anyone, and gain no execution priority by being early. The strategic content of limit-order placement — the substance of Section 2.1's queue arithmetic — simply does not exist for you.

**There is no depth to walk.** Because the broker quotes a size at a price, an order larger than that size is not walked through visible levels; it is either rejected, filled at a worse dealer-determined price, or filled in parts as the dealer's own quote moves. The cost of size still exists — it has to, since the dealer must lay the position off into a market that does have depth — but it is opaque, delivered as a wider quote or a worse fill rather than as a visible walk.

**The quote can be rejected after you accept it.** Many FX e-trading arrangements permit the liquidity provider a brief final check before confirming a trade against a streamed price. The practice, its permitted duration, and its disclosure obligations vary by provider and jurisdiction and have been the subject of continuing industry work; treat any specific description — including this one — as indicative and verify your own broker's current terms. Structurally, its effect is to make the quoted price an option the client does not fully own, which biases realised execution relative to quoted prices in a way no quoted-spread cost model can capture.

**The mid is a dealer's mid.** As Section 2.4 showed, a dealer's quotes are skewed by the dealer's own inventory. On a single stream there is no cross-sectional averaging to remove that skew, so the mid you record is contaminated by a position you cannot observe.

**Trap — Mid-only candles delete this entire chapter**

A price history stored as mid (or, worse, as a single "close") has thrown away the bid, the ask, and therefore the spread, the imbalance, the microprice, and every quantity this chapter has defined. A backtest run on such a history does not merely underestimate costs — it cannot represent the object that generates them. It will fill you at a price that no counterparty ever offered, at moments when the true spread was several times its median, and it will do so most generously exactly where your signal is most active, because signals fire on news and spreads widen on news. The fix is not a constant cost assumption bolted on afterwards; it is

a bid/ask history, measured per instrument and per condition, as the Playbook's Chapter 2 requires before any cost figure is quoted.

**Your project — What the bid/ask re-pull is actually buying you**

The Playbook's Stage −1a rebuilds the platform's cost surface from a fresh bid/ask candle pull across all 21 instruments rather than from mid candles. This chapter supplies the reason that step is not optional and the list of quantities it makes computable. From bid/ask candles you can measure, per instrument and per conditioning cell: the quoted spread and its distribution (not just its mean — the tail is what your fills see); the spread's behaviour around the rollover hour and scheduled releases, which Section 2.4 shows is the dealer's own $\mu\Delta$ estimate; and, if fill prices are retained, the *effective* half-spread of Definition 2.2 against the decision mid, which is the only number that belongs in an expectancy calculation [Course Book §14.1]. What you still cannot measure from a dealer stream is depth, queue position, or price impact beyond your clip, because none of those objects exists in your feed. Record what you can measure, mark the rest as unmeasured rather than assumed to be zero, and size positions so that the unmeasured part cannot dominate — the execution-realism discipline the Course Book insists on [Course Book §18.3].

## 2.6 Auctions: Opening, Closing, and Uncrossing

Continuous trading is bracketed, at most exchanges, by call auctions: one to start the day and one to end it. The reason is the same in both cases. Continuous matching works well when orders arrive steadily, and badly when a large amount of trading interest arrives simultaneously — at the open, after a night of accumulated news, or at the close, when a great deal of business must transact at one particular price. A call auction handles concentrated interest by refusing to match anything until all of it has been collected, then clearing the whole crossed portion at a single price.

Chapter 1 derived that clearing rule: cumulative demand $D(p)$ and supply $S(p)$, executable volume $V(p) = \min\{D(p), S(p)\}$, and a clearing price $p^*$ maximising $V$. Two layers sit on top of it in any real auction, and they are what this section adds.

**Tie-breaking.** The maximum of $V$ is frequently attained over a range of prices, because between two adjacent submitted limits neither $D$ nor $S$ changes. Venues resolve this by a published cascade, typically: (i) maximise executable volume; (ii) among ties, minimise the residual imbalance $|D(p) - S(p)|$; (iii) among remaining ties, choose the price closest to a reference — usually the last continuous trade or the previous close; and (iv) failing that, apply a final rule such as the midpoint of the tied range. The precise cascade is venue-specific and occasionally revised; verify the current rules for any venue you trade.

**Priority and rationing.** At the clearing price, one side is generally larger than the other, so orders on the heavy side must be rationed. Priority runs: market orders (those with no limit) first, then limit orders strictly better than the clearing price, then — pro rata or by time, depending on the venue — orders whose limit is exactly the clearing price. The last group is where rationing bites, and it is why an order priced exactly *at* the price you expect to clear is not the same order as one priced a tick through it.

Auction-specific order types exist for this purpose: instructions to trade at whatever price the auction sets (commonly named market-on-close or at-the-open), instructions to do so subject to a limit, and, at some venues, instructions that participate only to offset a published imbalance. The naming and availability of these types is venue-specific — verify current specifications rather than assuming any particular label exists at your venue.

During the call phase most venues disseminate an *indicative* clearing price and the current imbalance, updated as orders arrive. This is a genuine, if noisy, public signal: it tells the market

what the auction would clear at if it uncrossed now, and it exists to attract the offsetting orders that would reduce the imbalance.

**Worked Example 2.5 — Uncrossing a closing auction**

The closing auction for the hypothetical share of Table 2.1 receives the following orders. The last continuous trade of the session was \$27.53. All figures are computed exactly for this illustrative scenario.

| Side | Type | Quantity |
|---|---|---|
| Buy | market-on-close | 1,200 |
| Buy | limit \$27.56 | 800 |
| Buy | limit \$27.55 | 1,500 |
| Buy | limit \$27.54 | 600 |
| Sell | market-on-close | 900 |
| Sell | limit \$27.54 | 1,000 |
| Sell | limit \$27.55 | 1,400 |
| Sell | limit \$27.57 | 700 |

**(a) Cumulative demand and supply.** Market orders count at every price. A buy limit participates at prices at or below its limit; a sell limit at prices at or above its limit.

| $p$ | $D(p)$ | $S(p)$ | $V(p) = \min$ | $D - S$ |
|---|---|---|---|---|
| \$27.54 | 4,100 | 1,900 | 1,900 | +2,200 |
| \$27.55 | 3,500 | 3,300 | **3,300** | +200 |
| \$27.56 | 2,000 | 3,300 | 2,000 | −1,300 |
| \$27.57 | 1,200 | 4,000 | 1,200 | −2,800 |

Checking one row in full: at \$27.55, demand is the 1,200 market-on-close buys plus the 800 at \$27.56 (who will happily pay less) plus the 1,500 at \$27.55 (whose limit is exactly met), giving 3,500; the 600 bid at \$27.54 do not participate. Supply is the 900 market-on-close sells plus 1,000 at \$27.54 plus 1,400 at \$27.55, giving 3,300; the 700 offered at \$27.57 are not reached.

**(b) The clearing price.** $V$ is maximised uniquely at $p^* =$ \$27.55, where 3,300 units trade — more than the entire five-level book of Table 2.1 contained on both sides combined. The close prints two ticks above the last continuous trade.

**(c) Rationing the heavy side.** At \$27.55, $D - S = +200$: demand exceeds the available supply by 200 units, so 200 units of buy interest go unfilled. Priority determines whose. The 1,200 market-on-close buys fill first; the 800 bid at \$27.56 fill next, being strictly better than the clearing price; that consumes 2,000 of the 3,300 available, leaving 1,300 for the 1,500 resting exactly at \$27.55 — a fill ratio of $1{,}300/1{,}500 = 86.6667\%$ for that group, with the remaining 200 units unfilled. A buyer who wanted certainty of participation should have used a market-on-close order, or priced through the expected clearing level; a buyer who wanted price protection got it, and paid for it with an incomplete fill — the same trade-off as Worked Example 2.1(f), in a different mechanism.

**(d) A tie, and how it breaks.** Suppose instead that the 1,400-unit sell had been priced at \$27.54 rather than \$27.55. Then $S(27.54) = 900 + 1{,}000 + 1{,}400 = 3{,}300$ as well, and:

| $p$ | $D(p)$ | $S(p)$ | $V(p)$ | $\lvert D-S\rvert$ |
|---|---|---|---|---|
| \$27.54 | 4,100 | 3,300 | **3,300** | 800 |
| \$27.55 | 3,500 | 3,300 | **3,300** | **200** |
| \$27.56 | 2,000 | 3,300 | 2,000 | 1,300 |
| \$27.57 | 1,200 | 4,000 | 1,200 | 2,800 |

Two prices now tie at 3,300 units, and the volume criterion cannot separate them. The second rule — minimum residual imbalance — selects \$27.55, with 200 units of surplus against 800 at \$27.54. Had the imbalances also tied, the third rule would have chosen the candidate closest to the \$27.53 reference, which is \$27.54. Note that the two rules can point in opposite directions, which is exactly why the cascade is published in advance: an unpublished tie-break is an invitation to dispute a price on which an entire day's valuations depend.

### Why the close is the deepest moment of the day

The closing price is not just one more print. It is the price at which a large amount of business is contractually or operationally *required* to happen:

- **Index tracking.** A fund whose mandate is to track an index measured at the close must trade at the close; any other execution price introduces tracking error against its benchmark. When an index changes composition, tracking funds must buy the additions and sell the deletions, in size, at that day's close. The resulting demand is famously price-moving — the evidence that index inclusion moves prices, even absent any change in fundamentals, is precisely the evidence that demand curves for individual stocks are not flat (Shleifer, 1986; Harris, Gurel, 1986), and Chapter 9 returns to it.
- **Valuation.** Funds strike net asset values at closing prices; derivatives settle against them; margin is computed from them. Anyone who must transact at the price used for their own valuation has a mechanical reason to be in the closing auction.
- **Concentration begets concentration.** Because so much size is there, the close is where a large order finds the most natural counterparties, so more large orders go there. The share of daily volume executing in closing auctions has risen materially in many equity markets over recent years; treat any specific percentage as indicative and verify current figures for the venue you care about.

The opening auction is the mirror image and behaves differently. It clears overnight information against a book assembled with no intraday price discovery to guide it, so it is noisier, its indicative price moves more, and the continuous session that follows typically begins with wider spreads and larger price revisions than the rest of the day — the same $\Delta$ argument as Section 2.4, applied to accumulated overnight uncertainty.

Two footnotes for this book's other pillars. First, FX has no auction of either kind: there is no official close, and the benchmark rates that serve a similar administrative purpose are computed as averages over a short window of trades and quotes around a fixed time, not as an uncrossing. Structurally, that difference matters — a windowed average concentrates flow into the window without producing a single clearing price, so participants transacting there face a different problem from auction participants. Chapter 4 takes this up with the FX session structure. Second, the reader's equity CFDs reference exchange closing prices formed by exactly the mechanism above, which means a CFD position is marked and financed against a price the reader can neither participate in nor influence.

## 2.7 Exercises

**Exercise 2.1 — Walking a bid side**

A different hypothetical share trades with a one-cent tick and the following book. All figures are illustrative, computed exactly; none is market data.

| Bid side | | Ask side | |
|---|---|---|---|
| Price | Size | Price | Size |
| 18.42 | 150 | 18.44 | 200 |
| 18.41 | 250 | 18.45 | 350 |
| 18.40 | 600 | 18.46 | 150 |
| 18.39 | 300 | 18.47 | 600 |
| 18.38 | 800 | 18.48 | 250 |

(a) State the quoted spread in dollars and ticks, the mid, the touch imbalance, and the microprice, using Result 2.1 and confirming it against the direct formula.

(b) A market order to sell 500 units arrives. Compute the fills level by level, the total proceeds, and the average fill price.

(c) Compute the slippage against the quoted bid, and the effective half-spread and effective spread against the mid, each in dollars per unit, in ticks, and in total dollars. State the ratio of effective to quoted spread and the effective spread in basis points of the mid.

(d) State the book and mid immediately after the trade, and fit the linear-impact coefficient $\lambda$ of (2.1) to this execution.

(e) Suppose instead the trader had sent a limit order to sell 500 units at \$18.41 or better. State the quantity filled, the average price, and the completion rate, and say in one sentence what the price protection cost.

(f) A cost model estimates this trade's cost as half the quoted spread times the quantity. By what dollar amount, and what percentage of the true cost, does it err?

▷ working space

**Solution — Exercise 2.1**

**(a)** The touch is \$18.42 / \$18.44, so $s = \$0.02$, two ticks, and $m = \$18.43$. The imbalance is $I = (150 - 200)/(150 + 200) = -0.1428571$, so by Result 2.1

$$P_{\text{micro}} = 18.43 + 0.01 \times (-0.1428571) = \$18.4285714,$$

confirmed directly by $(18.42 \times 200 + 18.44 \times 150)/350 = 6{,}450.00/350 = \$18.4285714$. The microprice sits 0.1429 of a tick below the mid: the ask side is heavier, so the weighted price leans down.

**(b)** Cumulative bid depth is 150, 400, 1,000, so the order clears the first two levels and takes 100 units from the third:

| Level | Price | Filled | Proceeds |
|---|---|---|---|
| 1 | \$18.42 | 150 | \$2,763.00 |
| 2 | \$18.41 | 250 | \$4,602.50 |
| 3 | \$18.40 | 100 | \$1,840.00 |
| | | **500** | **\$9,205.50** |

Average fill price $\bar{p} = 9{,}205.50/500 = \$18.4110$.

**(c)** Slippage against the quoted bid: $18.42 - 18.411 = \$0.0090$ per unit, 0.90 of a tick, $500 \times 0.0090 = \$4.50$ in total. Effective half-spread against the mid: $18.43 - 18.411 = \$0.0190$ per unit, 1.90 ticks, \$9.50 in total; the effective spread is \$0.0380, 3.80 ticks, which is $3.80/2.00 = 1.90$ times the quoted spread. In relative terms, $10^4 \times 0.019/18.43 = 10.3093$ basis points one way and 20.6186 basis points round trip.

**(d)** Levels one and two are consumed and $600 - 100 = 500$ units remain at \$18.40, so the new touch is \$18.40 / \$18.44 — four ticks, double its former width — and the new mid is \$18.42, one tick below the old one. Fitting (2.1) in its selling form $\bar{p} \approx m - s/2 - (\lambda/2)Q$:

$$\lambda = \frac{2\left(m - \frac{s}{2} - \bar{p}\right)}{Q} = \frac{2(18.43 - 0.01 - 18.411)}{500} = \frac{2 \times 0.009}{500} = 3.60 \times 10^{-5},$$

so $1/\lambda = 27{,}777.78$ units per dollar of displacement. Check: $18.42 - \frac{1}{2}(3.6 \times 10^{-5})(500) = 18.42 - 0.009 = \$18.4110$, as computed.

**(e)** A limit at \$18.41 reaches only the first two levels: 400 units for proceeds of \$7,365.50, an average of \$18.4137500, a completion rate of $400/500 = 80\%$, with 100 units unfilled. The protection improved the average price by $18.41375 - 18.4110 = \$0.00275$ per unit on the units that traded, and cost the trader an unexecuted fifth of the order — which is expensive precisely in the states where the bid keeps falling and the residual has to be sold lower.

**(f)** The naive estimate is $(s/2)Q = 0.01 \times 500 = \$5.00$ against a true cost of \$9.50: an error of \$4.50, understating the true cost by $4.50/9.50 = 47.3684\%$. Nearly half the cost of this trade is invisible to a quoted-spread model, and the missing half is the part that grows with size.

**Exercise 2.2 — Roll's estimator on eight trades**

Eight consecutive transaction prices in a hypothetical share with a one-cent tick are

$$41.22,\ 41.24,\ 41.22,\ 41.23,\ 41.21,\ 41.23,\ 41.22,\ 41.24.$$

(a) Compute the seven price changes and the six lagged products, and state $\hat{\gamma}_1$ using the estimator of Result 2.3.

(b) Compute $\hat{s}_{\text{Roll}}$ in dollars, in ticks, and in basis points of a \$41.22 reference mid. If the

quoted spread over the window averaged four ticks, what fraction of it did trades actually pay?

(c) Recompute $\hat{\gamma}_1$ and $\hat{s}_{\text{Roll}}$ on demeaned price changes and state the difference in ticks.

(d) A ninth trade prints at \$41.28. Recompute $\hat{\gamma}_1$ and $\hat{s}_{\text{Roll}}$ on all nine prices, and state the change in the estimate in ticks and as a percentage. What does this tell you about applying Roll's estimator to short windows, or to any window containing a trend?

▷ working space

**Solution — Exercise 2.2**

**(a)** The changes are

$$+0.02,\ -0.02,\ +0.01,\ -0.02,\ +0.02,\ -0.01,\ +0.02,$$

and the six lagged products are

$$-0.0004,\ -0.0002,\ -0.0002,\ -0.0004,\ -0.0002,\ -0.0002,$$

every one negative — an unusually clean bid-ask bounce. Their sum is $-0.0016$, so $\hat{\gamma}_1 = -0.0016/6 = -0.0002666667$.
**(b)** $\hat{s}_{\text{Roll}} = 2\sqrt{0.0002666667} = 2 \times 0.0163299316 = \$0.0326598632$, that is 3.2660 ticks, or $10^4 \times 0.03266/41.22 = 7.9233$ basis points. Against a four-tick quoted spread this is $0.0326598632/0.04 = 81.6497\%$: trades paid about four-fifths of the quoted spread, the balance representing executions inside the quotes.
**(c)** The mean change is $+0.02/7 = \$0.0028571$. Demeaning before multiplying gives $\hat{\gamma}_1 = -0.0002585034$ and $\hat{s}_{\text{Roll}} = \$0.0321560819$, or 3.2156 ticks — 0.0504 of a tick below the raw figure. The window drifts upward only slightly, so the correction is small; it would not be on a window containing a genuine trend.
**(d)** The ninth price adds an eighth change of $+\$0.04$ and therefore a seventh product, formed from the last two changes $+0.02$ and $+0.04$: $(0.02)(0.04) = +0.0008$, positive because the price trended rather than bounced. The sum of products becomes $-0.0016 + 0.0008 = -0.0008$ over seven products, so $\hat{\gamma}_1 = -0.0001142857$ and

$$\hat{s}_{\text{Roll}} = 2\sqrt{0.0001142857} = \$0.0213808994 = 2.1381 \text{ ticks},$$

a fall of 1.1279 ticks, or 34.5346% of the original estimate, caused by a single additional print. The lesson is that Roll's estimator is a *variance-like* statistic computed from very few effective observations: it is extremely noisy in short windows, and any trend — or any run of same-signed trades, which order splitting guarantees — biases $\hat{\gamma}_1$ upward towards zero and the spread estimate downward, eventually making it undefined. Use it as a cross-check on directly measured spreads, never as a substitute where quote data exist.

**Exercise 2.3 — A Glosten–Milgrom quote with an asymmetric prior**

A dealer makes a market in a hypothetical share ahead of a binary outcome. The value will be either $V_H = \$12.00$ or $V_L = \$11.00$, and the public prior is $\pi = \mathbb{P}(V = V_H) = 0.60$. A fraction $\mu = 0.30$ of arriving traders are informed and trade in the direction of the truth; the rest buy or sell with probability $\frac{1}{2}$ each. Order-processing cost is $c = \$0.02$ per unit per side.

(a) Compute $\mathbb{P}(\text{buy})$, $\mathbb{P}(V_H \mid \text{buy})$ and the zero-profit ask; then $\mathbb{P}(\text{sell})$, $\mathbb{P}(V_H \mid \text{sell})$ and the bid. State the adverse-selection spread.

(b) Compare it with the symmetric-prior benchmark $\mu\Delta$ of Result 2.3, and explain in one or two sentences why the asymmetric prior gives the smaller number.

(c) Compute the quote midpoint and compare it with the prior mean $\mathbb{E}[V]$. Explain the sign of the difference.

(d) State the quoted bid, ask and spread once the processing cost is included, and the percentage of the quoted spread attributable to adverse selection.

(e) Verify the zero-profit condition conditional on a buy arriving, by computing $\mathbb{P}(\text{informed} \mid \text{buy})$ and the dealer's gain and loss on the two cases.

(f) A competing dealer quotes the same instrument at a spread of \$0.42 with the same $c$.

Using the symmetric-prior formula as an approximation, what informed fraction does that quote imply?

▷ working space

**Solution — Exercise 2.3**

Throughout, $\Delta = V_H - V_L = \$1.00$ and $\mathbb{E}[V] = 0.6 \times 12.00 + 0.4 \times 11.00 = \$11.60$.
**(a)** $\mathbb{P}(\text{buy} \mid V_H) = \mu + \frac{1-\mu}{2} = 0.30 + 0.35 = 0.65$ and $\mathbb{P}(\text{buy} \mid V_L) = \frac{1-\mu}{2} = 0.35$, so

$$\mathbb{P}(\text{buy}) = 0.6 \times 0.65 + 0.4 \times 0.35 = 0.39 + 0.14 = 0.53, \qquad \mathbb{P}(V_H \mid \text{buy}) = \frac{0.39}{0.53} = 0.7358491.$$

Hence $P_a = 11.00 + 1.00 \times 0.7358491 = \$11.7358491$. Symmetrically $\mathbb{P}(\text{sell}) = 0.6 \times 0.35 + 0.4 \times 0.65 = 0.21 + 0.26 = 0.47$ and $\mathbb{P}(V_H \mid \text{sell}) = 0.21/0.47 = 0.4468085$, giving $P_b = \$11.4468085$. The adverse-selection spread is

$$s_{\text{adv}} = 11.7358491 - 11.4468085 = \$0.2890405.$$

**(b)** The symmetric benchmark is $\mu\Delta = 0.30 \times 1.00 = \$0.30$, which exceeds the answer by $\$0.0109595$, or 3.6532%. A trade's informational content is largest when the market is most uncertain: at $\pi = \frac{1}{2}$ each trade shifts the posterior by the most, so the two conditional means are furthest apart and the spread is widest. With $\pi = 0.60$ the market already leans towards $V_H$, a buy confirms what is partly known, and the update — hence the spread — is smaller.
**(c)** The quote midpoint is $(11.7358491 + 11.4468085)/2 = \$11.5913288$, which sits $\$0.0086712$ *below* the prior mean of $11.60. The asymmetry arises because a buy is more likely than a sell (0.53 against 0.47) when the high state is more likely; the ask must therefore be set for a more probable, less informative event and the bid for a less probable, more informative one, and the two do not straddle $\mathbb{E}[V]$ symmetrically. The dealer is still unconditionally fair — each side has zero expected profit given the trade that arrives — but the midpoint of a Glosten–Milgrom quote is not an unbiased estimate of value except in the symmetric case.
**(d)** Adding $c = \$0.02$ per side gives a quoted ask of $11.7558491, a quoted bid of $11.4268085 and a quoted spread of $0.2890405 + 0.04 = \$0.3290405$. Adverse selection accounts for $0.2890405/0.3290405 = 87.8434\%$ of the quoted spread and processing costs for 12.1566%.
**(e)** $\mathbb{P}(\text{informed} \mid \text{buy}) = \mu\pi/\mathbb{P}(\text{buy}) = 0.30 \times 0.60/0.53 = 0.3396226$. On an informed buy the dealer sells at $11.7358491 something worth $12.00, losing $0.2641509; on an uninformed buy he sells at $11.7358491 something worth $\mathbb{E}[V] = \$11.60$, gaining $0.1358491. Expected profit:

$$0.3396226 \times (-0.2641509) + 0.6603774 \times (+0.1358491) = -0.0897124 + 0.0897124 = \$0.00.$$

**(f)** Using $s = \mu\Delta + 2c$ with $s = \$0.42$, $c = \$0.02$, $\Delta = \$1.00$:

$$\mu = \frac{0.42 - 0.04}{1.00} = 0.38.$$

The competitor's wider quote implies it believes more than a third of arriving flow is informed — eight percentage points more than our dealer assumes. Both cannot be right; the quote is a testable belief about the composition of the flow, and whichever dealer is wrong will discover it in the P&L identity of Result 2.4.

**Exercise 2.4 — A market maker's day in a volatile session**

The maker of Worked Example 2.4 faces a volatile session and quotes two ticks wide (half-width $h = 1$ tick) in clips of 100 units, with processing costs of $0.10 per execution. He completes 18 round trips: 14 benign, and 4 on which fair value jumps $J = 9$ ticks against the position before he flattens.

(a) Compute gross spread capture, the adverse-selection loss, processing costs, and the day's net P&L. State the net per round trip.

(b) Holding the flow counts fixed, express net P&L as a function of the half-width $h$ in ticks and solve for the breakeven quote. What is the net at a three-tick quote?

(c) Widening drives away price-sensitive uninformed flow but not informed flow. Recompute the net at a three-tick quote if benign round trips fall from 14 to 10, and again if they fall to 7. What does the comparison with (b) imply about the maker's correct response to a genuine information event?

(d) Returning to the two-tick quote and holding total round trips at 18, express net P&L as a function of the number $n$ of adverse fills and find the breakeven $n$.

▷ working space

**Solution — Exercise 2.4**

**(a)** A benign round trip earns the full two-tick spread; an adverse fill loses $J - h = 9 - 1 = 8$ ticks per unit — the jump net of the one tick of edge the maker was paid on entry. With 36 executions:

$$\begin{aligned} \text{capture} &= 14 \times 100 \times \$0.02 = \$28.00, \\ \text{adverse} &= 4 \times 100 \times \$0.08 = \$32.00, \\ \text{processing} &= 36 \times \$0.10 = \$3.60, \\ \Pi &= 28.00 - 32.00 - 3.60 = -\$7.60, \end{aligned}$$

a loss of $\$7.60/18 = \$0.4222$ per round trip. The maker is quoting twice as wide as in Worked Example 2.4 and losing money, because the adverse component has grown faster than the quote.

**(b)** With half-width $h$ ticks, capture is $14 \times 100 \times 2h \times \$0.01 = \$28.00h$ and the adverse loss is $4 \times 100 \times (9 - h) \times \$0.01 = \$36.00 - \$4.00h$, so

$$\Pi(h) = 28.00h - (36.00 - 4.00h) - 3.60 = \$32.00\,h - \$39.60,$$

zero at $h^* = 39.60/32.00 = 1.2375$ ticks, a quoted spread of 2.4750 ticks. Since quotes must sit on the tick grid, the maker's narrowest viable quote is three ticks, at which $\Pi = 32.00 \times 1.5 - 39.60 = +\$8.40$.

**(c)** At a three-tick quote with only 10 benign round trips, capture is $10 \times 100 \times \$0.03 = \$30.00$, the adverse loss is $4 \times 100 \times (9 - 1.5) \times \$0.01 = \$30.00$, processing is $28 \times \$0.10 = \$2.80$, and $\Pi = -\$2.80$. With 7 benign round trips capture is \$21.00, adverse loss is again \$30.00, processing is $22 \times \$0.10 = \$2.20$, and $\Pi = -\$11.20$. Widening therefore restores profitability only if the benign flow is inelastic; if uninformed traders withdraw faster than the spread widens, every wider quote is worse than the last, because the maker is left facing a flow whose informed share has *risen*. The correct response in that case is not a wider quote but a smaller one — less size, or no quote at all — which is exactly what liquidity withdrawal in a stressed market consists of, and why depth can vanish without any participant behaving irrationally.

**(d)** With 18 round trips of which $n$ are adverse, at $h = 1$ tick,

$$\Pi(n) = (18 - n) \times \$2.00 - n \times \$8.00 - \$3.60 = \$32.40 - \$10.00\,n,$$

so $n^* = 32.40/10.00 = 3.24$. At $n = 3$ the day earns \$2.40 and at $n = 4$ it loses \$7.60, reproducing (a). The maker's viability turns on a single fill either way — which is the practical content of Result 2.4: in market making, the variance of the outcome lives entirely in the count of informed counterparties, not in the volume traded.

**Exercise 2.5 — What a quoted-spread cost surface cannot see**

The platform's cost surface is to be built from measured bid/ask candles, which yield the *quoted* spread per instrument and conditioning cell. Suppose instead it were built only from quoted spreads and used directly as the cost input to Stage 0's expectancy test.

(a) List the distinct execution phenomena, each named and defined using this chapter's vocabulary, that a quoted-spread cost model omits. For each, state its sign — whether it makes true costs higher or lower than the model predicts — and say whether it can be measured from a dealer stream at all.

(b) Suppose a measured cell shows a quoted spread of 1.2 pips while the effective spread

actually paid on fills in that cell is 1.9 pips. For a rule trading eight round trips a day over 250 trading days at a clip of 100,000 units of a pair whose pip value is $10 per 100,000, quantify the annual error.

(c) State what you would record, per fill, to make the effective spread measurable going forward, and how you would treat the components that remain unmeasurable.

▷ working space

**Solution — Exercise 2.5**

**(a)** Six omissions, with signs.

*1. Size slippage (walking the book).* Any order larger than the displayed size at the touch fills at an average price worse than the quote (Result 2.2), and the cost grows quadratically in size by (2.1). Sign: true cost *higher*. On a dealer stream this is not directly measurable — there is no visible depth — but it is partially observable as the difference between the quote at decision time and the fill price actually received, provided both are recorded.

*2. Delay slippage.* The market moves between decision and execution. Sign: on average adverse for a signal-driven strategy, because signals fire when the price is moving. Measurable if the decision timestamp and quote are stored alongside the fill.

*3. Adverse selection on the trader's own fills.* Fills cluster at moments when the quoting side is willing to trade — disproportionately when the price is about to move against the taker (Definition 2.2's price-impact term). Sign: higher. Measurable ex post as the mid displacement over a fixed horizon after each fill.

*4. Fill probability on resting orders.* A limit order is filled preferentially in the states where its price was wrong; unfilled orders are not costless, since the intended trade is missing when it would have been profitable — which is exactly what implementation shortfall counts (Perold, 1988). Sign: higher, and invisible to any model built on filled trades alone. Measurable only if unfilled and cancelled orders are logged.

*5. Conditional spread widening.* The quoted spread in a cell is a distribution, not a number; a rule that trades on news trades in the right tail of that distribution (Section 2.4's $\mu\Delta$ argument). Sign: higher, and specifically higher for this reader's strategies. Measurable, and the fix is flow-weighting rather than averaging.

*6. Rejections and requoting.* Where the provider may decline after the client accepts, the realised distribution of fills is truncated adversely. Sign: higher. Measurable only if rejected attempts are logged — most retail systems discard them.

Only the first two are commonly modelled; items 3–6 are the ones that turn a marginally profitable backtest into a losing live system, and all four are recoverable from execution logs the platform is free to keep.

**(b)** The understatement is $1.9 - 1.2 = 0.7$ pips per round trip, which is $0.7/1.9 = 36.8421\%$ of the true cost. At $8 \times 250 = 2{,}000$ round trips a year that is $0.7 \times 2{,}000 = 1{,}400$ pips, and at \$10 per pip per 100,000 units,

$$1{,}400 \times \$10 = \$14{,}000 \text{ per year per 100{,}000-unit clip,}$$

an error larger than the entire expected profit of most marginal rules. Note that the error is a *bias*, not noise: it does not average out with more backtest data, and it grows linearly with turnover, so it penalises exactly the high-frequency variants a search over strategies tends to favour.

**(c)** Record, per fill: the timestamp and the full bid/ask quote at the moment the decision was made; the timestamp, price and size of each partial fill; the bid/ask quote at a fixed horizon after the last fill (one minute and five minutes, say); and, for orders that did not fully execute, the quantity unfilled and the reason. From these, the effective half-spread $(p - m_0)$, the price impact $(m_\tau - m_0)$ and the realised half-spread $(p - m_\tau)$ of Definition 2.2 follow directly, per fill, and can be aggregated into the same cells as the quoted-spread surface so that the two can be compared cell by cell. For what remains unmeasurable — depth beyond the broker's clip, queue position, the counterfactual price of an order never sent — do not impute zero. Carry it as an explicit haircut on expectancy, sized from the gap between quoted and effective spreads where both are observed, and require any rule to clear its hurdle net of that haircut before capital is committed [Course Book §14.7].

**Key idea — What to remember**

A limit order book is a queue of promises sorted by price and then by time, and everything a market order costs can be read off it by hand: walk the levels outward from the touch, weight the fills by quantity, and the average price you obtain is the true cost of your own size — a cost that exceeded the quoted spread by a factor of 2.28 in this chapter's central example, and whose size-dependent term grew quadratically while the spread term did not. The quoted spread, the slippage, and the effective spread are three different numbers, and only the last belongs in an expectancy calculation; it splits, trade by trade and without assumptions, into a realised half-spread that the liquidity supplier keeps and a price impact that he does not, which is why a market maker quoting one tick can earn an eighth of one. That spread exists because supplying liquidity costs something to process, risks something to hold, and loses something to counterparties who know more: the third component is the Glosten–Milgrom breakeven $\mu\Delta$, so a spread is a readable, invertible estimate of how likely the market thinks you are to be informed, and it widens at news because $\Delta$ widens, not because anyone is being opportunistic. Roll's estimator recovers the same effective spread from transaction prices alone via the negative autocovariance that the bid-ask bounce induces, which is worth knowing precisely because it fails — loudly, by going undefined — whenever trades trend or order splitting makes signs run. None of this machinery is visible in the reader's own feed: an FX CFD account sees one dealer's inventory-skewed quote with no book, no queue and no depth behind it, and a mid-only price history deletes even the spread, which is why the Playbook's bid/ask re-pull is a structural requirement rather than a refinement. Chapter 3 takes the three spread definitions introduced here and turns them into a measurement discipline: the full cost taxonomy, explicit and implicit, and the accounting that makes a backtest's cost line something a live account can actually reproduce.

# 3

# Transaction Costs, Measured Honestly

*A strategy's backtest and its account statement disagree, and the difference is not noise: it is a bill, itemised, with a name and a payee against every line. This chapter reads that bill. It separates the costs that print on a statement — commission, exchange fees, overnight financing — from the far larger ones that never do, and then makes the implicit ones measurable: the three spread definitions a microstructure researcher uses (quoted, effective, realised), the decomposition that turns their difference into an estimate of what the trader's own information cost him, and Perold's implementation shortfall, which prices in one number everything that happened between deciding to trade and finishing the trade. It states the square-root impact law that institutional execution desks live by, and then argues, with arithmetic, that at retail CFD size the law predicts an impact roughly one hundred and thirty times smaller than the spread — so that a retail cost model built on impact theory is calibrating the wrong thing entirely. What replaces it is a measured cost surface, session by session, plus the financing arithmetic that can quietly consume eight times the spread on a position held a month. The chapter closes on the argument that gives it its title: every technique here is market microstructure — the same machinery Chapter 2 pointed at dealers and order books — turned around and pointed at yourself, and it is the only part of microstructure a retail trader is uniquely well placed to measure, because he alone has the fills.*

## 3.1 The Cost Taxonomy

The first discipline of cost accounting is refusing to net things that have different causes. A single round trip in a single instrument can lose money to at least six mechanisms that are economically distinct, are paid to different counterparties, respond to different controls, and scale differently with size, holding period and market conditions. Adding them into one number called "slippage" and calibrating that number from a backtest is the standard retail practice, and it is the reason a book with a real edge can run for a year and end flat: an aggregate cannot tell you which lever to pull.

The primary split is between costs that appear on a statement and costs that do not.

**Explicit costs** are billed. They arrive as line items, are known in advance to within a published schedule, and can be reconciled exactly against a broker's confirmations. Commission (per share, per lot, or as a percentage of notional); exchange, clearing and regulatory fees on listed products; the market-data subscriptions that make a strategy runnable at all; stamp duties and financial-transaction taxes where a jurisdiction levies them; and, for any leveraged or derivative position carried past a daily cut-off, *financing* — the overnight interest charge or credit that Section 3.5 derives in full. Explicit costs are the easy half, and because they are the easy half they receive most of the attention that ought to go elsewhere.

**Implicit costs** are never billed, never itemised, and never appear in any document the broker sends. They are the difference between the price a trade could theoretically have been done at and the price it was actually done at, and they are extracted through the price itself, which is precisely why they are invisible. There are five, and keeping them apart is the whole of this chapter's method:

- **Spread** — the cost of demanding immediacy. A trader who must transact now buys at the ask and sells at the bid, and pays the difference between that price and the market's consensus. Chapter 2 derived where the spread comes from (order processing, inventory risk, adverse selection); Section 3.2 makes precise what fraction of it a given trade actually paid.
- **Market impact** — the movement in price *caused by the trade itself*. Buying consumes resting sell interest and shifts the book upwards; the later shares of a large order are bought at prices the order created. Impact splits into a temporary component that decays once the order stops, and a permanent one that does not (Section 3.3).
- **Slippage** — the movement in the quote between the instant an order is sent and the instant it fills, for reasons that have nothing to do with the order. On a retail platform this includes the latency of the round trip to the broker, requoting, and the re-pricing of the dealer's stream when the market moves in the interval.
- **Delay cost** — the movement between the moment a decision was *made* and the moment the order reached the market. A signal that fires on the close of an M30 bar and is acted on forty seconds later has already conceded whatever the price did in those forty seconds. This is the line that a backtest assuming fills at the signal bar's close silently sets to zero [Course Book §18.1.9].
- **Opportunity cost** — the profit forgone on the part of an intended position that was never established, because the price ran away, the order was cancelled, or the liquidity was not there. It is a real cost of a real decision, it is invisible in any record of executed trades, and Section 3.3 shows it can be a third of the total.

| Cost | On a statement? | Paid to | Principal driver |
|---|---|---|---|
| Commission & fees | Yes | Broker, venue, regulator | Trade count, notional |
| Financing / swap | Yes | Broker (rate differential ± markup) | Holding period, leverage |
| Spread | No | Liquidity supplier | Immediacy demanded; session, volatility, news |
| Impact | No | The resting orders consumed | Size relative to available depth |
| Slippage | No | Whoever was quoting when the market moved | Latency, quote volatility |
| Delay | No | Nobody — the price simply moved | Decision-to-order lag |
| Opportunity | No | Nobody — the trade did not happen | Fill rate, order aggressiveness |

**Table 3.1.** The cost taxonomy. The bottom five rows are invisible to accounting and, for most systematic books, dominate the top two.

Table 3.1 makes the structural point that governs the rest of the chapter: the costs that are easiest to know are the ones that matter least, and vice versa. This is not an accident of bookkeeping. Explicit costs are contractual, so a competitive market pushes them towards the cost of provision; implicit costs are informational, arising from the fact that trading reveals something and consumes something, and no amount of competition removes them. It is also why the spread is not merely a nuisance to be assumed away in a backtest: it is a priced characteristic of an asset, with wider-spread assets commanding higher expected returns in equilibrium precisely because investors demand compensation for the cost of getting in and out (Amihud, Mendelson, 1986).

### The half-spread convention, and how to avoid double-counting

More cost accounting is wrecked by this one convention than by any other single error, so it is worth stating carefully and once.

Let $b$ and $a$ be the best bid and ask, and let the *mid* be $m = (a + b)/2$. The *quoted spread* is $s = a - b$. A single trade that crosses the quote — a market buy filled at $a$, or a market sell filled at $b$ — transacts at a distance $s/2$ from the mid, not $s$. That distance, the *half-spread*, is the cost of one fill. A complete round trip, in and out, crosses twice and therefore pays $s/2 + s/2 = s$: one full quoted spread, if the quote is unchanged between the two legs.

So the quoted spread is a *round-trip* number and the half-spread is a *per-fill* number, and the two are routinely confused in both directions. A backtest that deducts a full spread on entry and another on exit has doubled the true cost; one that deducts a half-spread per round trip has halved it. The convention adopted throughout this chapter, and the one used in the microstructure literature (Harris, 2003; Hasbrouck, 2007), is to quote all spread measures on the round-trip scale — so that "an effective spread of 1.2 pips" means a round trip costs 1.2 pips and a single fill costs 0.6 — and to say "half-spread" explicitly whenever the per-fill number is meant.

**Trap — Deducting the spread twice**

The most common cost-model error is not underestimating the spread. It is applying it on the wrong scale — a full quoted spread charged on each of two legs, or a half-spread charged once for a round trip. The first makes every strategy look worse than it is by a factor of two and kills good research; the second makes every strategy look better by the same factor and kills accounts. Fix the convention once, write it down next to the cost function, and test it: simulate a round trip that opens and closes at unchanged quotes and confirm the modelled loss equals exactly one quoted spread.

### Round-trip accounting

Assemble the pieces into the identity that every later section refines. Consider a position of $N$ units, side $q = +1$ for long and $q = -1$ for short, opened at fill price $p_1$ when the mid was $m_1$, closed at fill price $p_2$ when the mid was $m_2$, and held for $H$ days. Write $c$ for commission per side (in currency, per unit) and $f$ for financing per unit per day (signed: negative when charged). The realised profit is

$$\underbrace{qN(p_2 - p_1)}_{\text{realised}} = \underbrace{qN(m_2 - m_1)}_{\text{paper, mid-to-mid}} - \underbrace{N\big[\,q(p_1 - m_1) \;+\; q(m_2 - p_2)\,\big]}_{\text{the two half-spreads paid}} \tag{3.1}$$

and the account's profit subtracts the explicit lines on top:

$$\Pi \;=\; qN(m_2 - m_1) \;-\; \tfrac{N}{2}\big(s_1^{\text{eff}} + s_2^{\text{eff}}\big) \;-\; 2cN \;+\; fNH, \tag{3.2}$$

where $s_k^{\text{eff}} = 2q_k(p_k - m_k)$ is the effective spread of leg $k$ — Section 3.2's central object, defined so that halving it gives the per-fill cost in accordance with the convention just fixed, and $q_k$ is the side of that individual leg ($q_1 = q$ on entry, $q_2 = -q$ on exit). Every symbol has a home: $qN(m_2 - m_1)$ is what a backtest using mid-price candles reports; the second term is the spread bill; the third is commission; the fourth is financing, which can be either sign.

**Result — The cost decomposition, in one line**

Backtest P&L minus account P&L, for a completed round trip, equals

$$\underbrace{\tfrac{N}{2}\big(s_1^{\text{eff}} + s_2^{\text{eff}}\big)}_{\text{spread}} + \underbrace{2cN}_{\text{commission}} - \underbrace{fNH}_{\text{financing}} + \underbrace{qN(m_1 - m_0)}_{\text{delay}} + \underbrace{(\text{unfilled}) \times (m_T - m_0)q}_{\text{opportunity}},$$

where $m_0$ is the mid at the moment of decision and $m_T$ the mid at the horizon over which the forgone position would have been held. The first three lines are the round trip actually done; the last two are the trades not done as intended, and they are exactly the lines Section 3.3 shows a fills-only record cannot see.

The Course Book states the same accounting from the position-sizing side, where the object of interest is the no-trade band it induces: a signal must clear the round-trip cost before acting on it is rational at all, which converts a continuous forecast into a discrete decision with a dead zone around zero [Course Book §1.6], and makes cost a first-class term in the objective rather than a post-hoc deduction [Course Book §14.2]. What that treatment takes as given — a number for the cost — is what this chapter produces.

## 3.2 Quoted, Effective, and Realised Spread

"The spread" is three different quantities that happen to share a name, and a cost model that uses the wrong one is wrong by a factor that this section will show can exceed four. The three are distinguished by *when* they are measured and *whose* money they describe. The quoted spread is an advertisement, observable before any trade occurs. The effective spread is a receipt, observable at the instant of a trade. The realised spread is a profit-and-loss statement, observable only after the market has had time to move. Each answers a different question, and their differences carry the information.

### Quoted spread: the advertised price of immediacy

The quoted spread $s^{\mathrm{q}} = a - b$ is the cheapest number to obtain and the least informative. It is a standing offer: the price at which the liquidity supplier will trade a *specified size* right now. Three things it does not tell you. It says nothing about the size available at those prices — a one-pip quote good for 50,000 units and a one-pip quote good for 5,000,000 units are the same number and radically different markets. It says nothing about whether the quote will survive the arrival of an order, which on a fast-moving screen is precisely when it matters. And it says nothing about what a trade actually transacted at, because trades can occur inside the quoted spread (price improvement) or, on a stale or requoted stream, outside it.

Its virtue is that it can be sampled without trading, at whatever frequency the feed allows, in unlimited quantity and at zero cost — which, as Section 3.6 argues, makes it the only spread measure a small account can afford to estimate densely.

For cross-instrument comparison the quoted spread is normally expressed in relative terms, $s^{\mathrm{q}}/m$, in basis points of the mid, since a 1.2-pip spread on EURUSD near 1.08 and a 1.2-pip spread on GBPUSD near 1.27 are not the same proportional cost.

### Effective spread: what the trade actually cost

**Definition — Effective spread**

For a trade executed at price $p$ at time $t$, with the prevailing mid $m_t$ and direction indicator $q = +1$ if the trade was buyer-initiated (the liquidity demander bought) and $q = -1$ if seller-initiated, the **effective spread** is

$$s^{\mathrm{eff}} = 2\,q\,(p - m_t),$$

and its relative form is $s^{\mathrm{eff}}/m_t$. The factor of two puts it on the round-trip scale of the quoted spread: the cost borne by the liquidity demander on this single fill is the half-effective spread, $q(p - m_t) = s^{\mathrm{eff}}/2$.

Three properties follow directly. When a trade executes exactly at the quote — a buy at $a$, a sell at $b$ — the effective spread equals the quoted spread, since $a - m = m - b = s^{\mathrm{q}}/2$. When it executes inside the quote, the effective spread is smaller, and the gap $s^{\mathrm{q}} - s^{\mathrm{eff}}$ is the price improvement received. And when it executes outside the recorded quote — because the quote had already moved, or a dealer requoted — the effective spread exceeds the quoted spread, which is the arithmetic signature of slippage as Section 3.1 defined it. The three cases are why $s^{\mathrm{eff}}$, not $s^{\mathrm{q}}$, is the number a cost model needs: it is the only one of the two that describes money that actually changed hands.

The absolute-value form $s^{\mathrm{eff}} = 2|p - m_t|$ is equivalent whenever the trade is on the correct side of the mid, and is the form usually quoted, but it discards the information that makes the next definition work, and it silently converts a fill on the wrong side of the mid — a buy below the mid, which does happen — into a positive cost when it was in fact a subsidy. Keep the signed form.

### Realised spread: what the liquidity supplier kept

The effective spread measures the demander's cost at the moment of trading, but it is not the supplier's profit, because the supplier is left holding a position whose value is about to change. If the buyer knew something, the mid will drift up after the trade, and the dealer who sold has lost on inventory exactly what he gained on the spread. To capture this, measure the same quantity again against a *later* mid.

**Definition — Realised spread**

For the same trade, and a horizon $\Delta$ chosen in advance, the **realised spread** is

$$s^{\mathrm{real}} \;=\; 2\,q\,(p - m_{t+\Delta}),$$

using the same direction indicator $q$ as the effective spread. It is positive when the liquidity supplier is still ahead $\Delta$ later — the portion of the demander's payment that the supplier retained — and negative when the post-trade drift has cost the supplier more than the spread earned.

Subtracting the two definitions gives an identity, not an approximation:

**Result — The spread decomposition**

$$s^{\mathrm{eff}} \;-\; s^{\mathrm{real}} \;=\; 2q(p - m_t) - 2q(p - m_{t+\Delta}) \;=\; 2q\,(m_{t+\Delta} - m_t) \;\equiv\; \text{price impact},$$

so that

$$\underbrace{s^{\mathrm{eff}}}_{\text{charged to the taker}} \;=\; \underbrace{s^{\mathrm{real}}}_{\text{kept by the maker}} \;+\; \underbrace{2q(m_{t+\Delta} - m_t)}_{\text{adverse selection}}.$$

The middle term measures the permanent revision of the market's consensus in the direction of the trade — which is to say, how much the trade revealed. It is Glosten and Milgrom's (1985) adverse-selection component of the spread, measured directly from a tape rather than inferred from a model, and Chapter 2's third spread component made observable.

Two modelling choices are buried in that identity and should be surfaced. The first is $\Delta$. Too short and the mid has not yet finished incorporating the trade's information, so adverse selection is understated; too long and unrelated price movement swamps the signal, so the estimator becomes noisy without becoming less biased. Five minutes is the conventional equity choice and is used throughout this chapter; the honest practice is to compute the decomposition at two or three horizons and report whether the conclusion is stable, because it sometimes is not (Hasbrouck,

2007). The second is $q$. On your own trades you know it exactly — you know which side you were on — which is a genuine and under-appreciated advantage of measuring your own costs rather than estimating a market's. On a public tape, $q$ must be inferred from the trade's position relative to the prevailing quote, and misclassification biases every quantity above.

**Worked Example 3.1 — Quoted, effective and realised spread on a six-trade tape**

The table records six hypothetical EURUSD trades from a single London morning: the prevailing bid and ask, the trade price $p$, the direction $q$, and the mid five minutes later. All figures are illustrative, constructed to exhibit the full range of behaviour; they are not market data. Compute, for each trade and on average, the quoted, effective and realised spreads and the implied adverse selection, in pips (1 pip $= 0.0001$).

| # | bid | ask | $p$ | $q$ | $m_t$ | $m_{t+5}$ | $s^{\mathrm{q}}$ | $s^{\mathrm{eff}}$ | $s^{\mathrm{real}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.08420 | 1.08434 | 1.08434 | $+1$ | 1.08427 | 1.08431 | 1.4 | 1.4 | 0.6 |
| 2 | 1.08429 | 1.08445 | 1.08429 | $-1$ | 1.08437 | 1.08433 | 1.6 | 1.6 | 0.8 |
| 3 | 1.08431 | 1.08443 | 1.08440 | $+1$ | 1.08437 | 1.08446 | 1.2 | 0.6 | $-1.2$ |
| 4 | 1.08450 | 1.08462 | 1.08450 | $-1$ | 1.08456 | 1.08449 | 1.2 | 1.2 | $-0.2$ |
| 5 | 1.08444 | 1.08462 | 1.08462 | $+1$ | 1.08453 | 1.08452 | 1.8 | 1.8 | 2.0 |
| 6 | 1.08438 | 1.08450 | 1.08441 | $-1$ | 1.08444 | 1.08440 | 1.2 | 0.6 | $-0.2$ |
| mean | | | | | | | 1.400 | 1.200 | 0.300 |

**Trade 1, in full.** The mid is $m_t = (1.08420 + 1.08434)/2 = 1.08427$, so the quoted spread is $1.08434 - 1.08420 = 0.00014 = 1.4$ pips. The buy filled at the ask, so

$$s^{\mathrm{eff}} = 2(+1)(1.08434 - 1.08427) = 2 \times 0.00007 = 0.00014 = 1.4 \text{ pips},$$

equal to the quoted spread, as it must be for a fill at the quote. Five minutes later the mid is 1.08431, four tenths of a pip higher:

$$s^{\mathrm{real}} = 2(+1)(1.08434 - 1.08431) = 2 \times 0.00003 = 0.00006 = 0.6 \text{ pips},$$

and the adverse-selection component is $1.4 - 0.6 = 0.8$ pips, which the identity confirms independently as $2(+1)(1.08431 - 1.08427) = 0.8$ pips.

**Trade 3, the interesting one.** The buy filled at 1.08440, three tenths of a pip inside the 1.08443 ask — price improvement of 0.3 pips — so the effective spread, 0.6 pips, is half the quoted 1.2. But the mid then rose nine tenths of a pip to 1.08446, above the trade price, giving

$$s^{\mathrm{real}} = 2(+1)(1.08440 - 1.08446) = -0.00012 = -1.2 \text{ pips}.$$

The liquidity supplier collected 0.6 pips of spread and immediately lost 1.8 pips to the subsequent drift: a net loss of 1.2 pips on the round-trip scale. This is what being adversely selected looks like in a single row.

**Trade 5, the mirror image.** Here the quote was wide (1.8 pips — the stream had backed away) and the buyer paid all of it, but the mid then *fell* a tenth of a pip, so the supplier kept more than he charged: $s^{\mathrm{real}} = 2(1.08462 - 1.08452) = 2.0$ pips against $s^{\mathrm{eff}} = 1.8$, and adverse selection is $-0.2$ pips. Uninformed flow at a wide quote is the market maker's best trade.

**Averages.** Over the six trades the mean quoted spread is 1.400 pips, the mean effective spread 1.200 pips, and the mean realised spread 0.300 pips, so mean adverse selection is $1.200 - 0.300 = 0.900$ pips — the identity holding in aggregate as it does trade by trade. In relative terms the mean effective spread is 1.10657 basis points of the mid.

**What the three numbers say.** Takers paid 85.71% of the advertised spread (1.200/1.400), the shortfall being price improvement on trades 3 and 6. Of what they paid, the supplier retained just 25.00% (0.300/1.200); the remaining 75.00% was handed straight back to the

market through post-trade drift. On one standard lot (100,000 units, pip value \$10.00), the mean cost of a single fill is the half-effective spread, 0.60 pips = \$6.00, and of a complete round trip 1.20 pips = \$12.00. (All figures verified in `verify_mk03.py`, §1.)

That 75% adverse-selection share is the single most useful diagnostic in the table, and it cuts both ways. Read from the supplier's side it explains why market making is not the free money the quoted spread suggests, and why dealers widen when they suspect informed flow — Chapter 2's inventory and information arguments, now with a number attached. Read from the *taker's* side, which is the reader's side, it says something sharper: on this tape, three-quarters of what takers paid was not a fee at all but a measurement of how right they turned out to be, five minutes later, on average. A trader whose fills systematically show *low* adverse selection — whose trades are followed by no drift in his own direction — is paying the spread and buying nothing with it. That is not a cost problem. It is an alpha problem wearing a cost problem's clothes, and Section 3.6 builds the measurement that tells the two apart.

## 3.3 Price Impact and Implementation Shortfall

Section 3.2 measured the cost of a single fill. A position of any size is not a single fill, and the costs that dominate institutional trading arise precisely from that fact: an order large enough to matter changes the price it is trying to obtain. This section develops the two ideas that follow — impact, and the accounting framework that prices *everything* that happens between deciding and finishing — and then, deliberately, dismantles the first one's relevance to a retail account, because pretending otherwise is the most consequential form of dishonesty available in this part of the subject.

### Temporary and permanent impact

Impact splits along the same seam as the spread decomposition, and for the same reason.

*Temporary impact* is the price concession required to persuade someone to take the other side right now. It is compensation for immediacy and inventory risk, it is bounded by the depth available, and it decays once the order stops: the book refills, the dealer lays off, and the price returns to where the fundamentals put it. *Permanent impact* is the revision of the market's consensus caused by the information the order revealed. It does not decay, because nothing has been learned that then gets unlearned.

That is exactly Result 3.2, one level up. The realised spread is the single-trade analogue of temporary impact — what the supplier keeps once the dust settles — and the adverse-selection term is the single-trade analogue of permanent impact. A trader who understands the six-row table understands impact; what changes with size is only the magnitude and the fact that the trade must be sliced.

Kyle's (1985) model gives the canonical formalisation: with a single informed trader, noise traders, and a competitive risk-neutral market maker who sets prices from observed net order flow, the equilibrium pricing rule is linear,

$$\Delta p \;=\; \lambda\, x,$$

where $x$ is net signed order flow and $\lambda$, "Kyle's lambda", is the inverse of market depth — the price move per unit of flow. The result matters here less as a formula to calibrate than as a statement of what impact *is* in equilibrium: not a friction, but the market's rational inference from the fact that someone wanted to trade. It follows that impact cannot be arbitraged away, and that a strategy which reliably avoids impact is a strategy whose trades carry no information.

### The square-root law, and what it is a law of

Empirically, the aggregate impact of a large order worked over time is markedly *concave* in size rather than linear: doubling the order does not double the impact. The regularity most widely used on execution desks summarises this as

$$\frac{\Delta P}{P} \approx Y \sigma_{\text{daily}} \sqrt{\frac{Q}{V}}, \tag{3.3}$$

where $Q$ is the total size of the parent order, $V$ is the instrument's average daily volume in the same units, $\sigma_{\text{daily}}$ is its daily volatility, and $Y$ is a dimensionless prefactor of order one — values around 0.5 to 1 are commonly quoted, and it is instrument-, venue- and period-dependent. The ratio $Q/V$ is the order's *participation rate* in the day's volume.

It is essential to be precise about the epistemic status of (3.3). It is not a theorem, and it is not derived here because it is not derived anywhere from primitives that everyone accepts; it is an *empirical regularity*, fitted across large samples of institutional parent orders, and it is quoted in this book with three explicit caveats. First, its calibration domain is institutional metaorders in liquid equities and futures, typically at participation rates from a fraction of a percent to a few percent of daily volume; nothing in the evidence base licenses extrapolation many orders of magnitude below that range. Second, it prices *impact only* — it is stated net of the spread, and adding the two is the correct treatment, not double-counting. Third, $Y$ is a fitted constant whose value varies enough between studies and markets that the formula is an order-of-magnitude instrument, not a precision one. All numerical uses below take $Y = 0.5$ and are labelled illustrative.

### Implementation shortfall

Impact, slippage, delay and opportunity cost are all measured against different reference prices, and the temptation to report each separately against whichever benchmark flatters it is overwhelming. Perold's (1988) answer was to define one number, anchored to the only moment that is not chosen after the fact.

**Definition — Implementation shortfall**

Construct two portfolios. The **paper portfolio** executes the full intended trade, instantly and costlessly, at the *decision price* $P_d$ — the price prevailing at the moment the decision to trade was made. The **real portfolio** is what actually happened: the fills that occurred, at the prices they occurred at, in the quantities they occurred in, net of explicit costs, with the unfilled remainder simply absent. Marking both to the same terminal price $P_T$, the **implementation shortfall** is

$$\text{IS} = \Pi_{\text{paper}} - \Pi_{\text{real}}.$$

It is a single number, in currency, that captures every cost of turning a decision into a position — including the cost of the part of the position never established.

Two features make it the right definition rather than merely a definition. It is *benchmark-honest*: $P_d$ is fixed before any trading occurs and cannot be gamed by a trader choosing when to start, unlike a VWAP benchmark, which rewards a trader for tracking whatever the day happens to do. And it is *complete*: because the paper portfolio holds the full intended size, an order that is abandoned when the price runs away contributes its full forgone profit to the shortfall, which no measure computed from executed fills can do.

The shortfall then decomposes. Insert the *arrival price* $P_a$ — the price at the instant the order actually reached the market — and the four components separate cleanly:

**Result — Four-way decomposition of implementation shortfall**

For a buy programme of intended size $Q$, executed size $Q_e = \sum_i n_i$ at fill prices $p_i$, unfilled $Q_u = Q - Q_e$, decision price $P_d$, arrival price $P_a$, terminal price $P_T$ and total commissions $C$:

$$\text{IS} = \underbrace{Q_e(P_a - P_d)}_{\text{delay}} + \underbrace{\textstyle\sum_i n_i(p_i - P_a)}_{\text{execution vs arrival}} + \underbrace{Q_u(P_T - P_d)}_{\text{opportunity}} + \underbrace{C}_{\text{explicit}}.$$

For a sell programme every price difference reverses sign. The first two terms sum to the execution cost measured against the decision price, $\sum_i n_i(p_i - P_d)$; splitting them at $P_a$ is what separates *the market moved before we could act* from *our own trading moved the market*.

**Worked Example 3.2 — Implementation shortfall, decomposed**

A portfolio manager decides at 09:30 to buy 100,000 shares of a stock trading at $P_d = \$40.00$. Order construction and routing take until 10:15, by which time the price is $P_a = \$40.04$. The order is worked through the day and fills in three tranches: 40,000 at \$40.06, 30,000 at \$40.12, and 20,000 at \$40.19. The remaining 10,000 shares are never filled; the order is cancelled at the close, with the stock at $P_T = \$40.35$. Commission is \$0.005 per share on executed volume. (A hypothetical scenario, computed exactly.)

**Step 1: the two portfolios.** The paper portfolio buys all 100,000 shares at \$40.00 and marks them at \$40.35:

$$\Pi_{\text{paper}} = 100{,}000 \times (40.35 - 40.00) = \$35{,}000.00.$$

The real portfolio paid $40{,}000 \times 40.06 + 30{,}000 \times 40.12 + 20{,}000 \times 40.19 = 1{,}602{,}400 + 1{,}203{,}600 + 803{,}800 = \$3{,}609{,}800.00$ for 90,000 shares (an execution VWAP of \$40.108889), plus commission $0.005 \times 90{,}000 = \$450.00$, and those shares mark at $90{,}000 \times 40.35 = \$3{,}631{,}500.00$:

$$\Pi_{\text{real}} = 3{,}631{,}500.00 - 3{,}609{,}800.00 - 450.00 = \$21{,}250.00.$$

The implementation shortfall is $35{,}000.00 - 21{,}250.00 = \$13{,}750.00$.

**Step 2: the decomposition.** Delay — the 4 cents the market moved before the order arrived, on the shares that eventually traded:

$$90{,}000 \times (40.04 - 40.00) = \$3{,}600.00.$$

Execution against arrival — what the trading itself cost, tranche by tranche:

$$40{,}000(0.02) + 30{,}000(0.08) + 20{,}000(0.15) = 800 + 2{,}400 + 3{,}000 = \$6{,}200.00,$$

the rising per-tranche concession (2, 8, 15 cents) being the visible signature of a book being walked up. Opportunity — the profit forgone on the 10,000 shares never bought:

$$10{,}000 \times (40.35 - 40.00) = \$3{,}500.00.$$

Commission, \$450.00. The four sum to $3{,}600 + 6{,}200 + 3{,}500 + 450 = \$13{,}750.00$, matching Step 1 exactly.

**Step 3: in basis points.** Against the decision value $100{,}000 \times 40.00 = \$4{,}000{,}000$:

| Component | Currency | Basis points |
|---|---|---|
| Delay | \$3,600.00 | 9.000 |
| Execution vs arrival | \$6,200.00 | 15.500 |
| Opportunity | \$3,500.00 | 8.750 |
| Commission | \$450.00 | 1.125 |
| Implementation shortfall | \$13,750.00 | 34.375 |

**Step 4: was the execution bad?** Apply (3.3) with $Y = 0.5$, a daily volatility of 1.80% and an average daily volume of 5,000,000 shares. The participation rate is $100{,}000/5{,}000{,}000 = 2.00\%$, so $\sqrt{Q/V} = 0.14142136$ and

$$\frac{\Delta P}{P} \approx 0.5 \times 0.0180 \times 0.14142136 = 0.00127279 = 12.73 \text{ bp},$$

about \$5,091 on the decision value. The desk actually paid 15.50 bp against arrival — 1.22 times the modelled impact, which for a single order worked against a rising market is unremarkable. What is *not* unremarkable is the rest of the bill: commission, the only line anyone was invoiced for, is 1.125 bp of a 34.375 bp total, i.e. 3.3% of the cost of the decision. (All figures verified in `verify_mk03.py`, §2–§3.)

The decomposition is not merely descriptive; it assigns blame to the right desk. Delay is a workflow cost, owned by whoever sits between the signal and the order. Execution against arrival is the trader's cost, and the only one an execution algorithm can improve. Opportunity cost is an *aggressiveness* cost, owned by whoever set the limit price — and it is worth noticing that it moves in the opposite direction to the execution cost. Trade more passively and the execution line falls while the opportunity line rises; trade more aggressively and the reverse. A desk optimising only the line it is measured on will drive the other one up, which is the standard pathology of single-benchmark trading and the reason the shortfall is defined as the sum.

Almgren and Chriss (2001) formalise exactly this tension for the trading trajectory itself. Trading a large order quickly concentrates impact but leaves little time for the price to move for unrelated reasons; trading it slowly spreads the impact but leaves the unexecuted remainder exposed to volatility for longer. With impact costs increasing in the trading rate and volatility risk increasing in the elapsed time, the two effects define an efficient frontier of expected cost against cost variance, and the optimal trajectory is determined by how much variance the trader will tolerate — there is no single "cheapest" schedule independent of risk preference. That framing, rather than any particular closed-form trajectory, is what carries over: execution is a portfolio problem, not a minimisation.

**Trap — The square-root law, applied where it does not live**

Take (3.3) to a retail FX trade and watch it collapse. One standard lot of EURUSD at 1.0844 is a notional of \$108,440. EURUSD spot turnover is, indicatively, of the order of a trillion dollars a day (as of the mid-2020s; verify against the current BIS survey). Daily volatility is, indicatively, around 0.5%. Then $Q/V \approx 1.0844 \times 10^{-7}$, $\sqrt{Q/V} = 3.293 \times 10^{-4}$, and

$$\frac{\Delta P}{P} \approx 0.5 \times 0.005 \times 3.293 \times 10^{-4} = 8.23 \times 10^{-7} = 0.00823 \text{ bp} = 0.0089 \text{ pips},$$

about one-hundredth of a pip. The round-trip effective spread in Worked Example 3.1 was 1.2 pips — roughly 134 times larger. To generate one pip of modelled impact you would need to trade about \$1.36 billion, some 12,547 standard lots, in a single order.

The trap is not that the law gives a small number. It is what a modeller does next. A cost model built around impact will conclude that a retail order is costless, tune its slippage parameter to near zero, and pass a backtest — while the actual costs, which are spread, spread widening, requoting and financing, are mechanisms (3.3) does not model at all and was never fitted to. The law is not wrong at retail scale. It is irrelevant there, and irrelevance imported into a cost model as a small number is indistinguishable, in a backtest, from a discovery.

## 3.4 Costs at Retail CFD Scale

Everything in Section 3.3 was built for a participant who moves the market. The reader does not. Three things change when the size falls by four orders of magnitude, and the change is not a matter of degree in any of them.

**Impact disappears, and takes the entire optimal-execution literature with it.** The trapbox above did the arithmetic: modelled impact at one standard lot is a hundredth of a pip. A retail order does not walk the book, does not need slicing, does not face an Almgren–Chriss trade-off, and gains nothing from a participation-rate schedule. This is genuinely good news, and it is routinely converted into bad news by cost models that, having correctly concluded that impact is negligible, conclude that *execution* is negligible.

**The spread stops being a parameter and becomes a process.** At institutional scale the quoted spread on a liquid instrument is a small, stable fraction of the total cost, and treating it as a constant is a defensible approximation. At retail scale it is essentially the *whole* cost, and it is not constant: it is a dealer's decision variable, re-set continuously in response to exactly the inventory and adverse-selection pressures Chapter 2 derived. It widens in thin sessions, widens sharply around scheduled data releases, widens at the daily rollover when the dealer's own funding rolls, and widens in stress — and each of those is a time when a systematic strategy's signals are more likely to fire, not less. Assuming an average spread when the strategy trades non-uniformly across the conditions that determine the spread is not an approximation; it is a bias, and it always runs in the optimistic direction.

**Financing arrives as a first-order term.** A leveraged position carried overnight pays or receives interest on the full notional, every calendar day, while the spread is paid twice per round trip regardless of duration. Section 3.5 shows the crossover is fast: on plausible mid-2020s rate differentials, a month of financing on the wrong side of a carry can cost eight times the round-trip spread. For any strategy holding longer than a day or so, financing is the largest single line on the bill and is completely absent from a backtest run on price data alone.

Underneath all three is the structural fact Chapter 1 established and Chapter 4 detailed: a CFD trade is a bilateral contract with one dealer, and the price is that dealer's quote, not a consensus discovered in a public book. The spread is therefore not a market outcome the reader observes but a price he is quoted, and the correct mental model is a supplier's price schedule — conditioned on time, instrument, volatility and, in a dealer's own risk management, on the client's observed behaviour. That is why it can be *measured* but not *derived*, and why the only defensible retail cost model is an empirical one.

**Your project — The cost surface is the retail cost model**

The platform's replacement of mid-only candles with a measured bid/ask cost surface is the direct consequence of this section. A single scalar "spread assumption" is the wrong object; the right one is a function

$$\widehat{c} \;:\; \big(\text{instrument},\ \text{session bucket},\ \text{volatility regime},\ \text{news proximity}\big) \longmapsto \text{distribution of round-trip effective spread},$$

estimated from the broker's own stored bid and ask, and from the account's own fills wherever they exist. Three design rules follow from this chapter. Store the *quote*, not the mid: the mid is a derived quantity and the spread cannot be recovered from a series of mids, which is what makes mid-only candles a permanent, unrecoverable loss of exactly the object being modelled. Estimate a *distribution* per cell, not a mean: the spread's conditional distribution is right-skewed, and a strategy that trades in the tail of the cell — into a release, at the rollover instant — pays the tail, not the mean. And apply the surface *point-in-time*: the cost charged to a backtested trade at timestamp $t$ must come from the cell as it would have

been estimated using data available before $t$, or the cost model itself becomes a leak [Course Book §18.3].

### What the bill actually looks like

The arithmetic below is the chapter's centre of gravity. It is a hypothetical but fully specified scenario, computed exactly, and its conclusion is the reason the chapter exists.

**Worked Example 3.3 — The annual cost bill of an M30 book**

A systematic M30 EURUSD strategy runs on a \$25,000 account, trading 20,000 units (0.2 standard lots, pip value \$2.00) per position, taking 750 round trips a year. Backtested on mid-price candles, it shows a gross expectancy of 1.9 pips per round trip. Its entries fall into three session buckets, and the platform's cost surface reports the following mean round-trip effective spreads for them (illustrative cell values, of the order a retail FX account would measure in the mid-2020s; verify against your own surface):

| Session bucket | $s^{\text{eff}}$ (pips) | Share (mix A) | Contribution |
|---|---|---|---|
| London–New York overlap | 1.1 | 30% | 0.330 |
| Single-session London or New York | 1.5 | 45% | 0.675 |
| Asia and late New York (thin) | 2.6 | 25% | 0.650 |
| Expected round-trip spread | | 100% | **1.655** |

**Rung 1: the book as it trades.** Annual gross P&L is $750 \times 1.9 \times \$2.00 = \$2{,}850.00$, an apparently healthy 11.40% of equity. The spread bill is

$$750 \times 1.655 \times \$2.00 = \$2{,}482.50,$$

which is 87.11% of the gross edge, leaving \$367.50. Then add financing. Twenty per cent of the trades — 150 of them — are open across the 17:00 New York rollover. The notional is $20{,}000 \times 1.0850 = \$21{,}700$; on a long EURUSD position with indicative mid-2020s rates of 3.25% (EUR) and 5.35% (USD) and a 1.00% broker markup, the daily charge is $\$21{,}700 \times 0.0310/365 = \$1.843014$, or 0.921507 pips. One roll in five lands on the Wednesday triple (Section 3.5), so the 150 rolls carry $150 \times (0.8 \times 1 + 0.2 \times 3) = 210.0$ days of financing:

$$210.0 \times \$1.843014 = \$387.03.$$

The net result is

$$\$2{,}850.00 - \$2{,}482.50 - \$387.03 = -\$19.53,$$

that is, −0.08% of equity. A strategy with a real, positive, mid-to-mid forecasting edge of nearly two pips a trade is, after costs, a machine for converting that edge into broker revenue.
**Rung 2: re-time, do not retune.** Suppose the entry logic is gated to shift the session mix to 45% overlap, 45% single-session, 10% thin — a scheduling change, with no modification to the model, the features, or the signal threshold. The expected round-trip spread falls to

$$0.45(1.1) + 0.45(1.5) + 0.10(2.6) = 1.430 \text{ pips},$$

the spread bill to $750 \times 1.430 \times \$2.00 = \$2{,}145.00$, and the net, financing unchanged, to $\$2{,}850.00 - \$2{,}145.00 - \$387.03 = \$317.97$, or +1.27% of equity.
**Rung 3: stop paying for time you do not use.** Now flatten before the rollover wherever the signal permits, cutting the share of positions held across it from 20% to 5%. The day-count falls to $37.5 \times 1.4 = 52.5$ days, the financing bill to $52.5 \times \$1.843014 = \$96.76$, and the net to

$$\$2{,}850.00 - \$2{,}145.00 - \$96.76 = \$608.24,$$

or +2.43% of equity.
**The ladder.** −0.08% → +1.27% → +2.43% of equity per year, with the forecasting model untouched at every step. (All figures verified in `verify_mk03.py`, §4.)

Two conclusions follow, and they are the reason this chapter sits where it does in the book.

The first is arithmetic. Divide the financing bill by the trade count: \$387.03/750 = \$0.516 per trade, which at \$2.00 per pip is 0.2580 pips. Add it to the expected spread and the *breakeven gross edge* is

$$1.655 + 0.2580 \;=\; 1.9130 \text{ pips per round trip,}$$

against an actual gross edge of 1.9000. The strategy is 0.0130 pips short — roughly one seventy-seventh of a pip — and that gap, multiplied by 750 trades and \$2.00, is the \$19.53 annual loss. The Course Book states this as the no-trade band [Course Book §1.6]: a forecast must clear the round-trip cost before acting on it has positive expectancy at all, and the band here is 1.91 pips wide. Everything the model does inside that band is work performed for the broker.

The second is about hit rates, and it is the diagnosis of a specific failure this platform has already lived through. Consider the same book run with a $+5/-7$ pip target-and-stop structure and a 62% hit rate — a figure that looks, and feels, like a working system. Its gross expectancy is

$$0.62 \times 5 - 0.38 \times 7 = 3.10 - 2.66 = 0.44 \text{ pips,}$$

against the 1.9130-pip breakeven: a net of −1.4730 pips per trade, or −\$2,209.53 a year, −8.84% of equity. To break even it would need a hit rate of

$$p^{*} = \frac{1.9130 + 7}{5 + 7} = 74.28\%,$$

against 58.33% at zero cost — so costs alone move the required hit rate by 15.94 percentage points. A trader who reports "I win 62% of my trades" has said nothing about whether he makes money, and this is the hit-rate illusion the Course Book dissects in full [Course Book §14.1]. The vocabulary of this chapter states its cause precisely: with an asymmetric payoff structure, the gross edge per trade is small in the same units as the round-trip effective spread, and a book that trades often enough to accumulate a convincing hit rate trades often enough to accumulate a fatal spread bill.

## 3.5 Financing and Carry Costs

A CFD position is leveraged by construction: the client posts margin, not notional, and the broker carries the economic exposure of the full position size. Somebody is funding the difference, and the charge for that funding is levied daily, for as long as the position is open. It is the one implicit-feeling cost that is in fact fully explicit — it appears on the statement, it is computed from a published formula, and it can be predicted to the cent before a trade is placed — and it is, for anything held longer than a session, the largest number in this chapter.

### Where the number comes from

An FX CFD is economically a pair of money-market positions. A long EURUSD position is long a EUR deposit and short a USD borrowing of equal value: the holder should earn the euro rate on the notional and pay the dollar rate on it. That difference is the same object Chapter 4's FX swap prices and Chapter 5 derives from covered interest parity — the forward points are the interest differential, expressed in price terms — and a retail financing charge is that differential handed to the client with a markup attached. It is not a separate invention of the CFD industry; it is the swap market's own arithmetic, retailed.

**Definition — CFD overnight financing**

For a position of $N$ units of the base currency at spot $S$ (notional $NS$ in the quote currency), with annualised base and quote rates $r_b$ and $r_q$, a broker's annualised administration markup $\alpha > 0$, and a day-count basis $D$ (conventionally 365 or 360 depending on the currency), the financing accrued per day is

$$f_{\text{long}} = \frac{NS\left[(r_b - r_q) - \alpha\right]}{D}, \qquad f_{\text{short}} = \frac{NS\left[(r_q - r_b) - \alpha\right]}{D},$$

positive meaning a credit to the client and negative a charge. The markup enters with the same negative sign on both sides: it is charged, never paid. Brokers publish their own exact formulation — markup applied per leg rather than to the net, currency-specific day-counts, tiered rates — so these expressions are the structure, and the reader must verify the current specification of his own account before using them.

The sign symmetry is worth pausing on, because it identifies what business the broker is in. Adding the two:

$$f_{\text{long}} + f_{\text{short}} = -\frac{2\alpha NS}{D},$$

independent of the rate differential entirely. A client holding both sides of the same position simultaneously earns and pays the differential exactly once each, netting to zero, and is left paying $2\alpha$ per year on the notional. The financing book is a spread business, exactly like the quoting book, and $\alpha$ is its spread.

### When the charge is applied, and why Wednesday is different

Financing is applied at a fixed daily cut-off — conventionally 17:00 New York, the same instant Chapter 4 identified as the FX day's rollover — and its day-count follows the spot market's *value date* convention, not the calendar. Spot FX settles T+2. Rolling a position past the cut-off therefore moves its value date forward by two business days from the new trade date, and the number of calendar days between the old and new value dates is what gets financed:

| Roll at 17:00 NY on | Old value date | New value date | Days financed |
|---|---|---|---|
| Monday | Wednesday | Thursday | 1 |
| Tuesday | Thursday | Friday | 1 |
| Wednesday | Friday | Monday | **3** |
| Thursday | Monday | Tuesday | 1 |
| Friday | Tuesday | Wednesday | 1 |

Wednesday's roll spans the weekend in value-date space, because Saturday and Sunday are not business days: Thursday plus two business days is Monday. Hence the "triple swap Wednesday" a retail platform's schedule advertises. The arithmetic that matters is the total: $1+1+3+1+1 = 7$ days per week, exactly the calendar. *The triple redistributes financing; it does not add any.* A trader who avoids the Wednesday roll and holds continuously through the rest of the week has not saved anything; a trader who is flat across the Wednesday roll and long the other four days genuinely pays four days rather than seven. Public holidays shift the pattern and can create double or quadruple rolls; these vary by currency and year and must be checked against the broker's published schedule rather than assumed.

**Worked Example 3.4 — A thirty-day financing bill, both directions**

A trader holds one standard lot ($N = 100{,}000$ AUD) of AUDUSD for thirty calendar days. Spot is $S = 0.6510$, so the notional is \$65,100. Take indicative mid-2020s policy-linked rates

of $r_{\text{AUD}} = 4.10\%$ and $r_{\text{USD}} = 5.35\%$, a broker markup $\alpha = 1.00\%$, and $D = 365$. The pip value is $0.0001 \times 100{,}000 = \$10.00$. (Illustrative scenario; rates and markups drift and must be re-verified.)

**The daily rates.** For the long, $(r_b - r_q) - \alpha = (0.0410 - 0.0535) - 0.0100 = -0.0225$; for the short, $(r_q - r_b) - \alpha = 0.0125 - 0.0100 = +0.0025$. So

$$f_{\text{long}} = \frac{65{,}100 \times (-0.0225)}{365} = -\$4.013014 \text{ per day},$$
$$f_{\text{short}} = \frac{65{,}100 \times 0.0025}{365} = +\$0.445890 \text{ per day}.$$

**Thirty days.** The long pays $30 \times 4.013014 = \$120.39$, which at \$10.00 per pip is 12.0390 pips. The short receives $30 \times 0.445890 = \$13.38$, or 1.3377 pips. The swing between the two sides is \$133.77, or 13.3767 pips, on a position whose round-trip spread might be 1.5 pips: *the direction of the carry is worth roughly nine times the cost of trading it*.

**Against the spread.** At an assumed round-trip effective spread of 1.5 pips (\$15.00), thirty days of long financing is 8.03 times the entire cost of entering and exiting the position. Any cost model that charges the spread carefully and omits financing is not approximately right for a position held a month; it is wrong by an order of magnitude.

**The markup's share.** The markup component alone is $65{,}100 \times 0.0100/365 = \$1.783562$ per day, which is 44.44% of the long's total charge. On the short side it is more striking: the gross differential would credit \$2.229452 per day, and the markup takes 80.00% of it, leaving \$0.445890. Holding both sides costs \$3.567123 per day — exactly $2\alpha NS/D$ — or \$1,302.00 a year, 2.00% of notional.

**The flip point.** Suppose the strategy's thesis is a 15-pip move on the long side, with a 1.5-pip round-trip spread. Net profit as a function of holding period $H$ in days is

$$\Pi(H) = 15.0 - 1.5 - 0.4013014\,H \text{ pips},$$

since $\$4.013014/\$10.00 = 0.4013014$ pips per day. At $H = 30$ this leaves 1.4610 pips (\$14.61) — a 15-pip forecast that was exactly right delivers under a pip and a half. Setting $\Pi(H) = 0$,

$$H^* = \frac{15.0 - 1.5}{0.4013014} = 33.64 \text{ days}.$$

Beyond thirty-four days the trade loses money *having been completely correct about the price*. (All figures verified in `verify_mk03.py`, §6.)

The flip point is the concept to carry away, because it converts financing from an accounting nuisance into a design constraint. Every leveraged strategy on the wrong side of a carry has a maximum holding period beyond which its own thesis cannot pay for the position, and that period is computable in advance from three numbers the trader already knows: the target in pips, the spread, and the daily financing rate. A signal generator with no holding-period discipline will occasionally hold past $H^*$, and those trades are guaranteed losers regardless of how well the model forecast.

The asymmetry runs the other way too, and is easy to under-exploit. On the short side of the same pair the financing is a credit: the position is *paid* to stay open, and time works for the trade rather than against it. This is not a free lunch — it is the carry trade, and Chapter 5 shows what it costs in crash risk — but it is a genuine and computable difference between the two sides of the identical instrument, and a strategy that treats long and short symmetrically has mis-specified its own P&L.

Two structural notes complete the picture, both stated as structure and not as specification.

Non-FX CFDs are financed on the same principle but with a different rate: an index, commodity or share CFD carries a benchmark cash rate plus a markup on the long side, with the short side receiving the benchmark less a markup, so that unlike FX there is typically no rate differential to make one direction cheap — both sides tend to be charged, and long-side financing on an index CFD is a persistent headwind for a buy-and-hold position. And share and index CFDs additionally pass through corporate actions: a dividend adjustment credits the long and debits the short on the ex-date, offsetting the mechanical price drop rather than creating profit, which is the mechanism Chapters 8 and 9 treat in full and which a P&L attribution must separate from genuine price return.

## 3.6 Measuring Your Own Costs

Everything above is a definition until something is logged. This section states the protocol, and then confronts the fact that a retail account cannot afford to estimate the whole cost surface from its own fills — which turns out to determine the design.

### The per-trade record

Nothing in Sections 3.2–3.5 can be reconstructed after the fact from a broker's trade confirmations, because confirmations record the fill and nothing else. The quantities that matter are all differences between the fill and something that existed at another moment, and if that other moment was not recorded when it happened, it is gone. The required record, per leg of every trade, is:

- **Decision:** timestamp $t_0$ at which the signal fired, and the mid $m_0$ at $t_0$. This is the anchor for delay cost and for implementation shortfall, and it is the field most often missing, because it belongs to the strategy process rather than the broker's.
- **Arrival:** timestamp $t_a$ at which the order was accepted, with the full quote $(b_a, a_a)$ at that instant — both sides, not the mid.
- **Fill:** timestamp $t$, price $p$, size, side $q$, order type (market, limit, stop, stop-out), and the quote $(b_t, a_t)$ prevailing at $t$.
- **Post-trade:** the mid $m_{t+\Delta}$ at two or three horizons — say one, five and thirty minutes — recorded by a background job, not reconstructed later from candles.
- **Context:** instrument, session bucket on the wall clock, prevailing volatility regime, and the time to and from the nearest scheduled release.
- **Explicit:** commission charged, and financing accrued while the position was open, taken from the statement rather than modelled.

From those fields every quantity in this chapter falls out by arithmetic: the half-effective spread $q(p - m_t)$ paid on each leg, the effective spread $2q(p - m_t)$, the realised spread $2q(p - m_{t+\Delta})$, the adverse-selection component as their difference, the delay cost $q(m_a - m_0)$, and the round-trip shortfall against $m_0$ including financing. Nothing here requires a model. It requires a schema and the discipline to write to it before the trade rather than after it.

**Your project — The measurement protocol, and what it is measuring**

This is the platform's cost-surface calibration, and it is worth being explicit about the estimand, because there is a subtlety that makes the exercise both easier and narrower than it looks. The fills in any cell are not a random sample of that cell's trading conditions: they are conditioned on the strategy's signal having fired, and the signal fires more often when volatility is elevated, which is also when spreads widen. So the mean effective spread computed from your own fills is *not* an estimate of the typical cost of trading that instrument in that session. It is an estimate of the cost of *your* trades in that cell — which is precisely

the number the cost model needs and precisely the wrong number to quote as a description of the market. Record the distinction in the surface's documentation, because the two will diverge and someone will eventually compare them.

Two further design points follow from the fields above. Aggregate by *order type* as well as session: a stop-out is a market order sent into a market already moving against the position and will show the worst effective spreads in the book, while a resting take-profit shows the best — and averaging them produces a number describing no order the strategy actually sends. And record *both* legs separately: entry and exit cost profiles differ systematically, and a single round-trip average conceals which one is the problem.

### How much data this needs, and why the surface must be layered

The per-cell estimate is a sample mean, so its precision follows the usual arithmetic. Suppose the per-trade half-effective spread within a cell has a standard deviation of 0.5 pips, and the target is to know the cell mean to within $\pm 0.1$ pips with 95% confidence. Then

$$n = \left(\frac{z\,\sigma}{d}\right)^2 = \left(\frac{1.96 \times 0.5}{0.1}\right)^2 = 96.04,$$

so 97 fills per cell. With 21 instruments and 8 session buckets, that is 168 cells and $168 \times 97 = 16{,}296$ fills. At the M30 book's 750 round trips a year — 1,500 fills — that is 10.86 years of live trading to populate the surface once, before any regime changes, before any conditioning on volatility or news, and before the broker revises its pricing.

The conclusion is not that the surface is impossible. It is that it cannot be built from fills alone, and must be layered by how expensive each layer's data is:

- **Layer 1 — quoted spread, from the feed.** Costs nothing but storage, and can be sampled at whatever rate the connection allows. Polling every five seconds across a 118-hour FX week yields 84,960 observations per instrument per week, which populates every cell of the surface densely within days, including cells the strategy has never traded. This layer gives the *shape* of the surface: how much wider Asia is than the overlap, how far a release moves the quote, how the rollover instant behaves.
- **Layer 2 — effective spread, from your own fills.** Sparse and slow to accumulate, but it measures the thing the account is actually charged. Its role is to calibrate the *level*: estimate the ratio $s^{\text{eff}}/s^{\text{q}}$, which absorbs price improvement and slippage and which is far more stable across cells than either quantity separately, and then reconstruct $\widehat{s}^{\text{eff}}(\text{cell}) = \widehat{\rho} \cdot \widehat{s}^{\text{q}}(\text{cell})$ using the dense Layer-1 estimate for the shape. In Worked Example 3.1 that ratio was 0.8571; the point is not the value but that a single pooled ratio needs far fewer fills to pin down than 168 separate means.
- **Layer 3 — realised spread and adverse selection.** The sparsest, requiring fills *and* post-trade mids. This layer does not belong in the cost model at all: the account pays the effective spread, not the realised one. It belongs in the diagnostics, where it answers a different question entirely.

That third layer's question is the one worth ending on. Section 3.2 observed that the adverse-selection component measures how much the market revised its view in the direction of a trade, shortly after the trade. Computed across your own fills, it is a direct, model-free test of whether your orders carry information: a book whose fills are consistently followed by drift in their own direction is being adversely selected *against the dealer*, which is what a genuine edge looks like from the counterparty's side. A book whose fills show adverse selection indistinguishable from zero is paying the full effective spread for immediacy and receiving nothing in return — and no amount of cost engineering fixes that, because the problem is not the cost.

**Trap — Calibrating costs on the trades you would like to have made**

Two failure modes are common enough to name. The first is calibrating the cost model on a *backtest's* fills — the timestamps and prices the simulator produced — rather than on live fills; this measures the simulator's assumptions, in a closed loop, and will always agree with itself. The second is estimating a cell's cost from all quotes in that cell while trading only its worst moments: if the strategy fires into releases, its costs come from the right tail of the cell's spread distribution and the cell mean understates them, sometimes by a multiple. Both errors are invisible in-sample and both produce a cost model that is optimistic exactly where the strategy is most active [Course Book §18.3].

## 3.7 Exercises

**Exercise 3.1 — The three spreads on a GBPUSD tape**

Five hypothetical GBPUSD trades are recorded below, with the prevailing quote, the trade price $p$, the direction $q$ ($+1$ buyer-initiated), and the mid five minutes later. Work in pips (1 pip $= 0.0001$).

| # | bid | ask | $p$ | $q$ | $m_{t+5}$ |
|---|---|---|---|---|---|
| 1 | 1.26840 | 1.26858 | 1.26858 | $+1$ | 1.26855 |
| 2 | 1.26852 | 1.26866 | 1.26852 | $-1$ | 1.26861 |
| 3 | 1.26860 | 1.26876 | 1.26870 | $+1$ | 1.26872 |
| 4 | 1.26870 | 1.26890 | 1.26870 | $-1$ | 1.26866 |
| 5 | 1.26862 | 1.26878 | 1.26878 | $+1$ | 1.26868 |

(a) Compute $m_t$, the quoted, effective and realised spreads, and the adverse-selection component, for each trade and on average.

(b) Which trade received price improvement, and how much?

(c) Trade 4 has both the widest quoted spread and the largest adverse selection. Explain what that combination says about why the quote was wide, and whether widening it worked.

(d) The account trades 50,000 units and is denominated in USD. What did the fill on trade 4 cost, and what is the mean cost of a single fill across the tape?

▷ working space

**Solution — Exercise 3.1**

**(a)** The mid is $(b+a)/2$; the quoted spread is $a-b$; the effective spread is $2q(p-m_t)$; the realised spread is $2q(p-m_{t+5})$; adverse selection is their difference, which must equal $2q(m_{t+5}-m_t)$.

| # | $m_t$ | $s^{\text{q}}$ | $s^{\text{eff}}$ | $s^{\text{real}}$ | adv. sel. |
|---|---|---|---|---|---|
| 1 | 1.26849 | 1.8 | 1.8 | 0.6 | 1.2 |
| 2 | 1.26859 | 1.4 | 1.4 | 1.8 | −0.4 |
| 3 | 1.26868 | 1.6 | 0.4 | −0.4 | 0.8 |
| 4 | 1.26880 | 2.0 | 2.0 | −0.8 | 2.8 |
| 5 | 1.26870 | 1.6 | 1.6 | 2.0 | −0.4 |
| mean | | 1.68 | 1.44 | 0.64 | 0.80 |

Worked in full for trade 4: $m_t = (1.26870 + 1.26890)/2 = 1.26880$, $s^{\text{q}} = 0.00020 = 2.0$ pips, and the sale filled at the bid so $s^{\text{eff}} = 2(-1)(1.26870 - 1.26880) = 0.00020 = 2.0$ pips. Five minutes later the mid is 1.26866, fourteen tenths of a pip *below* the trade price, so $s^{\text{real}} = 2(-1)(1.26870 - 1.26866) = -0.00008 = -0.8$ pips, and adverse selection is $2.0 - (-0.8) = 2.8$ pips, confirmed by $2(-1)(1.26866 - 1.26880) = 2.8$ pips. The identity $1.44 = 0.64 + 0.80$ holds in the means as it does row by row.

**(b)** Only trade 3: the buy filled at 1.26870 against an ask of 1.26876, an improvement of 0.6 pips, which is why its effective spread (0.4 pips) is a quarter of its quoted spread (1.6). Across the tape, takers paid $1.44/1.68 = 85.71\%$ of the advertised spread.

**(c)** A wide quote is a liquidity supplier's defence against adverse selection: he widens when he believes the probability of trading against someone better informed is elevated — Chapter 2's third spread component, and typically a release, a thin session, or a sharp move already underway. Trade 4 shows the defence being mounted and *failing*: the supplier charged 2.0 pips, twice the tape's cheapest quote, and still lost, because the mid then ran 1.4 pips against him, giving a realised spread of −0.8 pips. The lesson runs both ways. For the dealer, widening reduces but does not eliminate the loss. For the trader, the widest quotes are precisely where the information is, so a strategy that refuses to trade wide quotes is refusing to trade at the moments that carry the most information — a real trade-off, not an obvious win. Note also that adverse selection is 55.56% of the mean effective spread here, against 75% in Worked Example 3.1: the split is not a constant of nature and must be measured per book.

**(d)** The pip value is $0.0001 \times 50{,}000 = \$5.00$. The cost of a single fill is the *half*-effective spread, so trade 4 cost $2.0/2 = 1.0$ pip $= \$5.00$, and the mean fill across the tape cost $1.44/2 = 0.72$ pips $= \$3.60$. A complete round trip at the tape's average would cost 1.44 pips $= \$7.20$. (All figures verified in `verify_mk03.py`, §8.)

**Exercise 3.2 — Implementation shortfall on a sell programme**

A manager decides at 09:45 to sell 60,000 shares, with the stock at $P_d = \$25.00$. The order reaches the market at 10:20 with the stock at $P_a = \$24.96$. It fills 25,000 at \$24.92, 20,000 at \$24.86 and 5,000 at \$24.80; the remaining 10,000 are cancelled at the close, with the stock at $P_T = \$24.70$. Commission is \$0.004 per share executed.

(a) Compute the paper-portfolio gain, the real-portfolio gain, and the implementation shortfall.

(b) Decompose the shortfall into delay, execution against arrival, opportunity and commission, in currency and in basis points of the decision value.

(c) Express each component as a share of the total. Which line would an execution algorithm improve, and which would it make worse?

(d) The stock's average daily volume is 3,000,000 shares and its daily volatility is 2.2%. Using (3.3) with $Y = 0.5$, is the execution cost against arrival explained by impact?

▷ working space

**Solution — Exercise 3.2**

**(a)** The paper portfolio shorts all 60,000 at \$25.00 and marks at \$24.70: $60{,}000 \times (25.00 - 24.70) = \$18{,}000.00$. The real portfolio realises proceeds of $25{,}000(24.92) + 20{,}000(24.86) + 5{,}000(24.80) = 623{,}000 + 497{,}200 + 124{,}000 = \$1{,}244{,}200.00$ (VWAP \$24.884000) for 50,000 shares, which mark at $50{,}000 \times 24.70 = \$1{,}235{,}000.00$; commission is $0.004 \times 50{,}000 = \$200.00$. So $\Pi_{\text{real}} = 1{,}244{,}200 - 1{,}235{,}000 - 200 = \$9{,}000.00$ and the shortfall is $18{,}000 - 9{,}000 = \$9{,}000.00$ — exactly half the paper profit.

**(b)** Every price difference reverses sign relative to the buy case: a sell is hurt when the price falls before it trades. The decision value is $60{,}000 \times 25.00 = \$1{,}500{,}000$.

| Component | Formula | Currency | Basis points |
|---|---|---|---|
| Delay | $50{,}000(25.00 - 24.96)$ | \$2,000.00 | 13.3333 |
| Execution vs arrival | $\sum n_i(24.96 - p_i)$ | \$3,800.00 | 25.3333 |
| Opportunity | $10{,}000(25.00 - 24.70)$ | \$3,000.00 | 20.0000 |
| Commission | $0.004 \times 50{,}000$ | \$200.00 | 1.3333 |
| Shortfall | | \$9,000.00 | 60.0000 |

The execution line comes from $25{,}000(0.04) + 20{,}000(0.10) + 5{,}000(0.16) = 1{,}000 + 2{,}000 + 800 = \$3{,}800.00$, the per-tranche concession rising from 4 to 16 cents as the order works down through the available bids.

**(c)** The shares are: delay 22.22%, execution against arrival 42.22%, opportunity 33.33%, commission 2.22%. An execution algorithm owns only the second line, and can reduce it by trading more passively — resting on the bid rather than crossing — but doing so on a falling stock raises the third line, because more of the order goes unfilled at exactly the moment the price is running away. The two move in opposition, which is why implementation shortfall is defined as their sum and why optimising either alone is a way of moving cost into a line nobody is measuring. Note also that commission, the only invoiced item, is 2.22% of the bill.

**(d)** Participation is $60{,}000/3{,}000{,}000 = 2.00\%$ of ADV, so $\sqrt{Q/V} = 0.14142136$ and

$$\frac{\Delta P}{P} \approx 0.5 \times 0.022 \times 0.14142136 = 0.00155564 = 15.5564 \text{ bp}.$$

The realised execution cost against arrival was 25.3333 bp, some 1.63 times the modelled impact. Impact therefore explains roughly three-fifths of it and no more; the balance is the market's independent drift during a slow execution — the stock fell 26 cents from arrival to close, far more than any 60,000-share order can account for — plus, possibly, an over-aggressive schedule. The honest reading is that (3.3) sets a floor for what the execution should have cost, not a benchmark it failed. (All figures verified in `verify_mk03.py`, §9.)

**Exercise 3.3 — Financing both ways, and the flip point**

A EURUSD CFD position of 50,000 units is opened at $S = 1.0850$. Take indicative rates $r_{\text{EUR}} = 3.25\%$, $r_{\text{USD}} = 5.35\%$, a broker markup $\alpha = 1.00\%$, and $D = 365$. The account is in USD.

(a) Compute the daily financing for a long and for a short position, in currency and in pips.

(b) Compute the accrued financing over a 45-day hold in each direction.

(c) The strategy's thesis is a 35-pip move, and the round-trip effective spread is 1.3 pips. On the long side, at what holding period does the trade break even? What is its net result if held the full 45 days and the target is hit exactly?

(d) Answer (c) for the short side, and state in one sentence what the asymmetry means for position management.

(e) How much of the long's daily charge is the broker's markup rather than the interest differential, and what would the flip point be without it?

▷ working space

**Solution — Exercise 3.3**

**(a)** The notional is $50{,}000 \times 1.0850 = \$54{,}250$ and the pip value is $0.0001 \times 50{,}000 = \$5.00$. The long's annualised rate is $(0.0325 - 0.0535) - 0.0100 = -0.0310$ and the short's is $(0.0535 - 0.0325) - 0.0100 = +0.0110$, so

$$f_{\text{long}} = \frac{54{,}250(-0.0310)}{365} = -\$4.607534 \text{ per day} = -0.9215068 \text{ pips},$$

$$f_{\text{short}} = \frac{54{,}250(+0.0110)}{365} = +\$1.634932 \text{ per day} = +0.3269863 \text{ pips}.$$

**(b)** Over 45 days the long pays $45 \times 4.607534 = \$207.34$, or 41.4678 pips; the short receives $45 \times 1.634932 = \$73.57$, or 14.7144 pips.

**(c)** Net profit in pips as a function of holding period is $\Pi(H) = 35.0 - 1.3 - 0.9215068H$, so

$$H^* = \frac{35.0 - 1.3}{0.9215068} = 36.57 \text{ days}.$$

Held the full 45 days, even with the target hit exactly, the trade nets $33.7 - 0.9215068(45) = -7.7678$ pips $= -\$38.84$. A completely correct 35-pip forecast loses money because it was right too slowly.

**(d)** On the short side financing is a credit, so $\Pi(H) = 33.7 + 0.3269863H$, which is increasing without bound: at 45 days the short nets $33.7 + 14.7144 = 48.4144$ pips $= \$242.07$, some 44% more than the price move alone delivered. The asymmetry means holding-period discipline is direction-dependent: on the financed side a position has a computable expiry beyond which its own thesis cannot pay for it, while on the credited side time is an asset and premature exit forfeits a known, positive accrual.

**(e)** The markup component is $54{,}250 \times 0.0100/365 = \$1.486301$ per day, which is 32.26% of the long's \$4.607534 daily charge. Stripping it out leaves the pure differential, $54{,}250 \times 0.0210/365 = \$3.121233$ per day $= 0.6242466$ pips, and the flip point becomes $33.7/0.6242466 = 53.99$ days — the markup alone shortens the position's viable life from about fifty-four days to about thirty-seven, a reduction of a third. (All figures verified in `verify_mk03.py`, §10.)

**Exercise 3.4 — The annual cost bill under two session mixes**

A GBPUSD M15 strategy takes 1,200 round trips a year in 30,000-unit positions (pip value \$3.00, USD account). Its cost surface reports the following mean round-trip effective spreads (illustrative cell values, not market data): 1.4 pips in the London–New York overlap, 1.9 pips in single-session London, 3.2 pips in Asia, and 5.5 pips inside scheduled-release windows. Its trades are distributed 25%/35%/30%/10% across those four buckets (mix A). The gross edge is 6.5 pips on the release-window trades and 2.5 pips everywhere else.

(a) Compute the blended gross edge per trade and the expected round-trip spread under mix A, and hence the annual gross P&L, spread bill and net.

(b) Entries are re-gated so that half the Asia allocation moves to the overlap, giving mix B: 40%/35%/15%/10%. Recompute, and express the improvement as a percentage of net P&L.

(c) Should the release-window bucket be dropped altogether? Support the answer with its stand-alone arithmetic, and state what would change your mind.

(d) Compare the margin of safety — gross edge less expected spread — under the two mixes, as a share of gross. Comment on what that ratio implies about how much

cost-model error each configuration can survive.

▷ working space

**Solution — Exercise 3.4**

**(a)** The blended gross edge is $0.90(2.5) + 0.10(6.5) = 2.900$ pips. The expected round-trip spread under mix A is

$$0.25(1.4) + 0.35(1.9) + 0.30(3.2) + 0.10(5.5) = 0.350 + 0.665 + 0.960 + 0.550 = 2.525 \text{ pips.}$$

Annual gross P&L is $1{,}200 \times 2.900 \times \$3.00 = \$10{,}440.00$; the spread bill is $1{,}200 \times 2.525 \times \$3.00 = \$9{,}090.00$, which is 87.07% of gross; the net is \$1,350.00.
**(b)** Under mix B the expected spread is $0.40(1.4) + 0.35(1.9) + 0.15(3.2) + 0.10(5.5) = 0.560 + 0.665 + 0.480 + 0.550 = 2.255$ pips, a bill of $1{,}200 \times 2.255 \times \$3.00 = \$8{,}118.00$ (77.76% of gross) and a net of \$2,322.00. The saving is \$972.00, which is an increase in *net* P&L of $972/1{,}350 = 72.00\%$. The model, the features and the signal threshold are unchanged: the entire gain comes from choosing when to pay.
**(c)** No — not on these numbers. The release bucket is $1{,}200 \times 0.10 = 120$ trades carrying a gross edge of 6.5 pips against a mean cost of 5.5, so it nets 1.0 pip per trade, 120 pips, or \$360.00 a year. Dropping it forgoes \$2,340.00 of gross to save \$1,980.00 of cost, a net loss of \$360.00. But the margin is one pip against a 5.5-pip cost, so the decision turns entirely on the cost estimate: an 18.18% error in the release bucket's mean spread wipes the bucket out. That bucket is also the one whose spread distribution is most right-skewed, so its *mean* is the least representative summary of what any individual trade pays. What would change the answer: measuring the bucket's cost distribution rather than its mean and finding that a material share of its trades pay well above 6.5 pips; or finding that the 6.5-pip gross edge is itself estimated from mid-price candles that never had to be traded.
**(d)** Mix A leaves $2.900 - 2.525 = 0.375$ pips of margin, 12.93% of gross; mix B leaves 0.645 pips, 22.24%. The margin of safety is not just profit, it is tolerance: mix A dies if the cost surface understates the true spread by about an eighth, while mix B survives an error of more than a fifth. Since the cost surface is an estimate with real sampling error (Section 3.6), the configuration with the larger margin is safer by substantially more than the ratio of its expected profits suggests. (All figures verified in `verify_mk03.py`, §11.)

**Exercise 3.5 — Essay: why the high-hit-rate book died**

The platform's earlier discretionary and rule-based history contained a book that won a clear majority of its trades and lost money over the year. Using only this chapter's vocabulary — effective spread, adverse selection, breakeven gross edge, financing, the cost surface — explain in three or four paragraphs what happened, and state which single measurement, had it been taken at the time, would have identified the problem before the money was gone. Support the explanation with the arithmetic of a $+5/-7$ pip book at a 62% hit rate, 750 round trips a year, 20,000 units per trade, against Section 3.4's all-in cost of 1.9130 pips per round trip.

▷ working space

**Solution — Exercise 3.5**

The arithmetic first, because it settles the question. A book that wins 5 pips 62% of the time and loses 7 pips the rest has a gross expectancy of $0.62(5) - 0.38(7) = 3.10 - 2.66 = 0.44$ pips per round trip: at 750 trades and $2.00 per pip, $660.00 a year gross. The all-in cost is 1.9130 pips per round trip — 1.655 pips of expected effective spread under the session mix it actually traded, plus 0.2580 pips of financing — or $2,869.53 a year. Costs are therefore 4.35 times the gross edge, and the book nets $-1.4730$ pips a trade, $-$2,209.53 a year, or $-8.84\%$ of a $25,000 account. Nothing went wrong in the sense of a bug or a regime change. The book did exactly what it was designed to do and the design could not pay for itself.

The hit rate was never the relevant statistic, and this is where the chapter's vocabulary earns its place. With a $+5/-7$ payoff, the hit rate required to break even at zero cost is $7/(5+7) = 58.33\%$; at the measured all-in cost it is $(1.9130+7)/12 = 74.28\%$. Costs move the threshold by 15.94 percentage points, and the book's 62% — comfortably above the frictionless bar, comfortably below the real one — sat precisely in the gap where a system feels like it is working while the account drains. The Course Book calls this the hit-rate illusion [Course Book §14.1]; the microstructural statement of it is that the gross edge per trade, 0.44 pips, was smaller than the half-effective spread paid on a *single* fill, roughly 0.83 pips under that session mix. The book was never trading for a profit; it was trading for a fraction of the cost of trading.

Where did the money physically go? Worked Example 3.1's decomposition answers it. Of the effective spread paid, only the realised-spread portion is a fee for immediacy; the rest — 75% on that tape — is adverse selection, which is to say the market moving against the fill immediately after it, because the fill happened at a moment when the counterparty had good reason to widen. Applying that share to the 1.655-pip spread bill puts 1.2412 pips per trade, $1,861.87 a year, into the adverse-selection column: the book was systematically buying immediacy at exactly the moments when immediacy was most expensive — thin sessions, releases, the rollover — which is what a signal built to react to volatility will do unless it is explicitly gated. Financing did the rest, quietly, on the fifth of positions carried across the New York rollover.

The single measurement that would have caught it is the per-trade effective spread, logged live from the account's own fills against the broker's own quote at the fill instant, and aggregated by session bucket. Not the quoted spread, which the platform could see all along and which understates what was paid whenever the stream requotes; not the backtest's assumed slippage, which was a parameter chosen to make the backtest work; and not the annual P&L, which reports the damage after it is done. One column — $2q(p - m_t)$, computed at the moment of every fill and bucketed by hour — would have shown a mean cost of about 1.7 pips against a gross edge of 0.44 before the second month was out, and the comparison of those two numbers is the whole of the diagnosis. That column is what the cost surface exists to produce, and it is the reason mid-only candles were a structural defect and not a storage optimisation: they make $m_t$ recoverable and $p - m_t$ permanently unknowable. (All figures verified in `verify_mk03.py`, §12.)

**Key idea — What to remember**

Cost is not a number but a taxonomy, and the two lines a broker invoices — commission and financing — are almost never the two that decide whether a systematic book survives. The spread has three definitions that must not be confused: the quoted spread is an advertisement, free to sample and silent about what was paid; the effective spread, $2q(p - m_t)$, is what the trade actually cost and is the only one belonging in a cost model; and the realised spread, $2q(p - m_{t+\Delta})$, is what the liquidity supplier kept, so that their difference measures adverse

selection directly from a tape rather than inferring it from a model — on this chapter's six-trade example, three-quarters of everything the takers paid. Perold's implementation shortfall completes the accounting by pricing the trades that did *not* happen, and its four-way split — delay, execution against arrival, opportunity, commission — assigns each line to the decision that caused it; in the worked programme, the only invoiced item was 3.3% of a 34.4 basis-point bill. The square-root impact law that governs that world has a calibration domain, and must be held to it: at one standard lot of EURUSD it predicts about a hundredth of a pip, roughly 134 times smaller than the spread actually paid, so a retail cost model built on impact theory is precisely calibrated to the one mechanism that does not apply — what applies instead is a measured, session-conditioned cost surface and the financing arithmetic that gave a correct 15-pip forecast a thirty-four-day expiry, and that turned an M30 book from $-0.08\%$ to $+2.43\%$ of equity a year with the forecasting model untouched at every rung. Which is the chapter's real claim: every technique here — signing trades, differencing against a later mid, decomposing a shortfall, bucketing by session — is market microstructure, the machinery Chapter 2 aimed at dealers and order books, turned around and aimed at your own account, where you enjoy an advantage no researcher has, because you know your own $q$, your own $p$ and your own decision times exactly and need estimate none of them. Chapter 4 goes to the market where those costs are set: the foreign exchange market's tiered dealer structure, its quoting conventions and its liquidity clock, which is the same session grid this chapter's cost surface is indexed on, now explained rather than merely measured.

PART II

# The Foreign Exchange Market



LUNA LABS

# 4

# The Foreign Exchange Market

*This chapter is the platform's home ground: every OANDA quote the reader's system consumes, every pip of profit and loss its backtests compute, and every night's financing charge on an open position originate in the market described here. It maps the foreign exchange market's decentralised, tiered structure and its dominant instrument, the FX swap; derives pip-value arithmetic precisely enough to reconcile a broker statement under any of three account-currency conventions; states and uses the triangulation identity that pins cross rates to within a spread; and turns the folklore of "London is deep, Asia is thin" into the exact session clock — rollover, value dates, and daylight-saving wrinkles included — that the reader's own cost surface already treats as a governing variable. It closes by placing the reader's own broker precisely within the dealer chain his orders actually traverse.*

## 4.1 The shape of the market

Foreign exchange has no building. There is no single order book in which every EURUSD order in the world queues for execution, no consolidated tape recording "the" last trade, and no regulator-mandated best-bid-and-offer that every venue must match or beat. What exists instead is a tiered network of bilateral dealer relationships, electronic platforms, and give-up arrangements, all quoting the same handful of currency pairs against each other continuously, never quite agreeing on the price to the pip. This is the purest OTC case among the four pillars: an exchange enforces a single matching venue by construction, but nothing in FX plays that role, and every quote the reader's platform receives is, structurally, one dealer's opinion — a point this chapter returns to in §4.6.

Four tiers matter, coarsely ordered by proximity to the ultimate liquidity:

(i) **The interbank/interdealer core.** The handful of the largest global banks quote each other continuously, both bilaterally and through anonymous, order-book-like interdealer platforms that historically consolidated the spot majors' deepest liquidity. This tier sets the reference prices everyone downstream ultimately marks against.

(ii) **Prime brokerage.** A prime broker — typically one of the tier-one banks — extends its own credit and settlement relationship to a client (a hedge fund, a smaller bank, sometimes a large retail aggregator) so that client can trade with many dealers while every resulting trade settles, on the prime broker's books, as if it had traded with the prime broker alone (a "give-up"). This is what lets a client without its own network of bilateral credit lines still access tier-one liquidity.

(iii) **ECNs and multi-dealer platforms.** Electronic venues aggregate streaming quotes from several liquidity providers into a single composite book a client can trade against; they do not themselves take principal risk, but they compress the search cost of finding the best price across dealers that a single bilateral relationship cannot.

(iv) **Retail brokers.** The client-facing layer — OANDA among them — quotes individual retail traders a bid and an ask, sources its own hedging liquidity from some combination of the

tiers above, and either passes client flow through or warehouses it. §4.6 works through this tier in full; the reader's own orders never leave it.

This is a tiered, relationship-based, search market rather than a single anonymous auction — Duffie's (2012) general account of over-the-counter markets as venues where participants must find a counterparty before they can trade at all, rather than simply post an order and wait, describes FX's structure as well as it describes any other OTC asset class, notwithstanding FX's unusually deep and continuous liquidity at the top of the tier structure. Rime and Schrimpf (2013), writing from inside a BIS triennial survey, give the same structure its FX-specific anatomy: a small core of dealing banks internalising an increasing share of client flow electronically, prime brokerage widening who can reach that core, and non-bank market makers increasingly present in the interdealer tier once reserved for banks alone. The reader's own 21-instrument book inherits exactly this structure one layer further down, at whatever liquidity the retail broker of §4.6 itself sources.

**Definition — The BIS Triennial Survey**

Every three years the Bank for International Settlements coordinates central banks worldwide to survey FX and OTC derivatives turnover for one representative month, producing the only comprehensive, methodologically consistent measure of global FX activity. Because it is a snapshot taken once every three years, every scale figure it produces is a vintage, not a constant: state it with the survey's reference month attached, and expect the next survey to have moved it.

The most recent published survey (BIS, 2022) put average daily global FX turnover at roughly $7.5 trillion in April 2022 — up from about $6.6 trillion in the 2019 survey — an order of magnitude worth internalising even though the digits will already be stale by the time the next triennial round reports (indicative, April 2022 vintage; verify current specifications before citing this number as current). By instrument, that turnover splits roughly: around half in FX swaps, a little over a quarter in spot, roughly a sixth in outright forwards, and the remainder in currency swaps and options combined (BIS, 2022, indicative) — a split §4.2 explains rather than merely reports. By location, trading concentrates geographically far more than the "24-hour market" framing suggests: London alone accounts for a little under two-fifths of global turnover, more than the next several trading centres combined, with the United States a clear second and Singapore, Hong Kong and Japan forming the next tier (BIS, 2022, indicative). This is why §4.4's liquidity clock treats the London session, and above all its overlap with New York, as the day's structural peak rather than one region among several equally weighted.

**Trap — A survey figure is not a live feed**

Every number in this section is a snapshot of one April, three years stale at best by the time you read the next survey. Treat "$7.5 trillion a day" as an order of magnitude that tells you FX dwarfs every other asset class's turnover, not as a figure to cite with any more precision than the survey itself claims, and never as a substitute for checking the current BIS release when the number itself is what matters to an argument.

## 4.2 Instruments

Four instrument types cover essentially all FX trading; retail CFDs, including every position the platform holds, are a fifth, synthetic construction layered on top of the first.

**Spot.** A spot trade exchanges two currencies for value two business days after the trade date (T+2), the market's default settlement convention; USD/CAD is the standard exception, settling T+1, because the two currencies share a North American business day and the extra day buys

nothing. Nothing about "spot" means instantaneous delivery — it means the shortest standard value date the market quotes, and §4.4 returns to exactly what that value date implies for a position held overnight.

**Outright forwards.** A forward fixes today an exchange rate for a single future value date, with no intervening cash flow and no obligation until that date. The forward rate is not a free market forecast of the future spot rate; it is pinned by no-arbitrage to the spot rate and the two currencies' interest rates.

**Result — The no-arbitrage forward (covered interest parity, sketch)**

Let $S$ be the spot rate (domestic currency per unit of foreign currency), $F$ the forward rate for horizon $T$ (in years), and $r_d, r_f$ the simple annualised domestic and foreign risk-free rates over $T$. Compare two riskless routes from one unit of domestic currency today to domestic currency at $T$: invest it domestically, yielding $1 + r_d T$; or convert it to $1/S$ units of foreign currency, invest that abroad at $r_f$, and sell the resulting $(1/S)(1 + r_f T)$ units of foreign currency forward today at rate $F$, locking in $(F/S)(1 + r_f T)$ domestic currency with no FX risk. Both routes are riskless and start from the same capital, so no-arbitrage forces their payoffs to match:

$$1 + r_d T = \frac{F}{S}(1 + r_f T) \quad \Longrightarrow \quad F = S \cdot \frac{1 + r_d T}{1 + r_f T}.$$

Forward points are $F - S$; for small rates this is approximately $S(r_d - r_f)T$ (expand the fraction and drop the second-order $r_f T$ term). Chapter 5 derives this in full generality, states its continuous-compounding form, and builds the carry trade on the systematic failure of its *uncovered* cousin; here the point is narrower: once $S$, $r_d$ and $r_f$ are set, the forward rate is not a fresh negotiation, it is arithmetic.

**Definition — FX swap**

An FX swap is a single, simultaneously agreed transaction with two legs: exchange currency A for currency B at a near date (often spot), and exchange it back at a far date at a forward rate consistent with the no-arbitrage result above. The two legs net to zero currency risk — whoever delivers A and receives B at the near leg has already locked in reversing that exchange at the far leg — so economically an FX swap is a collateralised loan of one currency against the other for the tenor between the two dates, priced at an implicit interest rate equal to the forward points.

This is why FX swaps are the single largest instrument category in the BIS figures above: they are not a directional bet on a future exchange rate, they are the money market's cross-currency plumbing, and most of that activity is very short-dated — overnight and tomorrow-next tenors bridging a one- or two-day currency funding mismatch, renewed daily, rather than a medium-term hedge negotiated once and held. A bank funding a foreign-currency asset with domestic-currency liabilities, or a broker managing the currency composition of client margin balances overnight, reaches for a swap far more often than for an outright forward, because the swap is exactly sized to the funding gap and creates no residual FX exposure. §4.4 draws the line this chapter needs directly: the nightly "rollover" or "swap" charge on every CFD position the platform holds is the retail-sized, daily-renewed instance of precisely this mechanism, not a separate fee bolted onto the trade.

**Non-deliverable forwards (NDFs).** For currencies subject to capital controls or limited offshore convertibility, a forward that would require actual delivery of the restricted currency is replaced by an NDF: both parties agree a notional and a forward rate as usual, but at maturity they settle

only the cash difference between that rate and an official fixing, in a freely convertible currency (typically USD). Structurally an NDF is a forward with delivery replaced by a cash settlement formula; which currencies trade this way, and against what fixing, shifts with capital-control regimes and should be verified against current market practice rather than assumed fixed.

**Futures.** Exchange-traded FX futures (the CME's currency contracts are the deepest example) standardise contract size, value date and margining, and converge to spot at expiry; they carry a modest share of total FX turnover relative to the OTC instruments above, existing mainly because a subset of participants — funds and corporates that need exchange-cleared, margined, centrally-guaranteed exposure — value that structure over the OTC alternative's flexibility.

**Where retail CFDs sit.** A retail FX CFD — every position the SST platform holds — references none of the above directly. It tracks the prevailing OTC spot rate continuously, with the broker itself as the direct counterparty (§4.6), and never settles to an actual value date: instead of delivering currency, an open position is rolled forward nightly at a financing charge that replicates, in miniature and daily, the forward-points cost a real forward or swap would have embedded over the same horizon. A CFD is, in this precise sense, synthetic exposure to spot plus a rolling series of overnight FX swaps you never explicitly request.

## 4.3 Quoting conventions

**Definition — Base and quote currency**

A currency pair BASE/QUOTE is quoted as $S$ units of the quote currency per one unit of the base currency: EURUSD $= 1.0850$ means one euro (base) is worth \$1.0850 (quote). Buying the pair means buying the base currency and simultaneously selling the quote currency; selling it means the reverse.

Market convention further labels pairs by which currency is USD. **American terms** quote the foreign currency's dollar price, USD as the quote currency — EURUSD, GBPUSD, AUDUSD and NZDUSD are all American-terms pairs. **European terms** quote the dollar's price in the foreign currency, USD as the base — USDJPY, USDCHF and USDCAD. From a US-domestic vantage point the first group are *direct* quotes (domestic currency per unit of foreign) and the second *indirect* (foreign currency per unit of domestic); the labels are conventionally applied from that USD-centric reference frame throughout FX discourse, reflecting the dollar's role as the market's vehicle currency, not because every trader is American.

**Definition — Pip and pipette**

A *pip* is the standard quoting increment: 0.0001 of the quote currency for essentially every pair, except pairs quoted with the Japanese yen as the quote currency, where it is 0.01. A *pipette* (or fractional pip) is one tenth of a pip — the fifth decimal place for non-JPY pairs, the third for JPY pairs — and is how brokers including OANDA stream prices tighter than a whole pip, e.g. EURUSD 1.08503 quotes three pipettes (0.3 pip) past 1.0850.

The two-decimal-place exception for JPY pairs is not an arbitrary quirk; it keeps a pip an economically comparable increment across pairs quoted at very different absolute magnitudes. For EURUSD at $S = 1.0850$, a 0.0001 pip is $0.0001/1.0850 = 0.9217$ basis points of the rate; for USDJPY at $S = 149.50$, a 0.01 pip is $0.01/149.50 = 0.6689$ basis points — both of order one basis point, whereas a 0.0001 pip applied literally to USDJPY would be $0.0001/149.50 = 0.0669$ basis points, an order of magnitude too fine to matter. Placing the yen pip two decimals to the left of the non-JPY convention is what keeps "one pip" meaning roughly the same thing, economically, regardless of which currency happens to be quoted in the hundreds rather than close to one (both figures verified in `verify_mk04.py`, §1).

Every quote is two-sided: a *bid* (the price a dealer will buy the base currency at) and an *ask* or *offer* (the price it will sell at), with ask $>$ bid and the *spread* their difference. A retail platform's quote stream is one dealer's bid and ask, not "the" market's — §4.6 returns to exactly what that means for a backtest. Position size is conventionally expressed in *lots*: a standard lot is 100,000 units of the base currency, a mini lot 10,000, a micro lot 1,000 — structural conventions stable across the retail industry, unlike the spreads and margin rates built on top of them.

### Pip value

The value of a one-pip move depends on the pair, the position size, and — unless the account happens to be denominated in the quote currency — the prevailing exchange rate used to convert that value into the account's own currency.

**Result — Pip value, three conventions**

For a position of $N$ units of the base currency in a pair with pip size $\pi$ (0.0001, or 0.01 if the quote currency is JPY) and spot rate $S$ (quote per base), the value of one pip is $\pi N$ units of the *quote* currency. Converting that into the account's currency depends on which currency the account is denominated in:

- **Account currency = quote currency:** no conversion; pip value $= \pi N$, directly in account-currency terms.
- **Account currency = base currency:** divide by the spot rate; pip value $= \pi N / S$, since one unit of quote currency is worth $1/S$ units of base currency.
- **Account currency is neither:** multiply by the prevailing exchange rate $X$ between the quote currency and the account currency (units of account currency per unit of quote currency); pip value $= \pi N \cdot X$. When the broker does not stream that cross directly, $X$ is exactly the triangulated rate derived below.

**Worked Example 4.1 — Pip value under all three conventions**

A trader with a USD-denominated account holds a one-standard-lot ($N = 100{,}000$) position in each of three pairs. Find the value of one pip in each case.

(a) EURUSD, $S = 1.0850$ (quote currency USD = account currency).
(b) USDJPY, $S = 149.50$ (base currency USD = account currency).
(c) AUDNZD, $S = 1.0790$, with NZDUSD $= 0.5920$ (neither leg is USD).

**(a)** Quote currency is USD, the account currency, so no conversion is needed:

$$\text{pip value} = 0.0001 \times 100{,}000 = \$10.00 \text{ per pip.}$$

This scales linearly with size: \$1.00 per pip on a mini lot ($N = 10{,}000$), \$0.10 on a micro lot ($N = 1{,}000$).
**(b)** The pip is 0.01 (JPY quote currency); its value in JPY is $0.01 \times 100{,}000 = 1{,}000$ JPY. Converting to USD (the base currency here, so divide by $S$):

$$\text{pip value} = \frac{1{,}000}{149.50} = \$6.6890 \text{ per pip.}$$

**(c)** The pip is 0.0001; its value in NZD (the quote currency) is $0.0001 \times 100{,}000 = 10.00$ NZD. USD is neither leg of AUDNZD, so convert via the given NZDUSD rate:

$$\text{pip value} = 10.00 \times 0.5920 = \$5.9200 \text{ per pip.}$$

(All three figures verified in `verify_mk04.py`, §1.)

**Your project — Classifying the SST book by pip convention**

The platform's fourteen FX majors and crosses each fall into exactly one of the three pip-value conventions above, and which one depends on the account currency chosen when the OANDA account was opened, not on the pair alone. Every pair with JPY as the quote leg (USDJPY, EURJPY, GBPJPY and any other JPY cross the book carries) additionally takes the 0.01 pip regardless of account currency. Work through the full instrument list once, for the account currency actually in use, and the classification never needs re-deriving — it only needs re-checking if the account currency itself changes. This is exactly Exercise 4.1, run for real against your own account rather than a textbook example; the non-FX instrument classes (metals, crypto, indices, energy) carry their own contract-specific tick conventions, out of scope here and addressed where those instruments are introduced.

### Cross rates and triangulation

A *cross rate* is a pair quoted between two non-USD currencies, e.g. EURGBP or AUDNZD. Even where a broker streams a cross directly, its price is not independently free to move: it is pinned, to within the combined cost of trading the alternative route, by the two USD pairs that share a leg with it.

**Result — Triangulation**

For currencies A, B and a shared currency C, chain the unit conversions through C to get the synthetic A/B rate implied by A/C and B/C. Three cases, all the same reasoning:

- C is the *quote* currency in both given pairs (A/C and B/C): $A/B = (A/C)/(B/C)$.
- C is the *base* currency in both (C/A and C/B): $A/B = (C/B)/(C/A)$.
- C is the quote currency in one and the base in the other (A/C and C/B): $A/B = (A/C) \times (C/B)$.

This is the identical algebraic identity used, for a different purpose, in [Course Book §16.5]: EURGBP = EURUSD/GBPUSD holds to within the bid-ask spread precisely because it is the first case above (C = USD, the quote currency in both EURUSD and GBPUSD), and the Course Book invokes it there to show why a same-timestamp split across correlated crosses leaks information — the crosses are not merely correlated, they are algebraically tied together by exactly this identity.

If a directly-quoted cross drifts far enough from its synthetic value, a riskless round trip opens: convert A to B directly, then rebuild A synthetically via C at a better combined rate than the direct route offered, pocketing the difference with no net currency exposure at the end. This is triangular arbitrage, and it is precisely what keeps crosses pinned to their synthetic value in the first place — any deviation beyond the combined transaction cost of the two-leg alternative is competed away by whoever notices it first, which in a market this liquid is measured in milliseconds, not the time it takes a retail trader to open a calculator.

**Worked Example 4.2 — Triangulating EURGBP, and checking for arbitrage**

A dealer quotes EURUSD 1.0848/1.0852 and GBPUSD 1.2678/1.2682. Find the no-arbitrage bid-ask band implied for EURGBP, and check whether a direct EURGBP quote of 0.8551/0.8555 admits a riskless round trip.

**Synthetic band.** To sell EUR for GBP synthetically, route EUR→USD→GBP: sell EUR at the EURUSD bid, buy GBP with the resulting USD at the GBPUSD ask (since buying GBP means

paying the ask). This gives the synthetic *bid*:

$$\text{synthetic bid} = \frac{\text{EURUSD bid}}{\text{GBPUSD ask}} = \frac{1.0848}{1.2682} = 0.8554.$$

To buy EUR with GBP synthetically, route GBP→USD→EUR: sell GBP at the GBPUSD bid, buy EUR with the resulting USD at the EURUSD ask. This gives the synthetic *ask*:

$$\text{synthetic ask} = \frac{\text{EURUSD ask}}{\text{GBPUSD bid}} = \frac{1.0852}{1.2678} = 0.8560.$$

The no-arbitrage band for a direct EURGBP quote is therefore $[0.8554, 0.8560]$, five to six pips wide — itself simply the sum of the two underlying pairs' own spreads, propagated through the ratio.

**Arbitrage check.** A round trip is profitable only if the direct bid exceeds the synthetic ask (sell direct, rebuild synthetically, pocket the gap) or the direct ask sits below the synthetic bid (buy direct, unwind synthetically). Here direct bid $= 0.8551$ is below the synthetic ask 0.8560, and direct ask $= 0.8555$ is above the synthetic bid 0.8554: the direct quote's interval overlaps the synthetic band on both sides, so neither condition holds and no riskless round trip is available. (All figures verified in `verify_mk04.py`, §3.)

**Trap — The mid-price gap is not a profit**

Comparing *mid* prices alone in Worked Example 4.2 — synthetic mid $1.0850/1.2680 = 0.8557$ against direct mid $(0.8551 + 0.8555)/2 = 0.8553$ — shows a gap of about 3.8 pips, tempting enough that a naive comparison looks like free money. It is not: executing either leg of the round trip means crossing a real bid-ask spread three times over (the two USD legs and the direct cross), and Worked Example 4.2 already showed the actual executable bands overlap with room to spare. Mid-to-mid triangulation gaps are a diagnostic for how wide the two markets' combined spread is, not a trading signal; at retail size, the cost of the three legs needed to realise a triangular arbitrage almost always exceeds the entire gap before you have finished entering the second leg.

## 4.4 The trading week and the liquidity clock

FX trades continuously from Sunday 22:00 UTC, when Wellington and Sydney open the week, to Friday 22:00 UTC, when New York closes it — five consecutive 24-hour rotations with no scheduled close in between, unlike an exchange that shuts every evening and reopens the next morning. Taken literally that is $5 \times 24 = 120$ hours a week. The convention this book's companion volumes use throughout — including [Course Book §6.8]'s data-wall arithmetic, built on exactly this figure — is 118 hours: the 120 nominal hours less roughly two hours, in aggregate across the week's four intraday rollovers and the weekend reopen, of thinned-to-absent liquidity that a serious backtest should not treat as equivalent to a normal trading hour. The two figures do not contradict each other; 118 is 120 discounted for exactly the phenomenon the rest of this section quantifies.

For a trader on Australian Eastern Standard Time — the reader's own clock, with no daylight saving applied — the same week states even more intuitively: it opens Monday at 8am and runs, continuously but for the nightly thinning §4.4 describes below, until Saturday at 8am, since Sunday 22:00 UTC and Friday 22:00 UTC each land ten hours later on the AEST clock (verified in `verify_mk04.py`, §8). A position can therefore be very much alive on a reader's own Saturday morning; "the week is over" is a UTC or New York statement, not something the reader's own calendar announces for free.

### Session characteristics

Within the continuous week, liquidity is nowhere near uniform. Four broad regimes recur every day:

- **Asia** is comparatively thin and rangey: Tokyo is the largest single participant, but the number of active desks is smaller than at any other point in the day, so prices tend to sit inside recent ranges rather than trend, punctuated by occasional yen-specific volatility around Tokyo fixings.
- **London** is the single deepest session on its own, consistent with its outsized share of the BIS turnover figures in §4.1: the bulk of European institutional and corporate flow executes here.
- **The London–New York overlap** is the day's liquidity peak, full stop: both dominant centres' dealing desks are live simultaneously, spreads are typically at their tightest, and the largest share of the day's volume prints in this window.
- **New York, late** — after London has closed — is thin again: the initial burst of NY-only liquidity fades, no European desk is left to add depth, and Asia has not yet opened.

### Rollover, value dates, and why 5pm New York

> **Definition — Value date**
>
> The value date of an FX transaction is the calendar date on which the two currencies actually change hands. Spot's default value date is trade date plus two business days (T+2, skipping weekends); a forward's value date is whatever future date the two parties agree.

A CFD position never reaches its value date — it is never actually delivered — so something has to stand in for the value-date roll a real spot position would undergo every night simply by continuing to exist. That something is the nightly *rollover*: at 5pm New York, every open position is closed and reopened (economically, via the FX-swap mechanism of §4.2) at a rate adjusted by the forward points for one day, crediting or debiting the interest-rate differential between the two currencies. 5pm New York is not an arbitrary choice of clock hour; it is the close of the New York trading day and, by long-standing market convention, the point at which the wholesale FX market itself rolls value dates forward, so retail brokers including OANDA anchor their own rollover to the same instant rather than inventing a second, incompatible convention.

Because spot's value date is T+2 business days, the value date attached to a continuously-held position must skip weekends exactly as a fresh spot trade would, and this creates a specific, predictable asymmetry across the week.

**Worked Example 4.3 — Value dates across a trading week, and the Wednesday roll**

Take the trading week of Monday 6 July 2026 to Friday 10 July 2026. For a trade dealt on each day, compute the T+2 (business-day) value date, and find which night's rollover carries the largest interest-rate adjustment.

Skipping weekends, the value dates are:

| Trade day | Date | T+2 value date |
|---|---|---|
| Monday | 6 Jul | Wednesday 8 Jul |
| Tuesday | 7 Jul | Thursday 9 Jul |
| Wednesday | 8 Jul | Friday 10 Jul |
| Thursday | 9 Jul | Monday 13 Jul |
| Friday | 10 Jul | Tuesday 14 Jul |

The rollover that fires at day $D$'s 5pm New York close advances a continuously-held position's reference value date from day $D$'s row to day $D+1$'s row above. Three of those four nightly advances move the value date forward by a single calendar day (Monday→Tuesday: Wed 8 Jul → Thu 9 Jul; Tuesday→Wednesday: Thu 9 Jul → Fri 10 Jul; Thursday→Friday: Mon 13 Jul → Tue 14 Jul), so each is charged one day's interest differential. The Wednesday-to-Thursday roll is different: the value date jumps from Friday 10 Jul to Monday 13 Jul, three calendar days, because a value date cannot fall on a weekend and must skip straight over it. The rollover firing at 5pm New York on Wednesday is therefore charged *triple* the normal daily rate — the standard CFD "triple swap Wednesday" convention, and nothing more mysterious than T+2 arithmetic meeting a five-day trading week (all dates and deltas verified in `verify_mk04.py`, §4).

**Result — The pattern is calendar-structural, not a property of one week**

The delta sequence in Worked Example 4.3 — one, one, three, one calendar days — depends only on the weekday of each trade, never on which specific week it is: T+2 business-day counting always skips the same weekend at the same point in a Monday-to-Friday week, holidays aside. Exercise 4.3 confirms the identical pattern recurs for a different week entirely.

A position that is only ever open *during* a trading day and closed before that day's rollover pays no financing at all, regardless of which weekday it traded on: rollover is charged to whatever is open at the 5pm New York instant, not to positions merely dealt that day. Symmetrically, a position that is open at that instant is charged the full rate for however many value-date days that specific rollover advances, even if it was opened minutes before and closed minutes after — duration within the day is irrelevant; only whether the position straddled 5pm New York, and which night's straddle that was, decides the charge.

### Daylight-saving wrinkles

The 5pm New York anchor is a *local* time, so the instant it falls at in UTC — and hence in every other reader's own local clock — shifts with United States daylight saving: 22:00 UTC when New York is on Eastern Standard Time (winter), 21:00 UTC on Eastern Daylight Time (summer). London's own session boundaries shift independently with UK daylight saving. Because United States daylight saving (second Sunday of March to first Sunday of November) and United Kingdom daylight saving (last Sunday of March to last Sunday of October) do not start and end on the same date, the New York–London offset is not constant at five hours all year: it is five hours whenever both are on standard time or both are on daylight time, but drops to four hours during the two short windows each year when only one of the pair has switched. In 2026, for example, US daylight saving begins 8 March and UK daylight saving begins 29 March — three weeks where New York has already sprung forward and London has not — and UK daylight saving ends 25 October against the US's 1 November, one further week of mismatch in the other direction (dates computed from each country's own rule in `verify_mk04.py`, §6). Tokyo observes no daylight saving at all, and the reader's own AEST/AEDT shifts only how the UTC rollover instant reads on his own clock, never the underlying UTC boundaries themselves §4.4 otherwise uses — consistent with how the Playbook's own session-bucketing step treats the two calendars as independent of each other and of Sydney's.

### Weekend gaps

Because nothing trades between Friday's close and Sunday's reopen, the first quote of the new week can differ from the last quote of the old one by a jump no intrabar spread model captures — the market's way of pricing in everything that happened while it was shut, all at once, at the open.

A stop order resting through the weekend is filled at the first available price after the gap, not at its stated level, if the market reopens through it; this is a risk-management fact about position sizing over a weekend, not a pricing anomaly, and [Course Book §18.3]'s insistence that weekend and rollover gaps be priced explicitly rather than smoothed over by whatever a backtest engine does by default with a multi-day candle gap is the same discipline applied to historical simulation.

**Your project — The cost surface is this clock, made quantitative**

Everything qualitative in this section — Asia thin, London deep, the overlap peak, New York late thin, Wednesday's triple roll — is exactly what the platform's own session-bucketed cost surface exists to measure rather than assert. Eight buckets (rollover_hour, weekend_open, Sydney, Tokyo, the Tokyo–London and London–New York overlaps, London, NY_late) partition every week, with a separate, narrower rollover_flag isolating the swap-application instant at finer resolution than the bucket grid. Building that partition on the wall clock, resolved against each of the two independent daylight-saving calendars this section describes, is precisely the difference between a cost model that is right about *when* spreads widen and one that merely knows, on average, that they do.

**Trap — Read the clock exactly, not approximately**

Two readings of this section's numbers are easy to get wrong. First, the "two-hour" gap between 120 nominal hours and the 118-hour convention is not one contiguous two-hour closure somewhere in the week — it is an aggregate modelling allowance spread across four daily rollovers and the weekend reopen, and no single clock hour is "the" missing two hours. Second, "Wednesday's triple swap" is Wednesday in New York: at 5pm EST that is 22:00 UTC, which on the reader's own AEST clock (ten hours ahead, no daylight saving) is already *Thursday* 8am. A trader who checks his own calendar for Wednesday evening and finds nothing unusual has not found an exception to the rule; he has looked on the wrong day, on the wrong side of the international date line, for an event that is timestamped in New York, not Sydney.

**Worked Example 4.4 — A session-cost scenario: the same strategy, two sessions**

A strategy trades one standard lot of EURUSD with a fixed 10-pip take-profit and no stop beyond the platform's normal risk controls, incurring one round-trip spread cost per trade and nothing else. Cost it at an illustrative London–New York overlap spread of 0.6 pip and an illustrative Asian-session spread of 1.4 pip (indicative scenario spreads, not live OANDA data — verify current spreads before using either figure operationally), and find the annual dollar difference over 250 trades a year, all else equal.

Net pips per trade are the gross target less the one round-trip spread:

$$\text{overlap: } 10 - 0.6 = 9.4 \text{ pips}, \qquad \text{Asia: } 10 - 1.4 = 8.6 \text{ pips}.$$

At \$10.00 per pip (Worked Example 4.1a), that is \$94.00 per trade at the overlap against \$86.00 in Asia — a difference of \$8.00 per trade, or 0.8 pip, purely from *when* the trade executes, holding the strategy and its gross edge fixed. Over 250 trades a year the two session choices diverge by

$$250 \times \$8.00 = \$2{,}000.00 \text{ per year},$$

and, expressed as a share of the gross target given up to spread alone, the overlap surrenders 6% of the 10-pip target and Asia 14% — more than double, on a strategy whose entire economic edge is the same ten pips either way (all figures verified in `verify_mk04.py`, §5).

The liquidity clock is not colour commentary on the trading day; it is a cost multiplier a strategy's own hit rate never sees.

## 4.5 Who moves the price

No single agent sets an exchange rate; it emerges from the aggregate of who is trading, in which direction, and why. Four groups matter most.

**Order flow.** Evans and Lyons (2002), working from confidential interdealer trade data, found that signed order flow — the running net of buyer- versus seller-initiated volume — explains a large share of daily exchange-rate variation, far more than contemporaneous macroeconomic news alone, reversing a long tradition of models that treated the exchange rate as driven only by public information. The mechanism is not that order flow is itself news; it is that flow aggregates dispersed private information (about hedging need, about portfolio rebalancing, about who currently wants to be long or short) faster and more completely than any public data release can, so price moves as dealers infer what the flow they are seeing implies and adjust their own quotes accordingly.

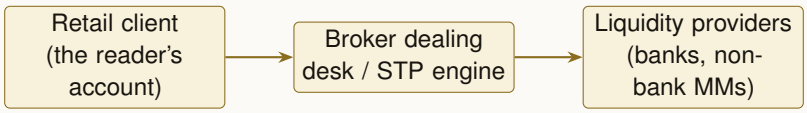
**Dealers managing inventory.** A dealer who has just bought a large client order is now net long and wants to lay that inventory off, typically in the interbank tier, well before a market move can turn a filled client order into a loss on the dealer's own book. A great deal of interdealer volume is exactly this — inventory passed from bank to bank as each tries to return to a flatter book, sometimes called "hot-potato" trading (Lyons, 2001) — which is why interbank turnover so far exceeds the underlying client-driven trades that ultimately originate it: one client trade can generate several interdealer trades as inventory is repeatedly passed along before it settles somewhere.

**Central banks.** Central banks trade FX for reasons no private participant shares: managing reserves, and occasionally intervening directly to influence the exchange rate itself. Sarno and Taylor (2001) catalogue the range of intervention styles — outright buying or selling to lean against a move, signalling intent through the size or timing of an intervention without necessarily moving the rate much on the day, and purely verbal intervention ("jawboning") that tries to shift expectations with no trade at all — and the sterilised/unsterilised distinction (whether the domestic money-supply effect of the FX operation is offset elsewhere) that determines whether an intervention is a monetary-policy action in disguise or a pure exchange-rate operation. The extreme case is not really intervention at all but its sudden withdrawal: on 15 January 2015 the Swiss National Bank discontinued, with no advance warning, the minimum EUR/CHF exchange rate it had defended for over three years (Swiss National Bank, 2015), and the franc repriced within minutes by a magnitude few participants had position limits or margin to survive — a full treatment, including what it did to leveraged retail accounts caught on the wrong side, waits for Chapter 15; the point to take now is only that a central bank's own commitment can be the single largest source of one-sided, correlated order flow the market ever sees, right up until the instant it stops being true.

**Benchmark fixes.** Asset managers and index-tracking funds that must value portfolios against a single reference rate concentrate their FX conversions into short, scheduled windows built around published fixes, the 4pm London fix (administered under the WM/Refinitiv methodology) being the best known. Concentrating large, largely price-inelastic flow into a short window creates a predictable liquidity and volume spike around the fix each day; the fixing window itself was widened by industry reform following a mid-2010s benchmark-rigging scandal, so its exact current width and administration should be verified against current methodology rather than assumed — what is structurally stable is only the function the fix serves, not its precise mechanics.

## 4.6 Retail FX and where OANDA sits

Every order the SST platform places enters the market through a single, short chain: client, then broker, then whatever liquidity the broker itself sources.




The client never reaches, and never needs to reach, the liquidity-provider tier directly; the broker is the counterparty to every client trade, full stop, regardless of what it subsequently does with the resulting exposure.

**Definition — A-book and B-book**

In an *A-book* model, the broker passes client flow through to its own liquidity providers (hedging it essentially trade-for-trade), earning a markup on the spread or a commission, and carrying little directional risk on client positions. In a *B-book* model, the broker internalises client flow, becoming the actual counterparty at risk: a client's loss is, directly, the broker's gain, and vice versa. Most brokers run a hybrid, routing flow they judge likely to be informed (large, fast-money, or historically profitable accounts) A-book, and warehousing flow they judge likely to be uninformed B-book — a mechanically rational segmentation given the adverse-selection logic of who trades against whom, not a claim about any individual client's skill.

Neither model is inherently improper; both are standard, disclosed industry practice, and a broker's choice of model for a given account says more about that broker's risk management than about the fairness of any single fill. What does not change across either model is the reason the spread exists at all: it is the retail price of immediacy, compensation to whichever party is standing ready to take the other side of your order right now, for the inventory risk and adverse-selection risk that immediacy exposes it to — the same components any market maker's quote must price in, retail or interbank.

Leverage available to a retail account is capped by regulation, and the cap varies by jurisdiction and has moved materially over the past decade; every figure here is indicative and should be verified against current rules before being relied on. Australia's ASIC restricts retail leverage tiered by instrument class, with major currency pairs capped around 30:1 (indicative, per ASIC's product intervention order taking effect in 2021); the EU's equivalent ESMA regime caps major pairs at the same 30:1; U.S. rules under the CFTC/NFA are looser, around 50:1 on majors. None of these numbers is a law of markets — each is a policy choice a regulator can and does revise, and the only safe practice is to check the current instrument specification rather than transcribe a textbook figure into a position-sizing calculation.

**Trap — Your quote stream is one dealer's view**

Every price the platform has ever stored is OANDA's own bid and ask (or their mid), not "the" market's. Another broker, at the same instant, may show a materially different spread, and occasionally a different mid, especially in the thin or fast conditions §4.4 describes. A backtest built on this feed is a backtest of what OANDA would have filled you at, not of some Platonic market price that happens to coincide with it — an honest description of the object being modelled, not a limitation to apologise for, but one that changes what "the price" in a backtest even means. Exercise 4.5 works through the consequence in full.

## 4.7 Exercises

**Exercise 4.1 — Pip-value drills on an AUD account**

Suppose the reader opens his OANDA account in AUD rather than USD. For a one-standard-lot ($N = 100{,}000$) position in each pair below, find the pip value in AUD.

(a) GBPAUD.

(b) AUDUSD, given AUDUSD $= 0.6510$.

(c) EURUSD, given AUDUSD $= 0.6510$ (no direct EURAUD quote is available).

In each case, state which of the pip-value result's three conventions (§4.3) applies, and comment on why (b) and (c) come out identically.

▷ working space

**Solution — Exercise 4.1**

**(a)** GBPAUD's quote currency is AUD, the account currency: convention one, no conversion needed.

$$\text{pip value} = 0.0001 \times 100{,}000 = \$10.00 \text{ AUD}.$$

**(b)** AUDUSD's base currency is AUD, the account currency: convention two. The pip value in the quote currency (USD) is $0.0001 \times 100{,}000 = \$10.00$, independent of the spot rate — the pair's own spot rate never enters the pip value expressed in its own quote currency, it only enters at the conversion step. Converting to AUD by dividing by the given rate:

$$\text{pip value} = \frac{10.00}{0.6510} = \$15.3610 \text{ AUD}.$$

**(c)** Neither leg of EURUSD is AUD: convention three. The pip value in USD is again $0.0001 \times 100{,}000 = \$10.00$ (EURUSD's own spot rate is not needed at all). Converting USD to AUD requires the reciprocal of the given AUDUSD rate, $1/0.6510 = 1.53610$ AUD per USD:

$$\text{pip value} = 10.00 \times \frac{1}{0.6510} = \$15.3610 \text{ AUD}.$$

(b) and (c) agree exactly because both reduce, after the first step, to "convert a \$10.00 pip value into AUD via the same AUDUSD rate" — which leg of the original pair happened to be USD is irrelevant once the pip's worth has been expressed in USD. (All three figures verified in `verify_mk04.py`, §2.)

**Exercise 4.2 — Triangulating EURCHF, and pricing the arbitrage**

A dealer quotes EURUSD 1.0836/1.0840 and USDCHF 0.8845/0.8849. USD is the quote currency in the first pair and the base currency in the second, so the triangulation result's *product* case (§4.3) applies: EURCHF = EURUSD × USDCHF.

(a) Derive the synthetic EURCHF bid and ask.

(b) A direct EURCHF quote of 0.9600/0.9604 is available. Does a riskless round trip exist, and if so, in which direction?

(c) If it exists, compute the round-trip profit, in EUR, on a EUR 1,000,000 notional.

▷ working space

**Solution — Exercise 4.2**

**(a)** To sell EUR for CHF synthetically (EUR→USD→CHF), sell EUR at the EURUSD bid and sell the resulting USD at the USDCHF bid:

$$\text{synthetic bid} = 1.0836 \times 0.8845 = 0.958444.$$

To buy EUR with CHF synthetically (CHF→USD→EUR), buy USD with CHF at the USDCHF ask and buy EUR with USD at the EURUSD ask:

$$\text{synthetic ask} = 1.0840 \times 0.8849 = 0.959232.$$

**(b)** The direct bid, 0.9600, exceeds the synthetic ask, 0.959232: selling EUR direct for CHF and rebuilding EUR synthetically nets a positive gap, $0.9600 - 0.959232 = 0.000768$ CHF per EUR (about 7.68 pips), so a riskless round trip exists in that direction — sell direct, buy back synthetically.
**(c)** Starting from EUR 1,000,000: sell it direct at 0.9600 for CHF 960,000.00; buy USD with that CHF at the USDCHF ask, $960{,}000/0.8849 =$ USD 1,084,868.35; buy EUR with that USD at the EURUSD ask, $1{,}084{,}868.35/1.0840 =$ EUR 1,000,801.06. The round trip nets

$$1{,}000{,}801.06 - 1{,}000{,}000.00 = \text{EUR } 801.06,$$

about 0.08% of the round-tripped notional, before any cost of actually executing the three legs — consistent with §4.3's warning ("The mid-price gap is not a profit") that a real arbitrage this size is competed away almost immediately once execution costs are counted. (All figures verified in `verify_mk04.py`, §3.)

**Exercise 4.3 — Rollover across a different week**

Take the trading week of Monday 12 October 2026 to Friday 16 October 2026.

(a) Build the T+2 value-date table for trades dealt on each day of the week.
(b) A position is held continuously from Monday's open to Friday's close. On which nights does it incur rollover financing, and which single night is charged triple?
(c) A second position is opened at 3pm New York on Thursday and closed at 2pm New York on Friday. How many rollovers does it incur, and is any of them tripled?
(d) A third position is opened at 3pm New York on Wednesday and closed at 10am New York the next day. How many rollovers does it incur, and is any of them tripled?

▷ working space

**Solution — Exercise 4.3**

**(a)** Skipping weekends exactly as in Worked Example 4.3:

| Trade day | Date | T+2 value date |
|---|---|---|
| Monday | 12 Oct | Wednesday 14 Oct |
| Tuesday | 13 Oct | Thursday 15 Oct |
| Wednesday | 14 Oct | Friday 16 Oct |
| Thursday | 15 Oct | Monday 19 Oct |
| Friday | 16 Oct | Tuesday 20 Oct |

identical in pattern to Worked Example 4.3, as the accompanying result box predicted.
**(b)** Every night Monday through Thursday is a rollover (Monday's close through Thursday's close, four rollovers in total); the value-date deltas are 1, 1, 3, 1 calendar days respectively, so the Wednesday-to-Thursday rollover (value date jumping from Friday 16 Oct to Monday 19 Oct) is the tripled one, exactly as before.
**(c)** This position is open at exactly one 5pm New York instant: Thursday's (it did not exist before Thursday's rollover, and it is closed before Friday 5pm, which in any case is the week's close, not a rollover). One rollover, the Thursday-to-Friday roll, delta one calendar day — not tripled.
**(d)** This position is open at exactly one 5pm New York instant too: Wednesday's (opened 3pm Wednesday, before that day's rollover; closed 10am Thursday, before Thursday's rollover would next fire). One rollover — and it is the Wednesday one, tripled, despite the position having been held for under 24 hours. Duration in wall-clock hours is irrelevant; only which rollover instants a position straddles decides the charge. (All dates and deltas verified in `verify_mk04.py`, §4.)

**Exercise 4.4 — Ranking the liquidity clock**

Six timestamps, all given in AEST (no daylight saving, the reader's own local clock):

| Label | AEST timestamp |
|---|---|
| (i) | Tuesday 01:00 |
| (ii) | Tuesday 18:30 |
| (iii) | Wednesday 21:00 |
| (iv) | Thursday 14:00 |
| (v) | Saturday 05:30 |
| (vi) | Thursday 08:30 |

Rank them from tightest to widest expected spread, with reasons.

▷ working space

**Solution — Exercise 4.4**

Converting each to UTC (subtract ten hours, no daylight saving on either side) and reading off the nominal session bucket:

| Label | UTC (nominal) | Bucket |
|---|---|---|
| (i) | Monday 15:00 | London–New York overlap |
| (ii) | Tuesday 08:30 | Tokyo–London overlap |
| (iii) | Wednesday 11:00 | London (pure) |
| (iv) | Thursday 04:00 | Tokyo (pure) |
| (v) | Friday 19:30 | NY late |
| (vi) | Wednesday 22:30 | rollover_hour (Wednesday's, the tripled one) |

Ranked tightest to widest: **(i)** $<$ **(ii)** $<$ **(iii)** $<$ **(iv)** $<$ **(v)** $<$ **(vi)**. (i) is the day's liquidity peak, both dominant centres live at once. (ii) still benefits from two sessions overlapping, though Tokyo adds less depth than New York does in (i). (iii) is the deepest single session on its own. (iv) is Asia, thin and rangey by comparison with any London-inclusive window. (v) has lost London's depth with nothing yet arrived from Asia. (vi) is not really a "session" at all: it is the specific instant liquidity is briefly withdrawn while books re-value for the roll, compounded here by being the week's triple-charged Wednesday instant — note in passing that (vi), read only as "Thursday 08:30 AEST," is easy to misfile as an unremarkable Thursday-morning timestamp unless it is first converted back to the New York Wednesday it actually is (Trapbox 4.4). (All UTC/AEST conversions verified in `verify_mk04.py`, §7.)

**Exercise 4.5 — Essay: what "no consolidated tape" does to a backtest's price**

In two or three paragraphs, explain why the absence of a consolidated tape in FX — unlike, say, a single-exchange equity market — changes what "the price" means for a historical backtest, and why that matters more, not less, at exactly the moments §4.4 identifies as thin.

▷ working space

**Solution — Exercise 4.5**

An equity backtest built on a single exchange's tape is, at minimum, arguable as a backtest of "the" price, because a consolidated tape is by construction the record of every trade that counted as a transaction in that market. FX has no such record: there is no single venue, so there is no single price, only however many dealers happen to be quoted at a given instant, each with its own bid and ask, agreeing closely most of the time and diverging exactly when it matters most. A backtest built on the platform's stored OANDA candles is therefore not an approximation to some more fundamental "true" FX price with a single small error term; it is a faithful record of one dealer's own historical quotes, and nothing in the data says how far another dealer's quotes sat from them at the same instant.

This is precisely why §4.4's liquidity clock matters to backtest honesty, not only to live cost budgeting: dealers' quotes are closest together exactly when liquidity is deepest (the London–New York overlap) and can diverge most exactly when it is thinnest (the Asian session, the rollover instant, the weekend reopen) — which is to say, cross-dealer dispersion is itself session-conditioned, in the same direction as the spread widening [Course Book §18.3] already documents for OANDA's own quotes in isolation. A strategy whose edge concentrates in thin sessions is therefore doubly exposed: its own broker's spread is wider there, *and* the assumption that its historical fills would have looked the same at a different dealer is weakest there too.

This is the same lesson Chapter 2's discussion of mid-only candles reaches from the opposite direction: collapsing a dealer's own bid and ask into a single mid-price series throws away exactly the object market microstructure studies, even before asking whether that dealer's mid agreed with anyone else's. "No consolidated tape" does not make a backtest worthless; it makes the backtest's price an explicit, falsifiable claim about one counterparty's historical willingness to trade, and the honest response is to keep that claim visible — store and use the bid/ask actually quoted, condition cost estimates on the session, and never report a result as if it were independent of whose quotes generated it.

**Key idea — What to remember**

The foreign exchange market is a tiered OTC dealer network with no consolidated tape, its turnover dominated not by speculative spot trading but by FX swaps — the money market's cross-currency plumbing, priced by covered interest parity, and the same mechanism that shows up on a retail statement as nightly rollover. Pip value is a single formula in three guises, distinguished only by whether the account currency happens to be the quote currency, the base currency, or neither; cross rates are pinned to their synthetic value, via exactly the algebraic identity [Course Book §16.5] uses for a different purpose, to within the cost of the alternative route. The trading week is genuinely continuous but very much not uniform: liquidity peaks at the London–New York overlap and troughs at the rollover instant that a T+2 value-date convention forces to carry triple weight every Wednesday in New York — Thursday morning, inconveniently, on the reader's own AEST clock. Every quote behind all of this is one dealer's own bid and ask, OANDA's included, which is what §4.6 means by describing a backtest's price as a claim about a counterparty rather than a fact about the market. Chapter 5 turns from this mechanical anatomy to the economics underneath it: why exchange rates move at all, what a rate differential predicts and over what horizon, and why the carry trade is a systematic bet against a parity condition that fails on average but not always.

# 5

# The Economics of Exchange Rates

*Chapter 4 described the machine; this chapter describes what drives it. An exchange rate is a relative price between two monies, and almost everything that moves it can be organised around a single arbitrage identity and the one place that identity is allowed to fail. Covered interest parity is derived here from no-arbitrage — it holds by construction, and where it visibly does not, the deviation is itself a tradeable price. Its uncovered cousin is not an arbitrage at all but a hypothesis about expectations, and it fails systematically: the carry trade is the industrial harvesting of that failure, together with the crash risk that is the failure's price. Purchasing power parity and the balance of payments anchor the rate over years; central-bank policy expectations dominate it over the days and weeks in which the platform actually forecasts. The chapter closes on the honest question — which of these drivers is exploitable at which horizon — and answers it with arithmetic rather than assertion, arriving at the conclusion the whole programme rests on: economics explains a great deal of exchange-rate variance and almost none of its conditional mean, so the tradeable residue is the distribution, not the point.*

## 5.1 Covered interest parity

Two currencies, two interest rates, a spot rate and a forward rate: four prices, and only three degrees of freedom. The fourth is determined by the other three, not by anyone's opinion about the future, and the argument that pins it is the cleanest no-arbitrage derivation in this book. It is worth doing slowly, because everything else in the chapter is defined by its relationship to this result — the carry trade is a bet against its uncovered analogue, the platform's nightly financing charge is its retail-sized instance, and the one persistent modern violation of it (§5.1) is a price of balance sheet rather than a free lunch.

### Notation, fixed once

Following Chapter 4's convention, a pair BASE/QUOTE is quoted as $S$ units of the quote currency per one unit of the base currency, and this chapter keeps the same labels rather than importing the textbook's ambiguous "domestic/foreign":

- $S$ — the spot rate, quote-currency units per one base-currency unit.
- $F$ — the outright forward rate for a value date $T$ ahead, same units.
- $r_q$ — the simple annualised interest rate at which the *quote* currency can be borrowed and lent over the period, on that currency's own money-market day-count basis.
- $r_b$ — the same for the *base* currency.
- $\tau_q, \tau_b$ — the accrual fractions of a year for the period, computed on each currency's own convention (USD, EUR and JPY money markets conventionally accrue ACT/360; AUD, NZD and GBP accrue ACT/365). These are *not* interchangeable, and §5.1 shows that assuming they are misprices a six-month forward by nearly two pips.

Write the two gross accrual factors as $A_q = 1 + r_q\tau_q$ and $A_b = 1 + r_b\tau_b$: one unit of quote

currency lent today returns $A_q$ units at $T$, and likewise for the base currency. All four rates are taken as riskless and available to the same agent in unlimited size — the idealisation whose failure §5.1 prices.

**Result — Covered interest parity**

With no bid-ask spread, no credit risk, no capital charge and free borrowing in both currencies,

$$\boxed{\frac{F}{S} = \frac{A_q}{A_b} = \frac{1 + r_q \tau_q}{1 + r_b \tau_b}} \qquad \Longleftrightarrow \qquad F = S \frac{1 + r_q \tau_q}{1 + r_b \tau_b}.$$

*Proof.* Consider two ways to convert one unit of quote currency today into quote currency at $T$, both of which are riskless and require no capital beyond that one unit.
*Route 1 (stay home).* Lend the unit at $r_q$. At $T$ it is worth $A_q$ units of quote currency. No exchange rate ever enters.
*Route 2 (go abroad, hedged).* Buy $1/S$ units of base currency spot. Lend them at $r_b$, so at $T$ you hold $A_b/S$ units of base currency — a quantity known *today*, since $A_b$ and $S$ are both known today. Simultaneously, today, sell exactly $A_b/S$ units of base currency forward at the rate $F$. At $T$ you deliver the matured base-currency balance into that forward and receive $(A_b/S)\,F$ units of quote currency.
Route 2 has no residual exchange-rate exposure: the forward sale was sized to the *known* maturity value of the base-currency deposit, not to an estimate of it, which is exactly what "covered" means. Both routes start from one unit of quote currency, neither requires further capital, and both terminate in a quantity of quote currency fixed at inception. If $A_q > (A_b/S)F$, borrow abroad and lend at home — that is, run Route 2 in reverse and Route 1 forward — and bank the difference with no capital and no risk; if $A_q < (A_b/S)F$, do the reverse. Neither is possible in equilibrium, so

$$A_q = \frac{A_b}{S} F \quad \Longrightarrow \quad F = S \frac{A_q}{A_b}. \qquad \blacksquare$$

*Continuous form.* If instead the rates are continuously compounded, $r_q^c$ and $r_b^c$, over a common year fraction $T$, the accrual factors become $e^{r_q^c T}$ and $e^{r_b^c T}$ and the identity reads

$$F = S\, e^{(r_q^c - r_b^c)T},$$

which is the form Hull (2021) states for a foreign currency treated as an asset paying a known continuous dividend yield $r_b^c$ — the same result in the language of forward pricing rather than of money markets.

Three features of the proof deserve emphasis, because each is a place where practitioners misread the result.

First, *no expectation appears anywhere*. The word "forward" invites the reading that $F$ is the market's forecast of $S_T$, and it is not: $F$ is a spot rate plus a financing adjustment, assembled from three prices all observable today. A dealer quoting a forward is not expressing a view. Whether $F$ nonetheless *happens* to forecast $S_T$ is the entirely separate question §5.2 takes up, and the answer is no.

Second, the derivation is symmetric in which currency you call home. Running the identical argument starting from one unit of base currency yields $S/F = A_b/A_q$, the same equation. This matters operationally: a reader who habitually thinks in AUD and a counterparty who thinks in USD must agree on the forward or one of them can arbitrage the other, and the identity is what guarantees they do.

Third, the identity is enforced by the FX swap, not by outright forwards. A dealer with a

mispriced forward is arbitraged by the combination *spot leg + opposite far leg*, which is precisely the FX swap of Chapter 4 — which is why forward pricing is done, in practice, by quoting the swap points directly and letting the outright fall out.

**Definition — Forward points**

The *forward points* (equivalently *swap points*) for a tenor are the difference $F - S$, expressed in pips of the pair. Rearranging Result 5.1,

$$F - S \;=\; S\,\frac{A_q - A_b}{A_b} \;=\; S\,\frac{r_q\tau_q - r_b\tau_b}{1 + r_b\tau_b},$$

so the points are positive — the base currency is at a forward *premium* — exactly when the quote currency's accrual exceeds the base currency's. Dropping the denominator gives the familiar linearisation $F - S \approx S\,(r_q\tau_q - r_b\tau_b)$, and imposing a common day count gives the textbook $F - S \approx S(r_q - r_b)T$; both understate the accuracy needed for a real quote, as §5.1 shows.

Market convention quotes points two-way, and the sign is carried by the *order* of the two numbers rather than by an explicit minus. If the quoted pair ascends — "27.5/29.0" — the points are added to the corresponding spot side; if it descends — "29.0/27.5" — they are subtracted. The rule is not arbitrary: it is the unique reading under which the resulting outright bid remains below the outright ask, so a dealer who applies it correctly never quotes a crossed forward market. Exercise 5.2 works both directions. Note also that the outright spread is the spot spread *plus* the points spread: a forward is quoted as two prices, each carrying its own bid-ask, and hedging a position forward therefore costs strictly more than hedging it spot.

**Worked Example 5.1 — A six-month AUDUSD forward, with day counts respected**

A hypothetical scenario, not market data. Spot AUDUSD is $S = 0.6520$ (USD per AUD). The six-month USD deposit rate is 4.60% on an ACT/360 basis; the six-month AUD deposit rate is 4.10% on ACT/365. There are 182 calendar days from the spot value date to the forward value date. Find the outright forward, the forward points, and the differential the points actually imply.

**Accrual factors.** AUDUSD has AUD as base and USD as quote, so $r_q = 4.60\%$ (USD) and $r_b = 4.10\%$ (AUD), with

$$\tau_q = \tfrac{182}{360} = 0.505556, \qquad \tau_b = \tfrac{182}{365} = 0.498630,$$

$$A_q = 1 + 0.0460 \times 0.505556 = 1.023256, \qquad A_b = 1 + 0.0410 \times 0.498630 = 1.020444.$$

**Forward.**

$$F = S\,\frac{A_q}{A_b} = 0.6520 \times \frac{1.023256}{1.020444} = 0.6520 \times 1.0027554 = 0.653797 \;\;(\text{quoted } 0.6538).$$

**Points.** $F - S = 0.001797$, i.e. +17.97 pips: the AUD trades at a forward *premium* of about eighteen points against the USD, because the USD accrual is the larger of the two. Note the direction — the *lower*-yielding currency of the pair trades at a forward premium, which is Result 5.1 restated and the seed of the whole carry argument in §5.2.

**Why the textbook approximation is not good enough.** The common shortcut $S(r_q - r_b)T$ with $T = 0.5$ gives

$$0.6520 \times (0.0460 - 0.0410) \times 0.5 = 0.001630 = 16.30 \text{ pips},$$

which is 1.67 pips light — on a pair whose spot spread might be one pip, an error of well over a spread. Almost all of the gap is the day-count mismatch, not the linearisation. Restating the USD rate on the AUD's ACT/365 basis gives $4.60\% \times 365/360 = 4.6639\%$, so the *comparable* differential is 56.39 basis points rather than the 50 the headline rates suggest, and

$$0.6520 \times 0.005639 \times 0.498630 = 0.001833 = 18.33 \text{ pips},$$

now only 0.37 pips heavy — the residual being exactly the denominator $A_b$ that Definition 5.1 drops.

**The differential the points imply.** Inverting, the annualised premium embedded in the quote is

$$\frac{1}{\tau_b}\left(\frac{F}{S} - 1\right) = \frac{0.0027554}{0.498630} = 0.005526 = 55.26 \text{ bp},$$

slightly below the 56.39 bp comparable differential because dividing by $A_b$ compounds the base leg. This number, not the difference between two policy rates on a screen, is the differential a hedger actually pays or receives.

**Continuous check.** Converting to continuous rates over $T = 182/365$, $r_q^c = \ln(A_q)/T = 4.6105\%$ and $r_b^c = \ln(A_b)/T = 4.0587\%$, and $F = 0.6520\, e^{(0.046105 - 0.040587)\times 0.498630} = 0.653797$ — identical to twelve decimal places, as it must be, since the continuous form is a re-parametrisation and not a different model. (All figures verified in `verify_mk05.py`, §1.)

The arbitrage the result forbids is worth pricing once, so that "no-arbitrage" is not merely a phrase. Suppose a dealer, in the scenario above, quoted the six-month forward at 0.6560 rather than 0.653797 — too high by 22 pips. Borrow \$1,000,000 USD at 4.60%; buy $1{,}000{,}000/0.6520 = 1{,}533{,}742.33$ AUD spot; lend it at 4.10% to mature at 1,565,097.91 AUD; and sell that exact quantity forward at 0.6560 for \$1,026,704.23. The USD loan matures at \$1,023,255.56. The difference, \$3,448.67, is riskless, requires no capital, and is known at inception — which is why the quote does not survive. That \$3,448 on a \$1m notional is 34.5 basis points, an enormous number by the standards of a market where the entire six-month premium is 55 basis points.

### When parity itself is the price: the cross-currency basis

The derivation assumed unlimited riskless borrowing in both currencies at a single rate. Since 2008 that assumption has been false in a way that shows up directly in prices. Measure the gap by rearranging Result 5.1 into a residual: the *cross-currency basis* is the spread that must be added to one leg's interest rate to make the observed $F$, $S$, $r_q$ and $r_b$ satisfy the identity. Before the global financial crisis this residual was small enough in the major pairs to be absorbed by transaction costs. Since then it has been persistently non-zero, typically negative against the US dollar — synthetic dollar funding obtained via the FX swap market costs more than direct dollar borrowing — with the gap widening at quarter-ends and balance-sheet reporting dates and blowing out sharply in dollar-funding stress, notably in March 2020 (BIS, 2020, indicative; magnitudes vary by pair, tenor and date, and should be verified against current data rather than assumed).

The mechanism is not a failure of the algebra but a failure of its premise. Executing the arbitrage above requires a balance sheet: someone must borrow in one currency, hold a deposit in another, and carry the resulting gross positions through a regulatory reporting date, consuming leverage-ratio capacity and counterparty credit lines that are neither free nor unlimited. That is precisely the setting of Brunnermeier and Pedersen (2009), in which arbitrageurs' capacity to enforce a price relationship is itself a function of their funding conditions, so that the relationship holds loosely in calm periods and visibly fails when funding tightens — exactly the observed time profile of the basis. Duffie's (2012) account of over-the-counter markets supplies the complementary point: enforcement requires a counterparty willing to intermediate, and intermediation capacity is a

scarce, priced input rather than a free background condition. The honest summary is that covered interest parity is an identity among four *frictionless* prices, and the basis measures the shadow price of the frictions.

**Trap — The differential on the screen is not the differential you can trade**

Three different numbers get called "the interest rate differential" and they are not equal. There is the policy-rate difference (a central-bank announcement); the money-market deposit-rate difference at your tenor (what §5.1 used); and the differential actually embedded in tradeable forward points, which differs from the second by the cross-currency basis and from the first by the whole term structure of expectations. For a covariate feed, pick one, define it explicitly, and never let a backtest silently substitute one for another across a sample period in which the basis moved — the substitution introduces a slow-moving, regime-dependent error into exactly the variable the model is being asked to learn from.

## 5.2 Uncovered interest parity and the carry trade

Result 5.1 required nothing of anybody's beliefs, because the forward sale removed the exchange-rate risk entirely. Delete that forward sale and the argument changes character completely. The investor who converts to the base currency, lends at $r_b$, and simply *waits*, planning to convert back at whatever spot rate prevails at $T$, is now exposed. If he is nonetheless indifferent between that and lending at home, it must be because the expected conversion rate compensates him exactly:

$$A_q \;=\; \frac{A_b}{S}\,\mathbb{E}_t[S_T] \qquad\Longleftrightarrow\qquad \frac{\mathbb{E}_t[S_T]}{S} \;=\; \frac{A_q}{A_b}.$$

This is *uncovered interest parity*. It is not a theorem. It is the conjunction of two substantive assumptions — that expectations are rational, and that investors demand no risk premium for bearing currency risk — and either can fail. Comparing it with Result 5.1 gives the compact statement that organises everything below:

**Result — UIP, restated against the forward**

Given covered interest parity — which arbitrage enforces — uncovered interest parity holds if and only if the forward rate is an unbiased forecast of the future spot rate:

$$\underbrace{\frac{F}{S} = \frac{A_q}{A_b}}_{\text{CIP (arbitrage)}} \quad\text{and}\quad \underbrace{\frac{\mathbb{E}_t[S_T]}{S} = \frac{A_q}{A_b}}_{\text{UIP (hypothesis)}} \quad\Longleftrightarrow\quad F \;=\; \mathbb{E}_t[S_T].$$

Since the left-hand identity is not in question, *every* empirical rejection of "the forward rate is an unbiased predictor of the future spot rate" is a rejection of UIP, not of CIP. In logs, with $s = \ln S$ and $f = \ln F$, UIP says

$$\mathbb{E}_t[s_{t+1} - s_t] \;=\; f_t - s_t \;\approx\; r_q\tau_q - r_b\tau_b,$$

i.e. the currency with the *higher* interest rate is expected to depreciate by exactly the differential, leaving no expected gain from chasing yield. Two approximations enter in moving to logs: the linearisation of the accrual ratio, and the Jensen gap between $\ln \mathbb{E}_t[S_T]$ and $\mathbb{E}_t[\ln S_T]$. Both are of order the variance of the log rate over the period and are small at these tenors relative to the effect being measured — but the second is why a careful statement of "unbiasedness" must say whether it is meant in levels or in logs.

The economics is intuitive and, stated that way, almost compelling: if Australian deposits pay 4.10% and Japanese deposits pay 0.50%, and everyone could see a free 3.60% in switching, they would switch until the Australian dollar was bid up to a level from which its expected decline exactly cancelled the pickup. The trouble is that this is not what the data do.

### The forward-premium puzzle

The canonical test regresses the realised change in the log spot rate on the forward premium:

$$s_{t+1} - s_t \;=\; \alpha + \beta\,(f_t - s_t) + \varepsilon_{t+1}. \tag{5.1}$$

Under UIP with rational expectations, $\alpha = 0$ and $\beta = 1$: a one-percentage-point forward premium should be matched, on average, by a one-percentage-point spot appreciation of the premium currency. Fama (1984) showed that estimates of $\beta$ are systematically below unity and frequently *negative* across major currencies and sample periods — the *forward-premium puzzle*. Negative $\beta$ says something far stronger than "UIP is imperfect": it says the high-interest-rate currency tends, on average, to *appreciate*, so the yield pickup and the spot move point the same way rather than cancelling.

Fama's own decomposition explains exactly what a negative slope requires, and it is worth reproducing because it converts a stylised fact into two testable statements about second moments. Decompose the forward premium into an expected depreciation and a residual risk premium,

$$f_t - s_t \;=\; p_t + \mathbb{E}_t[\Delta s_{t+1}], \qquad \Delta s_{t+1} \equiv s_{t+1} - s_t,$$

and impose rational expectations, $\Delta s_{t+1} = \mathbb{E}_t[\Delta s_{t+1}] + \eta_{t+1}$ with $\eta_{t+1}$ orthogonal to date-$t$ information. Then the population slope of (5.1) is

$$\beta \;=\; \frac{\operatorname{Cov}(\Delta s_{t+1},\, f_t - s_t)}{\operatorname{Var}(f_t - s_t)} \;=\; \frac{\operatorname{Var}\big(\mathbb{E}_t[\Delta s_{t+1}]\big) + \operatorname{Cov}\big(p_t,\, \mathbb{E}_t[\Delta s_{t+1}]\big)}{\operatorname{Var}(p_t) + \operatorname{Var}\big(\mathbb{E}_t[\Delta s_{t+1}]\big) + 2\operatorname{Cov}\big(p_t,\, \mathbb{E}_t[\Delta s_{t+1}]\big)},$$

the numerator following because $\eta_{t+1}$ is uncorrelated with anything dated $t$. The denominator is a variance and hence positive, so $\beta < 0$ requires

$$\operatorname{Cov}\big(p_t, \mathbb{E}_t[\Delta s_{t+1}]\big) \;<\; -\operatorname{Var}\big(\mathbb{E}_t[\Delta s_{t+1}]\big) \;<\; 0,$$

and, applying the Cauchy–Schwarz bound to that inequality,

$$\sigma_p\,\sigma_{\mathbb{E}[\Delta s]} \;\geq\; \big|\operatorname{Cov}(p, \mathbb{E}[\Delta s])\big| \;>\; \sigma^2_{\mathbb{E}[\Delta s]} \qquad\Longrightarrow\qquad \sigma_p \;>\; \sigma_{\mathbb{E}[\Delta s]}.$$

A negative Fama slope therefore forces two conclusions simultaneously: the risk premium moves *opposite* to expected depreciation, and the risk premium is the *more variable* of the two components of the forward premium. Most of what looks like a forecast in the forward rate is, on this reading, compensation for risk rather than a prediction — which is precisely why a systematic strategy can harvest it and precisely why doing so is not free.

Two caveats belong here rather than in a footnote. First, $\hat\beta$ is unstable: estimates vary materially with the pair, the sample window and the tenor, and several post-2008 samples give slopes closer to zero or above it, so no particular numerical slope should be treated as a constant of nature (indicative; re-estimate on your own sample rather than quoting a literature value). Second, the regression's $R^2$ is small in every sample — a few percent at monthly horizons is typical — so even a correctly signed, strongly negative slope describes a faint tilt in the conditional mean of a very noisy variable, not a forecast. Exercise 5.4 turns both caveats into arithmetic.

> **Definition — The carry trade**
>
> A *carry trade* is long the high-interest-rate currency and short the low-interest-rate currency

of a pair, financed so that the position is self-funding: borrow the low-yielder, lend the high-yielder, and accept the exchange-rate risk that UIP claims will exactly offset the yield pickup. Equivalently, in forward terms, it sells the forward-premium currency forward and buys the forward-discount currency forward. Its log excess return over one period is

$$x_{t+1} \;=\; (f_t - s_t) - \Delta s_{t+1},$$

the differential earned minus the spot move suffered, and by (5.1) its expectation is

$$\mathbb{E}[x_{t+1}] \;=\; -\alpha + (1-\beta)\,(f_t - s_t).$$

The factor $(1-\beta)$ is the *carry multiplier*: zero if UIP holds ($\beta = 1$); one if the spot change is unpredictable from the premium ($\beta = 0$, the case in which the random-walk benchmark of Meese and Rogoff (1983) is exactly right against this regressor, and in which the trade simply collects the differential); and greater than one whenever $\beta < 0$. The carry trade is thus, exactly and literally, a bet that $\beta \neq 1$, sized by how far from one it is.

**Trap — UIP is not a forecasting rule — in either direction**

UIP fails, so do not use it to forecast; but its failure is a statement about a long-run average across many currencies and many months, so do not invert it into a forecast either. "The differential is 3.6%, $\beta$ is about $-0.7$, therefore this pair will rise 6.1% this year" is a misreading of a regression whose $R^2$ is a few percent: the number produced is a conditional mean whose surrounding distribution is roughly an order of magnitude wider, and whose left tail is fatter than its right (Brunnermeier, Nagel and Pedersen, 2008). The correct use of a Fama slope is to size a diversified, risk-managed exposure to a premium, never to predict a single pair's path.

### Carry, in cash

The arithmetic below is a hypothetical scenario computed exactly, not market data, and it is deliberately run in full cash-flow form — borrow, convert, lend, convert back — rather than as a percentage shortcut, because the shortcut hides where the exposure actually sits.

**Worked Example 5.2 — Carry P&L on a long AUDJPY position**

Hypothetical scenario. Spot AUDJPY is $S_0 = 97.50$ (JPY per AUD). Twelve-month rates are $r_{\text{AUD}} = 4.10\%$ and $r_{\text{JPY}} = 0.50\%$, simple, on a matched day count for clarity, so the gross differential is 3.60%. A trader takes one standard lot: long $N = 100{,}000$ AUD, funded in JPY, held for one year.

**Setting up the two legs.** Borrow $NS_0 = 9{,}750{,}000$ JPY, convert to 100,000 AUD, and lend it. At $T$:

$$\begin{aligned}\text{AUD leg} &= 100{,}000 \times 1.0410 = 104{,}100.00 \text{ AUD},\\ \text{JPY debt} &= 9{,}750{,}000 \times 1.0050 = 9{,}798{,}750.00 \text{ JPY}.\end{aligned}$$

**Breakeven.** Closing at spot $S_1$ yields $104{,}100\,S_1 - 9{,}798{,}750$ JPY, which is zero at

$$S^* = S_0\,\frac{1 + r_{\text{JPY}}}{1 + r_{\text{AUD}}} = 97.50 \times \frac{1.0050}{1.0410} = 94.128242,$$

a depreciation of 3.4582%. This is not a coincidence: $S^*$ is exactly the one-year forward rate of Result 5.1, and the statement "the carry trade breaks even at the forward" is Result 5.2 read

backwards. Under UIP, $S^*$ is where the rate is *expected* to end up, and the trade's expected profit is zero.

**Three outcomes.** Decompose each into a spot leg (on the principal) and an interest leg (the AUD coupon converted at the terminal rate, less the JPY coupon):

| Scenario | $S_1$ | Spot move | Spot leg (JPY) | Interest leg (JPY) | Total |
|---|---|---|---|---|---|
| A: unchanged | 97.50 | 0.0000% | 0 | 351,000 | +351,000 |
| B: appreciates | 99.00 | +1.5385% | +150,000 | 357,150 | +507,150 |
| C: crash | 88.00 | −9.7436% | −950,000 | 312,050 | −637,950 |

Converting each total at its own terminal spot and expressing it against the AUD notional:

$$\text{A: } \tfrac{351{,}000}{97.50} = +3{,}600.00 \text{ AUD} = +3.6000\%, \quad \text{B: } \tfrac{507{,}150}{99.00} = +5{,}122.73 = +5.1227\%,$$

$$\text{C: } \tfrac{-637{,}950}{88.00} = -7{,}249.43 \text{ AUD} = -7.2494\%.$$

Scenario A returns the differential exactly — 3.60%, to the cent — which is the sanity check that the cash-flow construction is right.

**What the crash costs, measured in carry.** The 9.7436% spot fall is $9.7436/3.60 = 2.7066$ years of gross carry undone by the spot leg alone; net of the interest earned along the way, the 7.2494% loss is 2.0137 years of carry. One year's bad outcome erases two years' good ones. This is "up the stairs, down the lift" as arithmetic rather than as folklore.

**Two retail corrections.** First, financing. A retail CFD does not pay the interbank differential; the broker applies an administration markup against the client on both sides. At an indicative 1.00% p.a. markup (verify current specifications; brokers publish and revise these), the client's net carry is $3.60 - 1.00 = 2.60\%$, i.e. 2,600 AUD a year against a gross 3,600 — the markup consumes 27.78% of the entire economic edge before a single transaction cost is paid. Second, leverage. At an indicative 30:1 retail cap for major pairs the margin behind this position is $100{,}000/30 = 3{,}333.33$ AUD, and Scenario C's loss of 7,249.43 AUD is 2.1748 times that margin — the position is closed out well before the year ends, converting a paper drawdown into a realised one. (All figures verified in `verify_mk05.py`, §3.)

Two structural features of the decomposition deserve naming. The interest leg is not constant across scenarios — 351,000, 357,150, 312,050 JPY — because the AUD coupon is earned in AUD and converted at the *terminal* rate, so the carry itself is levered to the spot move. And the whole position's exposure is linear in $S_1$ with slope 104,100 rather than 100,000: the carry trade is fractionally *more* exposed to the crash than the notional suggests, because the accrued interest is itself denominated in the currency that crashes.

### Why carry pays, and when it does not

If the carry trade earns a positive average excess return, something must be bearing a risk. The modern literature's answer is that carry is compensation for a specific, identifiable and highly inconvenient risk profile.

Brunnermeier, Nagel and Pedersen (2008) documented that carry returns are strongly negatively skewed — long stretches of modest, steady accrual punctuated by rapid, large losses — and tied the asymmetry to funding conditions: carry positions are levered, crowded and financed, so a tightening in funding liquidity forces simultaneous unwinding, which moves the exchange rate against every holder at once, which triggers further unwinding. The crash is thus endogenous to the trade's own popularity, and its trigger is a funding shock rather than news about Australia or Japan. This is Brunnermeier and Pedersen's (2009) liquidity spiral operating in currency markets, and it is why the loss distribution's left tail cannot be estimated from the right-tail-symmetric

assumptions a Gaussian model would impose.

Menkhoff, Sarno, Schmeling and Schrimpf (2012) sharpened this into the form most useful to a systematic trader: carry returns load negatively on innovations to *global* FX volatility. High-interest currencies do badly precisely when volatility rises unexpectedly, and low-interest funding currencies do well then — so the carry premium is compensation for holding an asset that loses in exactly the states where a diversified investor most wants gains. Lustig, Roussanov and Verdelhan (2011) provide the cross-sectional counterpart: currency portfolios sorted on interest rates load on a common "slope" factor, so carry is a systematic risk exposure rather than a collection of independent pair-level bets, and diversifying across many pairs reduces idiosyncratic noise without reducing the factor exposure that generates both the return and the crash.

The operational content of that last sentence is the single most important thing this section has to offer a systematic FX book. Diversification across carry pairs does not diversify carry risk, because the pairs share the factor. What conditions the payoff is the volatility *state*, not the pair count.

**Your project — RORO and VolRegime are the carry-crash covariates**

The platform already carries a risk-on/risk-off composite and a volatility-regime state. The result above says what those series are structurally *for* in an FX book: not a generic "market mood" feature but the direct empirical handle on the conditional distribution of a differential-driven position. Menkhoff et al.'s (2012) finding is that the sign and magnitude of carry's payoff are conditional on volatility innovations, which means a model given a differential covariate and no volatility-state covariate is being asked to learn an unconditional average across two regimes whose payoffs have opposite signs — the worst possible specification, since the unconditional mean is positive and the conditional means straddle zero. Two concrete implications for the Stage-2 router [Course Book §13.8]: the differential feature should be interacted with, or routed by, the volatility state rather than entered additively; and the envelope's left tail in high-volatility states should be allowed to be materially fatter than a symmetric fit would produce, because the crash mechanism is a funding spiral, not a draw from the same distribution that generated the quiet months. The distributional forecast is the right object here for a reason that is economic, not merely statistical: the mean is regime-dependent and the tail is asymmetric, and no point forecast can represent either.

## 5.3 Long-run anchors: PPP, real exchange rates, and the balance of payments

Interest parity relates an exchange rate to *financial* prices. It says nothing about the level around which the rate wanders — CIP pins $F$ to $S$ but leaves $S$ free, and UIP, even if it held, would only pin the drift. The candidate anchors for the level come from the goods market and from the external accounts, and both operate on timescales measured in years.

### Purchasing power parity

The law of one price says an identical, freely tradeable good must cost the same in two countries once expressed in a common currency, since otherwise buying in the cheap market and selling in the dear one is profitable. Applied to a whole consumption basket, it becomes *absolute purchasing power parity*: with $P_d$ the domestic price of the basket, $P_f$ the foreign price in foreign currency, and $S$ the domestic-per-foreign exchange rate,

$$S = \frac{P_d}{P_f}.$$

Differencing in logs gives *relative* PPP, the far weaker and far more defensible claim that the exchange rate changes at the difference of the two inflation rates:

$$\Delta s \;\approx\; \pi_d - \pi_f.$$

**Definition — Real exchange rate**

The *real exchange rate* is the relative price of the two consumption baskets,

$$Q \;=\; \frac{S\,P_f}{P_d},$$

i.e. the cost of the foreign basket measured in domestic currency, divided by the cost of the domestic basket. Absolute PPP is exactly the statement $Q = 1$; relative PPP is the statement that $Q$ is constant, whatever its level. A real *appreciation* of the domestic currency ($Q$ falling) means domestic goods have become dearer relative to foreign goods, whether through a nominal appreciation, higher domestic inflation, or both. $Q$ — not $S$ — is the variable economic theory expects to be stationary.

The Big-Mac-style comparison popularised by the financial press is absolute PPP applied to a single, deliberately standardised good, and its pedagogical value is that its failures are instructive rather than embarrassing. A hamburger is not tradeable: its price embeds local rent, local wages, local beef, local regulation and local market power, none of which arbitrage across borders. The same objection scales up to the whole basket, and it is the standard explanation of why PPP fails badly in the short run. Non-tradeables are a large share of any consumer price index; transport costs, tariffs and distribution margins put a wedge around the law of one price even for tradeables; the baskets themselves differ in composition between countries; and productivity growth concentrated in the tradeable sector systematically raises the relative price of non-tradeables in fast-growing economies, so the equilibrium real exchange rate is not even constant.

Quantitatively, the finding that matters is not that PPP deviations exist but how slowly they decay. The consensus estimate — stable enough across methods to be worth stating, imprecise enough that the range matters — is a half-life of PPP deviations of roughly three to five years (indicative; the point estimate is sensitive to sample, panel structure and the treatment of the near-unit-root in $Q$, and the associated confidence intervals are notoriously wide). A four-year half-life means that a currency 20% above its PPP value is expected to close

$$1 - 0.5^{1/48} = 1.4337\% \quad \text{of that gap in a month,}$$

an expected move of $0.20 \times 0.014337 = 0.2867\%$, i.e. about 29 basis points. Over one year the closure is $1 - 0.5^{1/4} = 15.9104\%$ of the gap; over five years, 57.9552%. Hold on to those three numbers: §5.5 converts them into the chapter's central honest statement about predictability.

To make the arithmetic concrete, take a hypothetical scenario in which the US basket costs 100 USD, the equivalent Australian basket costs 118 AUD, and AUDUSD trades at 0.6520. Absolute PPP would put the rate at $S_{\text{PPP}} = P_d / P_f = 100/118 = 0.847458$ USD per AUD. The real exchange rate is $Q = 0.6520 \times 118/100 = 0.769360$: the Australian basket costs 76.94 USD against the American basket's 100, so Australia is about 23% "cheap" on this measure, and the AUD sits 23.06% below its absolute-PPP rate. Whether that number is a mispricing or an equilibrium reflecting different productivity structures, terms of trade and non-tradeable shares is exactly the question the level version of PPP cannot answer, which is why serious work uses *deviations of $Q$ from its own mean* rather than deviations from unity.

### Balance of payments: where the flows come from

The second long-run anchor is accounting rather than arbitrage. A country's balance of payments decomposes every cross-border transaction into a current account (trade in goods and services, plus net income and transfers) and a financial account (net purchases of assets), and by construction the two sum to zero net of reserve changes and statistical discrepancy. Every export receipt is a demand for domestic currency; every import payment is a supply of it; every foreign purchase of a domestic bond is a demand for domestic currency, and every domestic purchase of a foreign equity is a supply. The exchange rate is the price that clears the sum.

Three consequences matter for a trader rather than an economist. First, the composition of the financing matters more than its level: a current-account deficit financed by long-horizon direct investment behaves very differently from the same deficit financed by short-term portfolio inflows that can reverse in a week — the reversal risk is what makes the second configuration fragile. Second, for commodity exporters the current account is dominated by the terms of trade, which is the structural reason the Australian dollar co-moves with bulk commodity prices; this is a covariate relationship with a real economic mechanism behind it, not a curve-fit. Third, portfolio flows respond to the interest differentials of §5.4 and to global risk appetite, which is why balance-of-payments logic and carry logic are not competing explanations but the same mechanism observed at different frequencies.

Rey (2015) supplies the modern qualification that keeps this from being a purely bilateral story. Gross capital flows, leverage and asset prices across countries co-move in a *global financial cycle* driven substantially by conditions at the centre — US monetary policy and global risk aversion — rather than by each country's own fundamentals. Her conclusion is that a floating exchange rate does not, by itself, insulate a small open economy from that cycle: the classic trilemma collapses towards a dilemma, in which independent monetary policy requires managing the capital account, not merely letting the currency float. For a systematic FX book the practical implication is direct and slightly deflating: much of what looks like idiosyncratic AUD, NZD or CAD behaviour is a common exposure to a single global factor, so a portfolio of "different" currency positions may carry far less diversification than its instrument count suggests — the same warning §5.2 derived from the carry factor, arriving here by an entirely independent route.

## 5.4 Central banks: the driver that dominates the platform's horizons

PPP works over years and the balance of payments over quarters. At the horizons the platform actually forecasts — hours to days, with covariates updated daily — neither is moving fast enough to matter. What moves fast enough is the market's expectation of monetary policy, and the single most useful organising claim in this chapter is that at those horizons an exchange rate is best understood as the relative price of two expected policy paths.

The reasoning runs straight out of §5.2. Rearrange the log UIP relation into a level statement by iterating it forward: with $s_t$ the log spot rate and $i_q, i_b$ the two short rates,

$$s_t \;=\; \mathbb{E}_t\big[s_{t+H}\big] \;+\; \sum_{h=0}^{H-1} \mathbb{E}_t\Big[i^{(b)}_{t+h} - i^{(q)}_{t+h}\Big]\,,$$

so today's rate is a terminal expectation plus the cumulated expected differential over the intervening path. Two things follow immediately. First, the exchange rate is forward-looking in the strongest sense: it responds to *revisions* of the expected path, not to the current level of rates, which is already in the price. Second, the relevant summary statistic is not one policy rate but the whole expected differential path — which is exactly what a market-implied yield differential at a medium tenor measures.

**Definition — Policy surprise, and why only the surprise moves anything**

A scheduled policy decision has an *expected* component, already embedded in prices before the announcement, and a *surprise* component, the difference between the decision (plus its accompanying communication) and what was priced. Only the surprise can move an asset price at the announcement, because the expected part was priced in when it became expected. Bernanke and Kuttner (2005) made this operational by extracting the surprise from interest-rate futures rather than from surveys, and found equity prices respond to the unexpected component and essentially not at all to the expected one. Gurkaynak, Sack and Swanson (2005) showed that a single surprise number is insufficient: announcements move asset prices along *two* dimensions — a *target* factor, the surprise in the rate set today, and a *path* factor, the revision to the expected trajectory conveyed by the statement — with the path factor doing most of the work at longer maturities. The practical corollary is that a decision can match the consensus exactly and still move the currency several tenths of a percent, because the words changed the path even though the number did not.

Extracting the expected component is a small piece of arithmetic worth doing explicitly. A meeting-dated overnight index swap references the average overnight policy rate over the meeting period, so its rate relative to the prevailing policy rate measures how much of a move is priced. If the policy increment is $\delta$ and the meeting-dated instrument prices $m$ basis points of change, then under the simplifying assumption that the outcome is binary — move by $\delta$ or hold — the implied probability is

$$p = \frac{m}{\delta},$$

because $p\delta + (1-p)\cdot 0 = m$. The assumption is a real restriction: it fails whenever a larger increment or an intermeeting move is genuinely in play, and it silently ignores any term premium in the instrument, so treat $p$ as a market-implied *price*, not a probability estimate. Chapter 4's warning about quoting one dealer's view applies here too, at the level of whose curve you read it from.

### Which differential, at which tenor

The platform's workhorse macro covariate is a two-year government or swap rate differential, and the choice of tenor is neither arbitrary nor a matter of taste. Under the expectations hypothesis of the term structure, a two-year yield is approximately the average expected policy rate over the next two years plus a term premium, so a two-year *differential* between two currencies is approximately the average expected policy *differential* over the cycle ahead — the $\sum_h \mathbb{E}_t[i^{(b)} - i^{(q)}]$ term above, made observable. Shorter tenors are dominated by the next one or two meetings and are therefore jumpy and largely mechanical; much longer tenors are increasingly dominated by term premia and inflation-risk compensation, which move for reasons having nothing to do with the policy cycle. Two years is the tenor at which the ratio of policy-expectation signal to term-premium noise is highest.

The honest qualification is that the expectations hypothesis is itself rejected. Campbell and Shiller (1991) showed that yield spreads predict subsequent short-rate changes with the wrong sign at short horizons — a slope coefficient of the wrong sign in a regression that theory says should give one. The structural resemblance to Fama's (1984) forward-premium result of §5.2 is exact and not coincidental: in both cases a forward-looking price is decomposed into an expectation and a time-varying risk premium, and in both cases the risk premium turns out to be the more variable component, so the regression slope inherits the premium's dynamics rather than the expectation's. A two-year differential is therefore a *contaminated* measure of expected policy — good enough to be the best single macro covariate available, not good enough to be treated as the expectation itself.

**Worked Example 5.3 — A 25 basis-point surprise, priced through to spot**

An illustrative scenario, not a historical episode. Ahead of a scheduled meeting, meeting-dated OIS prices 15 bp of tightening against a policy increment of $\delta = 25$ bp. AUDUSD trades at $S_0 = 0.6520$. The central bank delivers 25 bp with unchanged forward guidance.
**Step 1 — the priced probability and the expected move.**

$$p = \frac{15}{25} = 0.60, \qquad \mathbb{E}[\text{move}] = 0.60 \times 25 + 0.40 \times 0 = 15 \text{ bp}.$$

**Step 2 — the surprise.** Delivered 25 bp against 15 priced:

$$\text{surprise} = 25 - 15 = 10 \text{ bp}.$$

Note what this does *not* say. The bank did exactly what the modal forecaster expected; the surprise arises entirely because the market had assigned 40% weight to a hold.
**Step 3 — pass-through to the two-year differential.** A single meeting's surprise does not shift the whole expected path one-for-one; some of it is a level shift the market expects to persist, some is treated as timing. Take an illustrative pass-through $\kappa = 0.80$ to the average expected policy rate over two years:

$$\Delta(\text{2y differential}) = 0.80 \times 10 = 8 \text{ bp}.$$

$\kappa$ is the empirically interesting quantity here and it is regime-dependent — estimate it on your own sample rather than adopting this illustrative value.
**Step 4 — the UIP-benchmark spot response.** From the iterated relation above, with the two-year-ahead expected level $\mathbb{E}_t[s_{t+H}]$ held fixed, a widening $\Delta$ of the differential sustained over $H$ years appreciates the spot rate immediately by

$$\Delta \times H = 8 \text{ bp} \times 2 = 16 \text{ bp} = 0.1600\%,$$

i.e. $0.6520 \times 0.0016 = 0.001043$, or $+10.43$ pips, to 0.653043.
**Step 5 — the counterfactuals that make the point.** Had the full 25 bp been unexpected, the same arithmetic gives 0.4000%, or 26.08 pips. The 15 bp already priced accounts for 0.2400%, or 15.65 pips — and that part of the move happened *before* the meeting, spread across whatever days the market re-priced it. The announcement itself is entitled only to the residual 10.43 pips, and $26.08 = 15.65 + 10.43$ exactly.
**Step 6 — the honest comparison.** At an indicative 10% annualised volatility, one unconditional daily standard deviation of AUDUSD is $10\%/\sqrt{252} = 0.6299\%$, or about 41.07 pips. The entire correctly-signed, correctly-sized response to a genuine policy surprise is therefore $10.43/41.07 = 0.2540$ of one ordinary day's noise. If a surprise of this magnitude arrived every single day and nothing else moved the rate, the $R^2$ of daily returns on it would be $0.2540^2 = 6.45\%$; but policy meetings occur roughly eight times a year (indicative — meeting frequencies differ by central bank and are periodically revised), or on 3.17% of trading days, so the unconditional contribution to daily $R^2$ is $0.0317 \times 0.0645 = 0.20\%$. The most reliably identified causal driver in international macroeconomics explains about one fifth of one percent of daily exchange-rate variance. (All figures verified in `verify_mk05.py`, §5.)

That last figure is not a defect of the example; it is the chapter's thesis arriving early. Announcement effects are real, replicable and causally identified, and they are still swamped — because they are rare, because most of the information is priced beforehand, and because the unconditional variance of a currency is enormous relative to any single scheduled event. What the announcement *does* reliably change is the conditional *variance*: the minutes around a scheduled

release are the reliably high-volatility, reliably wide-spread part of the day, which is a fact about the distribution rather than about its mean, and therefore exactly the sort of fact a distributional forecaster can use.

### Intervention and exchange-rate regimes

Not every central bank leaves the rate to the market. Regimes form a spectrum: a free float, in which the authority sets rates and lets the currency clear; a managed float, in which it occasionally leans against moves it judges disorderly; a crawling or banded peg, in which it commits to a path or a corridor; and a hard peg or currency board, in which it commits to a level and subordinates monetary policy to defending it.

Sarno and Taylor (2001) survey what official intervention actually achieves and the answer is carefully qualified. *Sterilised* intervention — buying or selling foreign currency while offsetting the effect on the domestic money supply — has no first-order effect through the money stock by construction, so any effect must come through one of three channels: a portfolio balance channel, in which changing the relative supply of the two currencies' assets changes their relative risk premia; a signalling channel, in which the operation conveys information about future monetary policy; and an order-flow or coordination channel, in which the operation works through exactly the microstructure mechanism Evans and Lyons (2002) identify. Their reading of the evidence is that sterilised intervention has at best temporary effects unless it signals a genuine change in policy, and that publicly announced, coordinated operations across several authorities are markedly more effective than unilateral secret ones — which makes sense on the signalling and coordination channels and not on the portfolio-balance one.

The regime spectrum matters for a systematic trader chiefly because of what happens at its discontinuities. A pegged or bounded rate has a return distribution that is compressed almost to nothing for as long as the commitment holds, and then is not. On 15 January 2015 the Swiss National Bank discontinued the minimum exchange rate of CHF 1.20 per euro it had maintained since September 2011 (Swiss National Bank, 2015), and the franc appreciated violently within minutes; prints in that window varied so much across venues that no single number should be quoted as "the" move, and Chapter 15 treats the episode properly. The structural lesson is available now. A managed rate does not remove risk; it warehouses it, converting a continuous distribution into a compressed body plus a catastrophic tail whose timing is a policy decision rather than a market process. Any volatility estimate fitted to the compressed body — any GARCH, any realised-vol measure, any conformal interval calibrated on the quiet period [Course Book §12.6] — is measuring the wrong distribution, and its confidence will be highest exactly when it is most wrong.

**Trap — A quiet series is not a safe series**

Realised volatility falls when a central bank suppresses it, and every risk model that reads volatility from prices will size positions *up* into a managed regime. That is the failure mode: the model is not wrong about the recent past, it is wrong about the mechanism generating it. Whenever an instrument's volatility is low because someone has committed to keeping it low, the honest treatment is a structural override — a hard position cap, or exclusion — and not a tighter interval. The platform holds no pegged pair today; the discipline matters anyway, because the same argument applies in weaker form to any managed float and to any period in which an authority is visibly leaning against a move.

## 5.5 What moves what, at which horizon

Everything above is true simultaneously, and that is the problem. Order flow moves the rate, policy surprises move the rate, carry premia move the rate, and purchasing power parity moves

the rate — but not at the same speed, not with the same reliability, and not in a way that any single model can absorb. The organising discipline is to ask of each driver not "is it real?" but "over what horizon does it accumulate enough signal to exceed the noise, and does the platform trade at that horizon?"

| Horizon | Dominant mechanism | Observable proxy | What is actually forecastable |
|---|---|---|---|
| Seconds–minutes | Order flow revealing dispersed private information; announcement jumps (Evans and Lyons, 2002; Gurkaynak, Sack and Swanson, 2005) | Trade-by-trade signed flow (not in the platform's feed set); the scheduled event calendar | Variance and spread around known events; direction, essentially not at all |
| Minutes–hours | Liquidity clock, dealer inventory management, session hand-overs (Harris, 2003; Chapter 4) | Hour-of-week, session bucket, rollover flag — all deterministic | Spread and variance seasonality, with high reliability; direction, no |
| Days–weeks | Revisions to expected policy paths; risk appetite; positioning crowding | 2y rate differential; RORO composite; VolRegime state; COT net positioning | Conditional variance and regime, strongly; conditional mean, faintly |
| Weeks–months | The carry risk premium and its volatility conditioning; the global financial cycle (Fama, 1984; Menkhoff et al., 2012; Lustig et al., 2011; Rey, 2015) | Level of the differential; global FX volatility innovations | An average excess return with strong negative skew — a premium, not a prediction |
| Months–years | Real-exchange-rate mean reversion; terms of trade; balance-of-payments flows | RER gap versus its own mean; terms-of-trade index; current-account balance | Level mean reversion, with $R^2$ climbing into double digits — at horizons the platform does not trade |

**Table 5.1.** What moves an exchange rate, at which horizon, and what part of the distribution is forecastable from it. The platform forecasts in the second and third rows and holds positions that occasionally reach the fourth.

The table's last column is the part that matters, and the fourth row's phrasing is deliberate. A premium is not a prediction. Knowing that a diversified carry book earns a positive average excess return tells you nothing about next Tuesday, in the same way that knowing equities earn an equity risk premium tells you nothing about next Tuesday — and unlike the equity case, the currency version's payoff is conditional on a volatility state you must estimate.

### How much predictability, exactly

The claim that long-horizon anchors are informative and short-horizon ones are not can be made arithmetic rather than rhetorical, using nothing but the PPP half-life of §5.3 and a volatility assumption.

**Worked Example 5.4 — The predictability of a long-run anchor, by horizon**

Illustrative arithmetic from stated parameters, not an empirical estimate. Assume: PPP deviations decay with a four-year half-life; the real-exchange-rate gap has a standard deviation across time of $\sigma_{\text{gap}} = 15\%$; and the nominal rate has an annualised volatility of 10%, scaling as $\sqrt{h}$.

Over a horizon of $h$ months, the expected fraction of the current gap that closes is

$$\lambda(h) \;=\; 1 - 0.5^{h/48},$$

so the predictable component of the $h$-month return has standard deviation $\lambda(h)\,\sigma_{\text{gap}}$, while the total return has standard deviation $\sigma_h = 10\% \times \sqrt{h/12}$. Treating the predictable component as the fitted part of a regression, the implied coefficient of determination is

$$R^2(h) \;=\; \left(\frac{\lambda(h)\,\sigma_{\text{gap}}}{\sigma_h}\right)^2 .$$

| Horizon | $\lambda(h)$ | Predictable s.d. | Total s.d. | $R^2$ |
|---|---|---|---|---|
| 1 day | 0.0687% | 0.0103% | 0.6299% | 0.027% |
| 1 month | 1.4337% | 0.2151% | 2.8868% | 0.555% |
| 3 months | 4.2397% | 0.6360% | 5.0000% | 1.618% |
| 1 year | 15.9104% | 2.3866% | 10.0000% | 5.696% |
| 3 years | 40.5396% | 6.0809% | 17.3205% | 12.326% |
| 5 years | 57.9552% | 8.6933% | 22.3607% | 15.115% |

The mechanism is entirely transparent: the signal grows roughly linearly in $h$ while the noise grows as $\sqrt{h}$, so the ratio improves as $\sqrt{h}$ until the mean reversion itself saturates. And the cost comparison finishes the argument. At an indicative 1.2-pip half-spread on AUDUSD at 0.6520, a round trip costs $2 \times 1.2 \times 10^{-4}/0.6520 = 0.0368\%$ of notional. The entire predictable daily move from the PPP anchor has a standard deviation of 0.0103% — 3.57 times *smaller* than the cost of acting on it. The anchor first clears one round trip somewhere between the daily and monthly rows. (All figures verified in `verify_mk05.py`, §6.)

Read the table as a statement about where economics lives relative to where the platform trades, and the conclusion is uncomfortable but clean. The drivers with the highest $R^2$ operate at horizons of years, where a CFD book financed nightly cannot hold position; the horizons at which the platform forecasts are dominated by drivers whose $R^2$ for the conditional *mean* is, on the arithmetic above and on §5.3's, a fraction of one percent.

This is Meese and Rogoff's (1983) result restated from first principles. They compared structural exchange-rate models — monetary models, portfolio-balance models, all built on precisely the economics of §§5.3–5.4 — against a naive random walk, out of sample, at horizons up to twelve months, and found the random walk won. The detail that gives the result its force is that they granted the structural models the *realised future values* of their own explanatory variables and the models *still* lost. The failure is therefore not a failure to forecast the fundamentals; it is that the mapping from fundamentals to the exchange rate is too unstable, at those horizons, to beat assuming no change at all.

**Key idea — The predictability statement, stated honestly**

Economics explains a great deal of exchange-rate *variance* and almost none of its conditional *mean* at tradeable horizons. Every driver in this chapter is real: the parity identity is enforced, the carry premium is measurable, policy surprises move rates causally, and real exchange rates do mean-revert. But the conditional mean each one delivers at a one-day horizon is a small fraction of one day's noise and, for the long-run anchors, smaller than one round trip's cost. What the same drivers deliver *reliably* is conditional second and higher moments — volatility clusters around scheduled events, spreads widen at known clock positions, carry's left tail fattens in high-volatility states, and managed regimes compress the body while loading the tail. A point forecast throws all of that away and keeps the one component the evidence says is weakest. This is the same conclusion the Course Book reaches from the statistical side [Course Book §1.3], arrived at here from the economic side, and the agreement of

two independent routes is the strongest form the argument takes.

**Your project — Reading Table 5.1 as a feed specification**

The table is, read column-wise, a specification for the platform's covariate set, and it says three things the current design should be checked against. First, the platform has no order-flow covariate and cannot acquire one from a retail feed — so the entire first row is structurally unavailable, and any residual predictability at that horizon is not addressable, which is a reason to stop looking rather than to look harder. Second, the deterministic covariates of the second row — hour-of-week, session bucket, rollover flag — are free, exactly known in advance, and predict variance rather than direction; they belong in the model as known-future inputs [Course Book §8.4] and should be evaluated against a variance target, not a directional one. Third, the third and fourth rows are where the macro feeds live, and their value is conditional: a differential covariate entered additively, without a volatility-state interaction, is being asked to estimate an unconditional average over two regimes whose payoffs differ in sign. Every macro series in the set (differentials, COT, terms of trade, RER gaps) must additionally be stored point-in-time with its true release lag [Course Book §18.2], because the economics of this chapter is measured in variables that are published late and revised afterwards — the single most common way a plausible macro covariate turns into a leak.

## 5.6 Exercises

**Exercise 5.1 — Derive covered interest parity, then price a forward**

(a) Without looking back at Result 5.1, construct the no-arbitrage argument yourself. Start from one unit of the *base* currency (not the quote currency), build the two riskless routes to base currency at $T$, and show that they force $S/F = A_b/A_q$. State explicitly, at the step where you do it, why the forward transaction must be sized to $A_b$ and not to 1.

(b) A hypothetical scenario. Spot USDJPY is $S = 152.40$ (JPY per USD). The three-month USD deposit rate is 4.35% on ACT/360 and the three-month JPY deposit rate is 0.45% on ACT/365, with 91 days from the spot value date to the forward value date. Compute the outright forward and the forward points in pips, and say which currency trades at a premium.

(c) Compute the linear approximation $S(r_q - r_b)\tau$ with $\tau = 91/360$, and then the version that first restates the USD rate on an ACT/365 basis. Both are wrong in the same direction here, whereas in Worked Example 5.1 the day-count-matched version was the better one. Explain the difference.

▷ working space

**Solution — Exercise 5.1**

**(a)** Start with one unit of base currency. *Route 1*: lend it at $r_b$, holding $A_b$ units of base currency at $T$. *Route 2*: sell it spot for $S$ units of quote currency; lend those at $r_q$, giving $SA_q$ units of quote currency at $T$; and today buy $SA_q/F$ units of base currency forward, delivering the matured quote balance into that forward. Both routes begin with one unit of base currency, are riskless, and require no further capital, so

$$A_b = \frac{SA_q}{F} \quad \Longrightarrow \quad \frac{S}{F} = \frac{A_b}{A_q},$$

which is Result 5.1 inverted — the identity is symmetric in which currency you start from, as it must be, since otherwise two traders on opposite sides of the same trade would disagree about a riskless price.

The sizing point: at $T$ the deposit is worth $A_b$ (or $SA_q$) units, not 1 (or $S$), and a forward struck on the smaller quantity leaves the accrued interest unhedged. That residual is small — of order $r\tau$ times the notional — but it is genuinely exposed to $S_T$, so the position would not be riskless and the no-arbitrage argument would not close. This is exactly the same effect that made the interest leg of Worked Example 5.2 vary across scenarios.

**(b)** USDJPY has USD as base and JPY as quote, so $r_q = 0.45\%$ (JPY, ACT/365) and $r_b = 4.35\%$ (USD, ACT/360):

$$\tau_q = \tfrac{91}{365} = 0.249315, \qquad \tau_b = \tfrac{91}{360} = 0.252778,$$

$$A_q = 1 + 0.0045 \times 0.249315 = 1.001122, \qquad A_b = 1 + 0.0435 \times 0.252778 = 1.010996,$$

$$F = 152.40 \times \frac{1.001122}{1.010996} = 152.40 \times 0.990233 = 150.911582 \ \ (\text{quoted } 150.912).$$

Forward points $= 150.911582 - 152.40 = -1.488418$ JPY, and with the JPY pip of 0.01 that is $-148.84$ pips. The USD — the higher-yielding currency — trades at a forward *discount*, which is Result 5.1 in its most useful mnemonic form: the high-yielder is always cheaper forward.

**(c)** The naive approximation with $\tau = 91/360$:

$$152.40 \times (0.0045 - 0.0435) \times 0.252778 = -1.502410 = -150.24 \text{ pips},$$

overstating the discount by 1.40 pips. Restating USD on ACT/365 gives $4.35\% \times 365/360 = 4.410417\%$, a comparable differential of $-396.04$ bp, and

$$152.40 \times (-0.0396042) \times 0.249315 = -1.504785 = -150.48 \text{ pips},$$

overstating by 1.64 pips — worse, not better.

The reason is that the naive expression makes *two* errors of opposite sign, and matching the day counts fixes only the smaller one. Decompose the gap from the naive figure to the exact one into its two steps:

| | Exercise 5.1 (USDJPY) differential ≈ 396 bp | Worked Ex. 5.1 (AUDUSD) differential ≈ 56 bp |
|---|---:|---:|
| Naive linearisation | −150.24 pips | +16.30 pips |
| + day-count effect | −0.2375 | +2.0324 |
| + denominator effect | +1.6366 | −0.3673 |
| = exact | −148.84 pips | +17.97 pips |
| Dominant term | denominator, 6.89× | day count, 5.53× |

The day-count effect scales with $r\tau\,(365/360-1)$ and so with the *level* of rates; the denominator effect scales with $(F-S)\times r_b\tau_b$ and so with the *differential*. At a 56 bp differential the points are tiny and the denominator barely matters, so matching conventions is the whole correction; at nearly 400 bp the points are two orders of magnitude larger and the denominator dominates by roughly seven to one. The lesson is that the linearisation is fit for a rough check and never for a quote: use the exact points expression

$$F-S = S\,\frac{r_q\tau_q - r_b\tau_b}{1+r_b\tau_b} = 152.40\times\frac{-0.009874}{1.010996} = -1.488418,$$

which reproduces the exact answer by construction and costs one extra division. (All figures verified in `verify_mk05.py`, §7.)

**Exercise 5.2 — Forward points, both conventions, and the differential they imply**

A dealer quotes EURUSD spot at 1.0836/1.0840 and three-month (91-day) swap points at 27.5/29.0.

(a) Determine, from the order of the two point figures alone, whether they are added or subtracted, and construct the outright forward bid and ask.

(b) Compare the outright spread with the spot spread. What does the difference cost a hedger, and why does it exist?

(c) From the mid rates, back out the annualised interest differential the quote implies, on an ACT/360 basis, and state which currency yields more.

(d) Repeat (a) for a quote of 29.0/27.5 and verify that the resulting forward market is not crossed.

(e) A corporate hedger rolls a EUR 1,000,000 exposure for the 91 days at the mid points. What are the points worth in USD?

▷ working space

**Solution — Exercise 5.2**

**(a)** The quote ascends ($27.5 < 29.0$), so the points are added:

$$\text{bid} = 1.0836 + 0.00275 = 1.08635, \qquad \text{ask} = 1.0840 + 0.00290 = 1.08690.$$

The EUR is at a forward premium of roughly 28 points, so by Result 5.1 the USD (the quote leg) is the higher-yielding currency.
**(b)** Spot spread $= (1.0840 - 1.0836)/10^{-4} = 4.0$ pips; forward spread $= (1.08690 - 1.08635)/10^{-4} = 5.5$ pips. The hedger pays 1.5 pips more to transact forward than spot, because a forward is two prices — a spot rate and a points quote — each carrying its own bid-ask, and the dealer must be compensated separately for the exchange-rate risk of the spot leg and the interest-rate and balance-sheet risk of the financing leg. The wider the tenor and the more constrained the dealer's balance sheet, the wider the points spread; this is the retail-visible face of the same friction that produces the cross-currency basis of §5.1.
**(c)** Mid spot $= 1.08380$; mid points $= 28.25$; mid forward $= 1.086625$. The period premium is

$$\frac{1.086625}{1.08380} - 1 = 0.00260657 = 0.260657\%,$$

and annualising on ACT/360 over 91 days,

$$0.00260657 \times \frac{360}{91} = 0.0103117 = 103.12 \text{ bp}.$$

The USD yields roughly 103 basis points more than the EUR at three months — approximately, since this is a simple annualisation of a ratio and ignores the day-count asymmetry the two currencies would actually apply, both of which are second-order at this tenor but not at a year.
**(d)** A descending quote means subtract:

$$\text{bid} = 1.0836 - 0.00290 = 1.08070, \qquad \text{ask} = 1.0840 - 0.00275 = 1.08125,$$

a spread of 5.5 pips, positive and therefore uncrossed. Note the mechanism: on subtraction the *larger* number is applied to the bid and the smaller to the ask, which is precisely the assignment that preserves bid $<$ ask. That, and not any deeper convention, is why the ordering carries the sign; a dealer who added a descending quote would produce a crossed forward market and be arbitraged instantly.
**(e)** $1{,}000{,}000 \times 28.25 \times 10^{-4} = \$2{,}825.00$ USD over the 91 days. For a long-EUR hedger this is a benefit (the EUR is bought back cheaper forward, in effect earning the differential); for a short-EUR hedger it is a cost of the same size. Either way it is the same quantity the CFD trader of Worked Example 5.2 meets as a nightly financing line. (All figures verified in `verify_mk05.py`, §8.)

**Exercise 5.3 — Carry P&L under three outcomes, and the crash that erases a year**

A hypothetical scenario. Spot NZDJPY is $S_0 = 90.00$ (JPY per NZD). Twelve-month rates are $r_{\text{NZD}} = 4.75\%$ and $r_{\text{JPY}} = 0.50\%$, simple, matched day counts. A trader goes long $N = 200{,}000$ NZD funded in JPY and holds for one year.

(a) State the gross carry and compute the two maturity legs.
(b) Compute the breakeven spot rate, and say what else that number is.
(c) Compute the total P&L in JPY and in NZD, and as a percentage of the NZD notional,

for $S_1 \in \{91.50,\ 90.00,\ 79.00\}$, decomposing each into a spot leg and an interest leg.

(d) Express the crash outcome in years of carry, two ways: on the spot leg alone, and net.

(e) The trader runs the position at a self-imposed 20:1 gearing. What does the crash do to the margin behind it?

▷ working space

**Solution — Exercise 5.3**

**(a)** Gross carry $= 4.75\% - 0.50\% = 4.25\%$ p.a. Borrowing $200{,}000 \times 90.00 = 18{,}000{,}000$ JPY and converting:

$$\text{NZD leg} = 200{,}000 \times 1.0475 = 209{,}500.00,$$
$$\text{JPY debt} = 18{,}000{,}000 \times 1.0050 = 18{,}090{,}000.00.$$

**(b)**

$$S^* = 90.00 \times \frac{1.0050}{1.0475} = 86.348449,$$

a depreciation of 4.0573%. It is also, exactly, the one-year forward rate implied by Result 5.1 on these rates, and therefore — under UIP — the rate at which the market is supposed to *expect* NZDJPY to be trading in a year. The carry trade is the wager that this expectation is systematically too pessimistic about the high-yielder.

**(c)** With total $= 209{,}500\,S_1 - 18{,}090{,}000$, spot leg $= 200{,}000(S_1 - 90)$, and interest leg $= 9{,}500\,S_1 - 90{,}000$:

| $S_1$ | Spot move | Spot leg (JPY) | Interest leg (JPY) | Total (JPY) | In NZD (%) |
|---|---|---|---|---|---|
| 91.50 | +1.6667% | +300,000 | +779,250 | +1,079,250 | +11,795.08 (+5.8975%) |
| 90.00 | 0.0000% | 0 | +765,000 | +765,000 | +8,500.00 (+4.2500%) |
| 79.00 | −12.2222% | −2,200,000 | +660,500 | −1,539,500 | −19,487.34 (−9.7437%) |

The middle row returning exactly 4.25% is the check that the construction is correct. Note again that the interest leg shrinks in the crash — 660,500 against 765,000 — because the NZD coupon is earned in NZD and converted at the fallen rate.

**(d)** On the spot leg alone: $12.2222/4.25 = 2.8758$ years of carry destroyed. Net of the interest earned: $9.7437/4.25 = 2.2926$ years. Either way, a single adverse year undoes between two and three good ones — and the asymmetry is structural, since the carry accrues at a bounded rate while the spot loss is unbounded.

**(e)** At 20:1 the margin is $200{,}000/20 = 10{,}000$ NZD. The crash loss of 19,487.34 NZD is 1.9487 times that margin: the account is closed out long before the year ends, so the trader never collects the 660,500 JPY of interest the table credits him with, and the realised loss is capped at the margin only because the broker's close-out intervened. The table's arithmetic assumes the position survives to maturity; leverage is precisely the mechanism by which it does not. (All figures verified in `verify_mk05.py`, §9.)

**Exercise 5.4 — Reading a Fama slope into a carry book**

A researcher estimates (5.1) on a long sample of monthly data for a major pair and obtains $\hat{\beta} = -0.70$ with $\hat{\alpha} \approx 0$ and $R^2 \approx 3\%$. The current annualised differential is 3.60% and the pair's annualised volatility is 10%.

(a) What is the carry multiplier, and what expected excess return does the slope imply? Compare with the UIP and random-walk counterfactuals.

(b) What Sharpe ratio does that imply, and what does the $R^2$ tell you about how much confidence to place in it?

(c) At an indicative 1.2-pip half-spread on AUDUSD at 0.6520, how many round trips a year would consume the entire expected edge? What does that imply about how the position should be managed?

(d) What leverage turns this into a 15% annual return, and what is wrong with the ques-

tion?

(e) Name three reasons the estimate should not be used as a forward-looking parameter without further work.

▷ working space

**Solution — Exercise 5.4**

**(a)** The multiplier is $1 - \hat{\beta} = 1.70$, so $\mathbb{E}[x] = 1.70 \times 3.60\% = 6.12\%$ per annum. Under UIP ($\beta = 1$) the multiplier is zero and the expected excess return is 0.00%; under a random walk in spot with respect to the premium ($\beta = 0$) it is one and the expected excess return is exactly the differential, 3.60%. The negative slope says the high-yielder tends to appreciate, so the trade collects the differential *and*, on average, a favourable spot move worth a further 2.52%.
**(b)** Sharpe $= 6.12/10 = 0.612$ before costs and before any allowance for skew. The $R^2$ of 3% implies a correlation of $\sqrt{0.03} = 0.1732$ between the premium and the subsequent spot move: 97% of the variance of the outcome is unexplained. The slope is a statement about a conditional mean estimated from a very noisy relationship, so the 0.612 is the centre of a wide sampling distribution, not a property of the strategy. The Sharpe is also the wrong summary statistic for a payoff with the negative skew of §5.2: two strategies with the same mean and variance are not equally attractive when one of them delivers its losses in a small number of very bad months.
**(c)** A round trip costs $2 \times 1.2 \times 10^{-4}/0.6520 = 0.036810\%$ of notional, so

$$\frac{6.12\%}{0.036810\%} = 166.26 \text{ round trips per year}$$

consume the entire expected edge. Two conclusions follow. First, at a couple of round trips a week the strategy is dead on costs alone, which means a carry exposure must be implemented as a *held* position with infrequent rebalancing, not as a signal traded on every refresh. Second, the arithmetic is a reminder that this is a premium harvested by patience: the edge accrues with time in the position and is destroyed by turnover, the exact opposite of the platform's short-horizon forecasting mode [Course Book §14.2].
**(d)** $15\%/6.12\% = 2.4510$ times. The question is wrong because leverage scales the crash by the identical factor: at 2.4510×, Worked Example 5.2's 7.2494% adverse year becomes 17.77%, and the negative skew means the realised drawdown distribution is worse than that scaling suggests. Levering a positively-skewed premium and levering a negatively-skewed one are different acts, and only the first is well described by "more return for more volatility".
**(e)** Three reasons, in increasing order of seriousness. (i) *Instability*: Fama-slope estimates vary materially by pair, tenor and subsample, and several post-2008 samples give slopes near or above zero — so $-0.70$ is a sample statistic, not a constant. (ii) *Regime-dependence*: Menkhoff et al. (2012) show the payoff loads negatively on global FX volatility innovations, so a single unconditional slope averages across states whose conditional means differ in sign; the model that matters is $\beta$ conditional on the volatility state, and the platform already carries the state variable that conditions it. (iii) *Selection*: a slope chosen after examining several pairs and windows is a maximum over a search, and its in-sample value overstates what will be realised — the deflation for selection [Course Book §15.4] applies here in full, and a naive Sharpe of 0.612 from a searched specification is not a Sharpe of 0.612. (All figures verified in `verify_mk05.py`, §10.)

**Exercise 5.5 — Map the drivers onto the platform's covariate feeds and their typing**

Take each economic driver in this chapter and place it in the platform's feed architecture. Produce a table with one row per driver giving: the observable series that proxies it; its covariate typing in the Temporal Fusion Transformer's sense [Course Book §8.4] — *static*, *known-future*, or *observed-past*; the horizon at which it carries information (Table 5.1); and the specific leakage risk it introduces. Then state, in a sentence each, the two rows where a careless implementation would most likely produce an apparently profitable but false backtest.

▷ working space

**Solution — Exercise 5.5**

| Driver | Proxy series | Typing | Horizon | Leakage risk |
|---|---|---|---|---|
| Policy decisions | Meeting calendar (dates/times) | Known future | Minutes | None: the calendar is published years ahead |
| Scheduled releases | Release calendar (dates/times) | Known future | Minutes | None for dates; the *values* are a different row |
| Liquidity clock | Hour-of-week, session bucket, rollover flag | Known future | Minutes–hours | None; purely deterministic, but must respect daylight-saving shifts |
| Financing | Broker's published financing rate for the coming day | Known future | Days | Using an average or current rate instead of the day's published one |
| Policy expectations | 2y rate differential | Observed past | Days–weeks | Release lag; also substituting policy-rate differences for market-implied ones |
| Realised macro data | CPI, payrolls, trade balance prints | Observed past | Days–weeks | Release lag *and* revision: the vintage you can see today is not what was seen then |
| Positioning | COT net non-commercial positioning | Observed past | Weeks | Publication lag: the data are as-of Tuesday and released later in the week |
| Risk appetite | RORO composite; VolRegime state | Observed past | Days–weeks | Normalisation leakage if the composite is standardised over the full sample |
| Carry premium | Level of the differential; global FX volatility | Observed past | Weeks–months | Regime-averaging: an unconditional loading hides sign-flipping conditional means |
| Long-run anchor | RER gap versus its own mean; terms of trade | Observed past | Months–years | Release lag, heavy revision, and a mean estimated using future data |
| Instrument identity | Pair, pip convention, contract spec | Static | — | Survivorship, if the instrument list is the one available today |

**The two most dangerous rows.** First, *COT positioning*: the report is compiled as of Tuesday's close and published several days later, so a naive join on the as-of date hands the model Tuesday's positioning on Tuesday — a clean, plausible-looking, and entirely fictitious edge of exactly the kind that survives a casual backtest because nothing about the resulting equity curve looks anomalous. Second, *realised macro data*: series such as GDP, the current account and even CPI are revised, sometimes substantially, so a modern database query returns the *revised* value for a historical date, and a model trained on it has been shown numbers that did not exist at the time. Both are release-lag or revision leakage in the Course Book's taxonomy [Course Book §18.2], and both are fixed the same way: store every macro series with an explicit availability timestamp and join on that timestamp, never on the reference date.

**Two structural remarks.** The known-future rows are the ones this chapter argues are most reliably useful — not because they predict direction, which they do not, but because they predict *variance* and *spread* with high reliability and at zero data cost. And the last row is the one most often forgotten: a covariate set built from the 21 instruments OANDA offers today is conditioned on those instruments having survived to today, which is a selection the backtest cannot see.

**Key idea — What to remember**

Covered interest parity is not a theory but an identity: given a spot rate and two interest

rates, the forward rate is arithmetic, enforced by an FX swap, and the persistent post-2008 cross-currency basis measures the price of the balance sheet the arbitrage consumes rather than a flaw in the algebra. Its uncovered cousin is a hypothesis about expectations, it fails systematically in the direction Fama (1984) identified, and the carry trade is the industrial harvesting of that failure — earning $(1 - \beta)$ times the differential in exchange for a negatively skewed payoff that loses precisely in high-volatility states, which is why the platform's RORO and VolRegime series are the correct conditioning variables for any differential-driven exposure rather than optional extras. Purchasing power parity and the balance of payments anchor the rate over years and are close to useless over days; policy expectations dominate the horizons the platform trades, but only through their surprise component, and even a correctly-signed, correctly-sized surprise response is a quarter of one ordinary day's noise. The synthesis is that every driver in this chapter is real and none of them delivers a point forecast at a tradeable horizon: economics explains the variance, the regime and the shape of the tail far better than it explains the mean, which is why the distributional envelope is not a modelling preference but the only object the evidence supports [Course Book §1.3]. What remains, and what the rest of this part supplies, is the machinery for acting on that conclusion: the contract through which the platform actually holds these exposures, the financing and margin arithmetic that turns this chapter's carry into a nightly line item, and the positioning and sentiment feeds of Chapter 7 that make the regime conditioning argued for here into something a model can actually be given.

# 6

# FX CFDs and the Retail Dealing Model

*Chapter 4 mapped the market the reader's quotes come from and Chapter 5 the economics that move them; this chapter turns the glass around and looks back through it from the dealer's side. It is the chapter in which the reader has the most at stake, because every mechanism in it is applied to his own account by his own counterparty, nightly, whether or not he knows the mechanism exists. It follows a retail FX CFD order from the liquidity providers whose streams feed the quote, through the aggregation and markup that turn those streams into the price he sees, into the routing decision that determines whether his trade is hedged or warehoused, and on into the dealing engine's frictions — last look, rejection, asymmetric slippage, inventory skew. It then rebuilds the spread from the dealer's own profit and loss, derives the financing charge from the interest differential that generates it, works a margin call and a stop-out end to end, and states plainly what all of this obliges the platform to do differently.*

## 6.1 The counterparty reality

Start with the single structural fact from which everything else in this chapter follows: a retail FX CFD trade never touches the interbank market. Not indirectly, not through a pipe, not in attenuated form — never. When the reader clicks buy, no order is routed to a venue, no queue is joined, no counterparty is matched. A bilateral contract is created between him and his broker, referencing a price the broker itself published, and that is the entirety of what has happened. Whatever the broker subsequently does in the wider market is the broker's own separate business, conducted on the broker's own account, and is invisible to and legally distinct from the client's contract.

Chapters 1 and 3 established the wrapper between them. A contract for difference is an agreement between two parties to exchange, on closing, the difference between an instrument's opening and closing price multiplied by a contract size, with no delivery of the underlying and no ownership of it at any point — a bilateral OTC contract with the broker whatever it happens to reference, which is what makes an equity CFD an OTC instrument even though the share it tracks trades on an exchange. It is margined rather than funded — the client posts a fraction of the notional as collateral, not the notional itself — so the broker is financing the remainder and charges for doing so. The client's asset, in other words, is not a currency position. It is a *claim on his broker*, denominated in the profit and loss of a reference price the broker also supplies. Two exposures come bundled into one instrument and they are worth separating in the mind: exposure to EURUSD, which is the exposure the client wants, and exposure to the broker's continued solvency and continued willingness to honour the contract, which is the exposure he acquires whether he wants it or not.

**Trap — The instrument is not the market**

"I am long EURUSD" is a description of the reader's economic exposure and a misdescription

of his legal position. He is long a contract with one company, whose value is defined by reference to that company's own published EURUSD quote. If the company stops publishing, or publishes something eccentric, or fails, the euro is entirely unaffected and his position is not. Every statement in this chapter about spreads, fills, financing and stop-outs is a statement about that company's conduct of its own book, not about the foreign exchange market, and the two are related only as closely as the company's business model requires them to be.

### The ancestry of a quote

The price on the client's screen is the end of a short manufacturing chain, and each stage of that chain does something specific to it.

**Definition — Liquidity provider**

A *liquidity provider* (LP) is a wholesale counterparty — a dealing bank or a non-bank electronic market maker — that streams continuous, executable two-way prices to a broker under a bilateral credit or prime-brokerage relationship. Each LP quotes its own bid and ask in its own size, updates them at its own frequency, and reserves its own rights over whether to accept an order it has quoted. A retail broker typically maintains streams from several LPs simultaneously; the streams are not identical and are not intended to be.

**Definition — Aggregation**

*Aggregation* is the construction of a single synthetic book from several LP streams by taking, at each instant, the highest available bid and the lowest available ask across providers (and, for larger sizes, by stacking the streams into depth tiers). Because the best bid and the best ask need not come from the same provider, the aggregated width is weakly narrower than any individual provider's width, and usually strictly narrower. Aggregation creates no liquidity; it selects, continuously, across liquidity that already exists.

**Definition — Markup**

A *markup* is the increment a broker adds to the aggregated price before publishing it to clients: a shift of the bid down and the ask up, in the broker's own units, applied per side. Symmetric markup widens the client's spread while leaving the mid unchanged; asymmetric markup — more on one side than the other — widens the spread *and* displaces the mid, which is the mechanical form the inventory skew of §6.3 takes.

Figure 6.1 is the chain in full. It is worth walking through with numbers, using a hypothetical instant on EURUSD constructed for this chapter (all arithmetic in `verify_mk06.py`, §1). Three LPs stream, respectively, 1.08508/1.08514, 1.08509/1.08516 and 1.08507/1.08513 — widths of 0.6, 0.7 and 0.6 pips. Aggregating, the best bid is LP2's 1.08509 and the best ask is LP3's 1.08513, so the composite is 1.08509/1.08513: a width of 0.4 pips, 0.2 pips tighter than the tightest single provider, and a width no individual LP is actually offering. The broker then applies a markup of, say, 0.5 pips per side, publishing 1.08504/1.08518 — a retail width of 1.4 pips, 3.5 times the composite width, of which the markup is 71.43%. The mid is untouched at 1.085110, because the markup was symmetric. Had the broker instead marked up 0.2 pips on the bid and 0.8 on the ask, the width would still be 1.4 pips but the published mid would sit at 1.085140, three-tenths of a pip above the composite mid: the same spread, a different centre, and a systematically different mark on every open position that references it.

Three consequences deserve stating before anything else in the chapter is built on them. First,

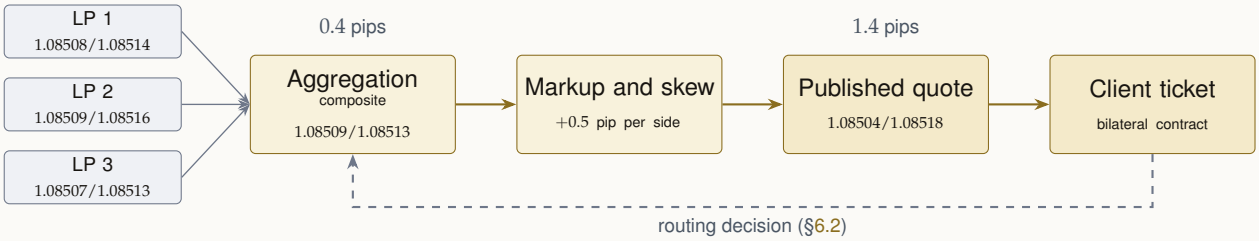



**Figure 6.1.** The ancestry of a retail quote, with the hypothetical EURUSD instant of §6.1. Aggregation narrows the width below any single provider's; markup widens it again, by more; the dashed return path is the routing decision of §6.2, which the client never observes.

the retail width is a *decision*, not a discovery: it is composite width plus markup, and the markup is set by the broker for its own reasons, per instrument, per session, and in some arrangements per client segment. Second, the composite width is itself a decision about how many LPs to connect and on what commercial terms, so "the spread widened" can mean either that the wholesale market thinned or that the broker chose to charge more, and from the client's side the two are indistinguishable. Third, and most consequential for the platform, the published mid is not a neutral object. It is the midpoint of a quote that has already been manufactured, and if the markup is ever asymmetric the mid inherits that asymmetry. A backtest run on stored mids is therefore not running on a wholesale reference price with the spread stripped out; it is running on the centre of one dealer's manufactured quote, which §6.7 treats as a binding constraint on what the platform can honestly claim.

## 6.2 A-book, B-book, and the hybrid

Once the client's ticket exists, the broker faces a decision the client never sees: what to do with the exposure it has just acquired. There are exactly two primitives and one combination.

**Definition — A-book**

In *A-book* handling, the broker offsets the client's trade in the wholesale market — essentially trade for trade, or in netted batches — so that the exposure it took on from the client is passed through to one or more LPs. Its revenue is then the markup of §6.1 plus any explicit commission, less what it pays the LP in spread and in the operational cost of hedging. It carries little directional risk on the client's position; what it does carry is execution risk on the hedge, credit exposure to its LPs, and the cost of posting margin against its own wholesale positions. The client's profit or loss is, to the broker, economically irrelevant.

**Definition — B-book**

In *B-book* handling, the broker *internalises* the trade: it takes the other side and keeps it, warehousing the resulting exposure on its own book rather than passing it out. Its revenue is then the markup *plus the client's loss*, or the markup minus the client's gain. It carries the full directional risk of the position it has warehoused, mitigated by whatever netting it achieves against other clients' opposing positions and by whatever residual it chooses to hedge. Internalisation is the wholesale market's own dominant practice at the dealer tier — Chapter 4 recorded the BIS's finding that dealing banks internalise a large share of client flow rather than hedging it externally (BIS, 2022; Rime and Schrimpf, 2013) — and at the retail tier it is the same operation conducted on smaller tickets.

**Definition — Hybrid routing**

*Hybrid routing* assigns handling per client, per instrument, or per ticket, according to a profile the broker maintains: flow it judges likely to be informed or systematically profitable is routed A-book, and flow it judges likely to be uninformed is warehoused B-book. The profile is built from the account's own history — realised profitability, holding period, order size, latency of entry relative to news, correlation with other clients' flow — and is updated continuously. Essentially every sizeable retail broker operates some version of this; pure-A and pure-B firms are the exceptions rather than the poles of a spectrum most firms sit between.

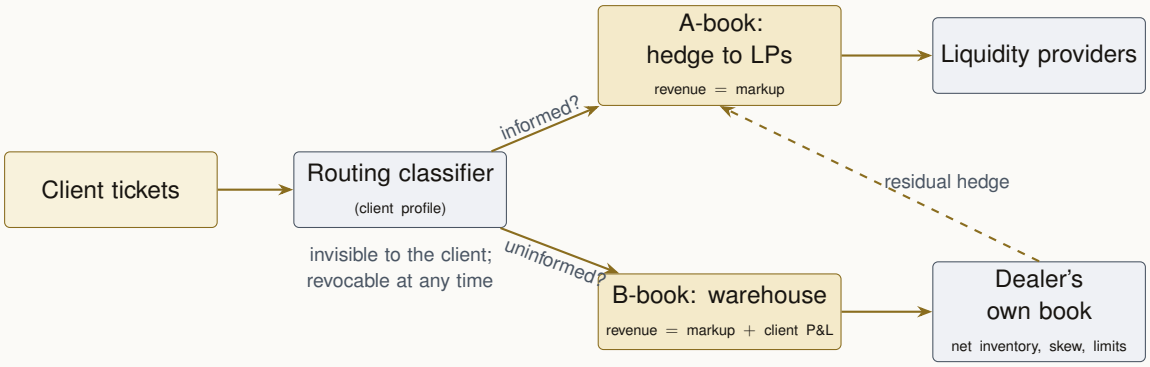



**Figure 6.2.** A-book, B-book and the hybrid. The classifier is internal, its inputs are the client's own order history, and its output is disclosed nowhere on the ticket. The dashed arrow is what makes the two books one business: the warehoused inventory's residual is hedged through the same LP relationships the A-book uses.

### The honest economics

The reason B-booking exists, and the reason it is profitable, is not mysterious and does not require anyone to behave badly. It is the adverse-selection arithmetic of Chapter 2 read in reverse. A market maker's spread must cover the losses it takes to informed counterparties, and the size of that cover is the probability of facing an informed trader times the size of the information at stake — $\mu\Delta$ in the Glosten–Milgrom decomposition (Glosten and Milgrom, 1985). If a population of counterparties has a low $\mu$, then quoting them a spread calibrated to a higher $\mu$ is profitable in expectation, and the profit is exactly the excess of the charged spread over the realised adverse selection. Retail FX flow is, in aggregate and over long samples, low-$\mu$ flow: it is not trading on order-flow information the dealer lacks, it is not front-run by a research process the dealer cannot see, and its systematic biases (Chapter 7 documents the positioning-data version of the same observation) are the kind a counterparty can price. That is what makes the population profitable to face. It is a statement about a population, and it says nothing whatever about any individual account.

Warehousing is also less risky than the phrase "the broker takes the other side of your trade" makes it sound, for a reason that is purely statistical. A broker with thousands of clients holds, at any instant, a large book of offsetting positions: some clients long EURUSD, some short. What it warehouses is not the gross of client positions but the *net*, and the net is a small fraction of the gross precisely because client opinions differ. If gross client exposure in a pair is $G$ and the net is $I$, the dealer's directional risk is governed by $I$ while its spread revenue is governed by $G$, and the ratio $G/I$ is the leverage of the business model. A dealer whose clients disagree with each other earns spread on every ticket and carries risk on almost none of it. This is the same internalisation logic the wholesale tier runs on, and it is why the question "does my broker want me to lose?" is less illuminating than the question "how big is my broker's net position, and in which direction?"

The conflict of interest is real, structural, and worth stating carefully. Under B-book handling the broker's revenue is increasing in the client's loss; under A-book handling it is not. That is a conflict in the plain sense: the two parties' interests are opposed on the contract's outcome, which is not true of a pure agency relationship. The conflict is disclosed in most jurisdictions' required client documentation, it is well known to regulators, and retail CFD provision has attracted sustained regulatory attention over the past decade — leverage caps, negative-balance protection mandates, marketing restrictions and mandatory loss-rate disclosures are all responses to it, and §6.6 returns to the specifics. Documented enforcement actions against individual firms exist in several jurisdictions. None of that supports the stronger claim that a particular broker manipulates a particular client's prices, and this chapter makes no such claim about any firm. The defensible position is the narrow one: the incentive exists, it is structural rather than accidental, it is regulated rather than hidden, and the correct response is to measure one's own fills rather than to trust or accuse.

Two further points make the routing decision harder to reason about than it first appears. It is invisible: nothing on a fill confirmation distinguishes an A-booked ticket from a B-booked one, because the client's contract is identical in both cases and the routing is an internal treasury decision made after the contract exists. And it is revocable: an account's classification can change without notice, in either direction, so a stable pattern of fills over one period is weak evidence about the next. The diagnostics of §6.3 are therefore statements about what the reader's fills have looked like recently, not about a permanent property of his account.

**Trap — "My broker hunts my stops"**

Sort this claim into its plausible and implausible halves, because they have very different answers.

*Structurally plausible.* Spreads widen in thin hours, at the rollover instant and around scheduled releases, for the dealer-side reasons §6.4 derives. A resting stop is triggered by the price on the side that reaches it — a long's stop by the bid, a short's by the ask — so a spread that widens around an unchanged mid moves both trigger levels outward and can fire a stop the mid never reached. Nothing about that requires targeting: it is what a wider quote mechanically does to every resting order in the book, and it happens at exactly the times a systematic strategy is most likely to have orders resting. Stops clustered at round numbers and obvious technical levels also sit where genuine liquidity effects concentrate, so they get hit disproportionately without anyone aiming at them.

*Implausible at scale.* Per-client price manipulation — a feed bent for one account to reach one stop — is a different claim entirely. It requires a per-client quote path, leaves an audit trail across the firm's own systems and its clients' statements, is a straightforward regulatory offence in every jurisdiction the reader would trade from, and buys the firm a few pips on one ticket against an existential enforcement risk. The economics do not work; a wider spread applied to everyone earns more, legally.

*The sober version.* Both the plausible mechanism and the implausible one predict "my stop got hit and price came back". They differ in what else they predict, and the difference is measurable from the reader's own data: the plausible mechanism predicts that stop-outs cluster in the same session buckets and news windows in which the platform's own measured spread curves widen, for *every* instrument and independent of which side he was on. Targeting predicts clustering conditional on his position. He already has the cost surface needed to tell them apart, and it is the only instrument that can.

## 6.3 The dealing engine

Between the published quote and the filled ticket sits a piece of software with rules of its own. Those rules are where the difference between a backtest's assumed fill and an account's actual fill is manufactured, and they are worth naming individually, because each leaves a different statistical fingerprint in the reader's own fill history.

### Last look

> **Definition — Last look**
>
> *Last look* is a liquidity provider's contractual right, having streamed an executable price and received an order against it, to pause for a short interval, re-observe the market, and then either accept or reject the order. The pause is typically measured in milliseconds. Two distinct justifications are advanced for it: a *credit and validity* check, verifying that the counterparty and the ticket are good, and a *price* check, verifying that the market has not moved between the quote and the order's arrival by more than a stated tolerance. The mechanism belongs to the wholesale layer, exercised by LPs against the broker; the retail client's contract with the broker is separate, and what reaches him is the broker's own decision about how to handle a wholesale rejection.

Last look has been the subject of sustained industry and regulatory scrutiny, and the concern is easy to state precisely: an option to reject after observing the subsequent market is, formally, a free option written by the order sender to the quoter. If the tolerance is applied symmetrically — rejecting equally when the market has moved in the LP's favour and against it — the option has little value and the check is doing the risk-control work it claims. If it is applied asymmetrically — accepting when the market has moved in the LP's favour, rejecting when it has not — the quoted price is not the price at which the sender can actually transact, and the difference is a transfer from sender to quoter that never appears in any spread statistic. The same logic applies to the hold window itself: a longer pause is worth more to the quoter and less to the sender, for reasons unrelated to credit checking. Industry codes of conduct through the 2010s and 2020s addressed exactly this, moving towards symmetric application and disclosed hold windows; the practice remains in use, remains disclosed at the wholesale tier, and remains contested. This book takes no position on whether any firm applies it improperly, and states the structure rather than an accusation — but the structure alone establishes the point the platform needs: *a streamed price is a conditional offer, and the condition is not under the sender's control*. This is the microstructural reason Chapter 4 insisted that no quote in this market is "the" price; here the same fact reappears as a property of executability rather than of consensus.

### Requotes and rejections

What the broker does when its own hedge is rejected, or when its risk system declines a ticket, reaches the client in one of three forms. A *requote* returns a new price and asks the client to accept or decline it, converting a fill into a decision under a clock. A *rejection* declines the order outright, leaving the client unfilled and — crucially — still exposed to whatever he was trying to trade into or out of. A *slipped fill* accepts the order at a price different from the one requested. The three are not interchangeable: the first two leave a strategy unhedged for a further round trip of latency, a risk for which a backtest that always fills has no representation at all.

The rate at which these occur is not constant. It rises in exactly the conditions in which wholesale liquidity thins and quote lifetimes shorten — news, rollover, the thin Asian window, stress — because that is when the right to decline is worth most to whoever holds it. This matters more than its unconditional average suggests, because the conditional rate is what a strategy actually experiences: a system whose signals fire preferentially into fast markets faces the

conditional rejection rate and never the unconditional one. The point generalises the selection problem §6.4 makes quantitative for spreads.

### Asymmetric slippage as a diagnostic

Slippage symmetric in sign is a nuisance. Slippage systematically adverse is a cost, and unlike the spread it is invisible on the quote. The distinction is testable from fill data alone, and the test is simple enough to run on a few hundred tickets.

Take a hypothetical sample of 400 market orders constructed for this chapter (arithmetic in `verify_mk06.py`, §2). Suppose 210 filled worse than requested, by an average of 0.42 pips; 60 filled better than requested, by an average of 0.11 pips; and 130 filled exactly at the requested price. The mean signed slippage is

$$\frac{210 \times (-0.42) + 60 \times (+0.11) + 130 \times 0}{400} = \frac{-81.6}{400} = -0.2040 \text{ pips per fill},$$

which on a 1.4-pip quoted width is a 14.57% uplift on the advertised cost of trading and raises the effective one-way cost from the 0.70 pips a half-spread model assumes to 0.9040 pips. But the mean is the smaller half of the finding. Of the 270 orders that slipped at all, 210 slipped adversely; under a null of symmetric slippage the expected count is 135 with standard deviation $\sqrt{270 \times 0.25} = 8.2158$, so the observed count sits

$$z = \frac{210 - 135}{8.2158} = 9.1287$$

standard deviations from that null. A number like that is not a sampling accident, and interpreting it requires no hypothesis about intent: an execution channel that fills you when the market has moved your way and slips you when it has not is, whatever its internal mechanics, collecting an option premium from you. Note also what the test does *not* need — no knowledge of the LP panel, no hold-window disclosure, no routing information. It needs the requested price and the filled price, both of which the client already has.

> **Key idea — The diagnostic is the deliverable**
>
> The reader cannot observe his broker's routing decision, its LP panel, its hold windows or its markup schedule. He can observe every fill he has ever received. Three statistics computed from that record — mean signed slippage, the adverse-to-favourable count ratio among slipped fills, and the rejection rate conditional on session bucket and news proximity — jointly characterise the execution channel he actually faces, without a single assumption about what happens inside it. Everything in this section exists to explain what those three numbers are measuring, and Exercise 6.4 asks what each of them would have to look like to change his mind about which dealing model he is facing.

### Inventory limits and quote skew

A dealer warehousing client flow accumulates a net position and does not want an unbounded one. The classical response, from the inventory strand of microstructure theory (O'Hara, 1995; Amihud and Mendelson, 1986), is not to hedge every ticket — hedging costs the LP's spread — but to *shade the quote* in the direction that recruits the offsetting flow. A dealer long the base currency lowers both bid and ask: the lower ask makes buying cheaper for clients, and the lower bid makes selling to him less attractive. A dealer short the base currency raises both. The width need not change at all; what moves is the centre. This is the asymmetric markup of §6.1, applied for a reason.

**Definition — Inventory skew**

*Skew* is a signed displacement $\delta$ applied to both sides of a dealer's published quote, so that a quote of width $a - b$ becomes $[\,b + \delta,\ a + \delta\,]$. Its sign is opposite to the sign of the dealer's net inventory $I$ in the base currency: long inventory skews down, short inventory skews up. Its magnitude follows a desk rule — typically increasing in $|I|$, capped, and accompanied by a hard limit beyond which the residual is hedged externally rather than skewed away.

**Result — The dealer's three levers**

Facing an accumulating position $I$, a dealer has exactly three responses, and they trade off against one another. *Skew* the quote to recruit offsetting flow: free of transaction cost, but slow, uncertain, and it lowers expected revenue per ticket on the side being discouraged. *Widen* the quote to slow the arrival of new flow in both directions and raise the compensation per ticket: fast, but it drives price-sensitive clients elsewhere. *Hedge* the residual externally: immediate and certain, but it costs the LP's half-spread on the hedged amount — a realised loss taken to remove an unrealised risk. Real desks use all three, ordered by cost: skew first, widen second, hedge last, with the hedge reserved for inventory beyond a hard limit. Every widening the reader observes in his own feed is one of these three levers being pulled, and §6.4 supplies the arithmetic that decides how far.

**Worked Example 6.1 — A dealer inventory ledger, six client trades**

A hypothetical EURUSD desk begins flat and quotes the 1.4-pip width manufactured in §6.1 — a half-spread of 0.7 pips either side of its own mid. Its desk rule skews the quote by 0.1 pip per 250,000 EUR of net inventory, capped at 0.5 pips, and hedges externally only when $|I| \geq$ 1,000,000 EUR. Six client tickets arrive. Every figure below is computed in `verify_mk06.py`, §3, for this illustrative scenario; none of it is market data.

| # | Client | Size (EUR) | Mid | Dealt | Dealer $I$ | Avg price | Skew |
|---|---|---|---|---|---|---|---|
| 1 | buys | 200,000 | 1.08505 | 1.08512 | −200,000 | 1.085120 | 0.08 |
| 2 | buys | 300,000 | 1.08517 | 1.08524 | −500,000 | 1.085192 | 0.20 |
| 3 | sells | 150,000 | 1.08515 | 1.08508 | −350,000 | 1.085192 | 0.14 |
| 4 | buys | 400,000 | 1.08544 | 1.08551 | −750,000 | 1.085362 | 0.30 |
| 5 | buys | 250,000 | 1.08556 | 1.08563 | −1,000,000 | 1.085429 | 0.40 |
| 6 | sells | 600,000 | 1.08554 | 1.08547 | −400,000 | 1.085429 | 0.16 |

**Reading the ledger.** A client buy is a dealer sale, so the dealer's inventory $I$ is the negative of the clients' net. The dealt price is the mid plus 0.7 pips on a client buy and minus 0.7 pips on a client sell. The average-price column carries the dealer's inventory at weighted average cost: a ticket that increases $|I|$ updates the average, a ticket that reduces $|I|$ leaves the average alone and realises the difference against it. Skew is $0.1 \times |I|/250{,}000$ pips, capped at 0.5; at ticket 5 the position reaches the 1,000,000 EUR hard limit and the hedging rule of Result 6.3 is triggered.

**Total profit and loss.** Summing cash flows (the dealer receives on sales, pays on purchases) gives \$434,163.50 against a closing position of −400,000 EUR, marked at a final mid of 1.08570 for −\$434,280.00:

$$\text{P\&L} = 434{,}163.50 - 434{,}280.00 = -\$116.50.$$

The desk lost money on the batch.

**Why, exactly.** The result decomposes exactly into two terms. Spread capture is the half-spread earned on every ticket,

$$1{,}900{,}000 \times 0.00007 = \$133.00,$$

a full 0.700 pips per unit traded with not one pip given away. Inventory carry is the position held across each move in the mid, leg by leg:

| After # | Position | Mid from | Mid to | Leg P&L |
|---|---|---|---|---|
| 1 | −200,000 | 1.08505 | 1.08517 | −$24.00 |
| 2 | −500,000 | 1.08517 | 1.08515 | +$10.00 |
| 3 | −350,000 | 1.08515 | 1.08544 | −$101.50 |
| 4 | −750,000 | 1.08544 | 1.08556 | −$90.00 |
| 5 | −1,000,000 | 1.08556 | 1.08554 | +$20.00 |
| 6 | −400,000 | 1.08554 | 1.08570 | −$64.00 |
| | | | Total | −$249.50 |

and $133.00 - 249.50 = -116.50$ reproduces the cash-flow answer to the cent. Per unit traded, the desk earned 0.700 pips of spread and lost 1.313 pips of inventory, netting −0.6132 pips. **What the scenario is showing.** Clients were net buyers of 400,000 EUR into a mid that rose 6.5 pips across the window. The desk was therefore short into a rally, and its short grew precisely as the rally proceeded — which is what warehousing directional client flow means when the flow happens to be right. Every pip of the advertised spread was collected and the desk still lost, on 1.9 million of turnover, exactly the spread it would have earned on a further 832,143 EUR of client round trips at the same 1.4-pip width. This is the concrete form of the point §6.2 made abstractly: B-booking is profitable in expectation over a low-$\mu$ population and is not a riskless business on any particular day, in any particular instrument, against any particular batch of tickets. The skew ladder in the final column is the desk trying to prevent exactly this outcome, and on this sample it did not recruit enough offsetting flow in time.

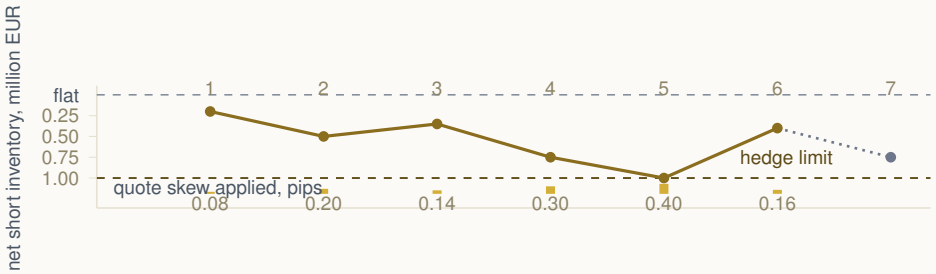



**Figure 6.3.** The inventory path of Worked Example 6.1 (gold, plotted downward because the desk is short throughout) with the skew it applied beneath. The hedge limit binds at ticket 5; the dotted continuation is the first ticket of Exercise 6.1. Hypothetical scenario, computed in `verify_mk06.py`, §3.

## 6.4 Where the spread comes from

Worked Example 6.1 produced a decomposition without naming it. Generalise it. Let $q_i$ be the size of ticket $i$, $h_i$ the half-spread charged on it, $m_i$ the mid at which it dealt, and $I_i$ the dealer's net inventory after it. The dealer's profit on a batch of client flow is then, exactly and as an accounting identity,

$$\Pi \;=\; \underbrace{\textstyle\sum_i |q_i|\, h_i}_{\text{spread capture}} \;+\; \underbrace{\textstyle\sum_i I_i \,(m_{i+1} - m_i)}_{\text{inventory carry}},$$

which is precisely what the ledger computed. Adverse selection is not a third term; it is the *conditional mean* of the second. Writing $a_i \geq 0$ for the expected loss per unit to whatever information ticket $i$ carried, and $\varepsilon_i$ for the residual,

$$I_i\big(m_{i+1} - m_i\big) \;=\; -\,|q_i|\,a_i \;+\; \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid \mathcal{F}_i] = 0,$$

so that $\mathbb{E}[\Pi] = \sum_i |q_i|\,(h_i - a_i)$: the dealer profits exactly to the extent that the half-spread he charges exceeds the information he is charged for. Splitting the realised inventory carry into these two pieces is a modelling act, not an accounting one — the ledger sees only their sum — but the split is the whole of the economics, because the pieces behave completely differently. The residual $\varepsilon_i$ is mean-zero noise the dealer can diversify across clients, instruments and hours until it is negligible; the drift $-|q_i|a_i$ is a one-directional drain that no amount of netting removes, because it is present in every ticket in the same direction.

Chapter 2 derived the adverse-selection term from first principles: facing a population in which a fraction $\mu$ of arriving orders is informed about a quantity worth $\Delta$, a competitive market maker's breakeven half-spread contains $\frac{1}{2}\mu\Delta$, so the quoted width is $\mu\Delta$ plus twice whatever order-processing cost $c$ the dealer bears per side (Glosten and Milgrom, 1985). Chapter 4 then placed the retail broker in the dealer chain that population belongs to. Nothing in either result is retail-specific — which is exactly why it applies here. What is retail-specific is the parameter values, and the interesting fact about them is that $\mu$ and $\Delta$ are not constants. They are functions of the clock.

**Result — The dealer's breakeven width, decomposed**

Writing $\iota$ for the inventory-risk premium the dealer requires per side — compensation for holding an unhedged position over the expected time to offset it — the breakeven quoted width in a given market state is

$$s = \underbrace{\mu\Delta}_{\text{adverse selection}} + \underbrace{2c}_{\text{order processing}} + \underbrace{2\iota}_{\text{inventory risk}}.$$

All three components are state-dependent, and in the states that matter they move *together and in the same direction*, which is why observed spread distributions are so much more right-skewed than any single component would suggest. Anything the dealer quotes below $s$ loses money in expectation against that state's flow; anything above it is competitive rent, bounded by what rival brokers quote.

### Why spreads widen, from the dealer's side

The reader's own feed widens at three recurring times, and Chapter 4's liquidity clock said *when*. Result 6.4 says *why*, and the answers are different in each case.

**Scheduled news.** A release is a scheduled increase in $\Delta$: the quantity about to become public is worth a known amount and will be revealed at a known instant. It is also an increase in $\mu$, because the population of traders active in the seconds around a release is selected for wanting to trade on precisely that information — the uninformed have no reason to be there. Both factors in the product rise at once, so $\mu\Delta$ rises multiplicatively rather than additively. Simultaneously, the dealer's LPs face the same problem and widen their own streams, so the composite width of §6.1 widens underneath the markup. A retail spread that goes from 1.4 to 8 pips across a release is not a broker seizing an opportunity; it is three effects compounding, only one of which the broker controls.

**The rollover instant.** At 5pm New York, LPs are rolling their own value dates, desks are changing shift, and the wholesale market's depth genuinely thins for an interval measured in minutes. The dealer's $\iota$ rises because the expected time to offset an unwanted position rises — fewer counterparties are quoting — and the number of LP streams contributing to the composite falls, so the aggregation gain of §6.1 shrinks at the same time. This is a supply-side thinning, not an information event, and it has the distinctive property of being perfectly predictable: it happens at the same instant every trading day.

**Thin Asian hours.** Here the mechanism is inventory risk per unit time. Arrival intensity is low, so a position taken on at 02:00 UTC in a European cross may wait a long time for offsetting

flow; $\iota$ scales with that expected waiting time. Adverse selection is not obviously higher, and $\Delta$ may in fact be lower — but a dealer required to hold an unhedged position five times as long requires roughly five times the compensation for doing so, and the width reflects that even in the absence of any informational threat.

There is a fourth case worth separating because it is the one that damages accounts: **stress**. In genuine dislocation all three components blow out simultaneously, LP streams are withdrawn rather than merely widened, and the dealer's own funding constraints bind (the funding-liquidity channel of Brunnermeier and Pedersen, 2009: a dealer whose margin requirements rise cannot warehouse the inventory it would normally warehouse, so it must widen or stop quoting). Chapter 15's case studies are the extreme end of exactly this mechanism, and §6.6 is what it does to a levered client.

**Worked Example 6.2 — A two-regime spread mixture, and what the width must cover**

Model the state of a hypothetical EURUSD market as a two-point mixture: a *calm* regime occupying a fraction $1-p$ of the clock and a *stressed* regime occupying $p$. Assign Result 6.4's parameters per regime, with an order-processing cost of $c = 0.10$ pips per side common to both. All arithmetic in `verify_mk06.py`, §4; the parameter values are chosen to illustrate the mechanism, not measured from any feed.

| Regime | Weight | $\mu$ | $\Delta$ (pips) | $\mu\Delta$ | $2c + 2\iota$ | Breakeven $s$ |
|---|---|---|---|---|---|---|
| Calm | 0.93 | 0.06 | 3.0 | 0.18 | $0.20 + 0.30$ | 0.68 pips |
| Stressed | 0.07 | 0.40 | 15.0 | 6.00 | $0.20 + 2.40$ | 8.60 pips |

**(a) One width for both states.** A dealer constrained to quote a single width all day must cover the time-weighted mean:

$$\bar{s} = 0.93 \times 0.68 + 0.07 \times 8.60 = 0.6324 + 0.6020 = 1.2344 \text{ pips}.$$

Two things about that number. It is 1.8153 times the calm-regime breakeven, so a client trading only in calm conditions is paying an 81.5% surcharge that is pure insurance against a state he never trades in. And $0.6020/1.2344 = 48.77\%$ of the dealer's entire expected cost is generated by the 7% of the clock spent in stress. That asymmetry is the reason state-contingent quoting exists: it is better for the dealer *and* better for the calm-hours client, and the only party worse off is the client who trades in the stressed state, who was previously being subsidised.

**(b) The dispersion.** $\mathbb{E}[s] = 1.2344$ and $\mathbb{E}[s^2] = 5.607232$, so

$$\mathrm{Var}(s) = 5.607232 - 1.2344^2 = 4.083489, \qquad \mathrm{sd}(s) = 2.0208 \text{ pips},$$

a coefficient of variation of 1.6370: the standard deviation of the cost exceeds its mean. No single number summarises a distribution shaped like this, which is precisely why a scalar "average spread" assumption in a backtest is not a mild approximation.

**(c) Flow is not uniform in time.** Suppose flow arrives 2.5 times faster in the stressed state than in the calm one. The dealer's costs are incurred per unit *traded*, not per unit of clock, so the relevant weights are volume weights:

$$w_{\text{calm}} = \frac{0.93 \times 1.0}{0.93 \times 1.0 + 0.07 \times 2.5} = \frac{0.9300}{1.1050} = 0.841629, \qquad w_{\text{stress}} = 0.158371,$$

giving a volume-weighted breakeven width of

$$0.841629 \times 0.68 + 0.158371 \times 8.60 = 1.9343 \text{ pips},$$

which is 1.567 times the time-weighted 1.2344. A dealer who set his single width from the clock rather than from his own tickets would underprice his flow by more than fifty per cent, and 70.41% of his realised costs would arrive in a state occupying 7% of the day.

**(d) The same error, from the client's side.** Now read (c) backwards. A strategy whose signals fire disproportionately into fast markets — a breakout system, a news-reactive system, anything triggered by a volatility expansion — places, say, 25% of its trades in the stressed state rather than the 7% a uniform trader would. Its realised average width is

$$0.75 \times 0.68 + 0.25 \times 8.60 = 2.6600 \text{ pips},$$

which is 2.1549 times the unconditional mean, an understatement of 1.4256 pips *per trade* for anyone who costed the strategy at the average. On a strategy with an edge of two or three pips per trade, that single modelling choice is the difference between a positive and a negative expectancy [Course Book §14.1], and it is not a small error made in the conservative direction — it is a large error that always runs optimistic.

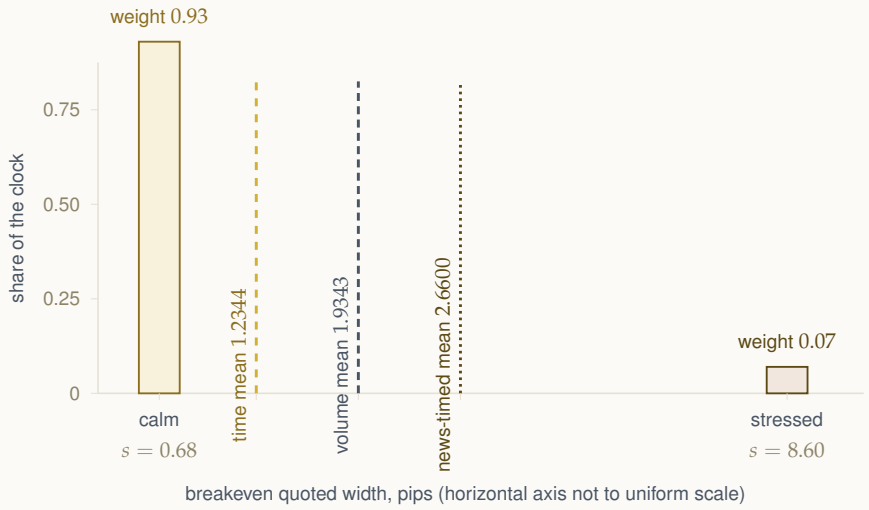



**Figure 6.4.** The two-regime mixture of Worked Example 6.2. Bar heights are the regimes' shares of the clock; horizontal position is the breakeven width each requires. The three vertical lines are three different "average spreads", all correct answers to different questions, differing by a factor of 2.15 from first to last. Hypothetical parameters, computed in `verify_mk06.py`, §4.

**Your project — This is the microfoundation of the cost surface**

Chapter 3 argued that the platform's replacement of mid-only candles with a measured bid/ask cost surface was forced by the empirical fact that a retail spread is a process, not a parameter. This section supplies the mechanism underneath that fact, and the mechanism dictates the surface's shape more tightly than the empirical argument alone did.

Result 6.4 says the width is a function of $(\mu, \Delta, \iota)$, and the three recurring widenings above say those parameters are driven by session, news proximity and volatility state — which is precisely the conditioning set the surface already uses. That is a non-trivial confirmation: the covariates were chosen empirically and turn out to be the dealer's own state variables.

Worked Example 6.2(b) says the per-cell object must be a *distribution*, because the coefficient of variation exceeds one and a mean is not a sufficient statistic for a right-skewed cost. Part (c) says the estimator must be *volume-weighted within the cell* — weighted by the platform's own trade times, not by clock time — or it inherits exactly the dealer's own 1.567× underpricing

error in mirror image. And part (d) prices the consequence of getting this wrong for a signal family the platform actually runs. The per-asset-class differences the Course Book's master table records [Course Book §22.4] sit on top of this same structure: what changes across v1 to v4 is which state variables drive $\mu$, $\Delta$ and $\iota$, not that they do.

**Trap — The average spread is the wrong average**

There are at least three defensible "average spreads" for a single instrument on a single day and they differ by more than a factor of two in Worked Example 6.2: the time-weighted mean (1.2344 pips), the volume-weighted mean over all market flow (1.9343), and the mean weighted by *your own strategy's* trade times (2.6600). Only the third is the number that belongs in your expectancy calculation, and it is the only one that cannot be computed from a spread series alone — it requires the strategy's own timestamps. Any cost figure quoted without saying which weighting produced it is uninterpretable, and the reporting convention that suppresses the question is the one that flatters the result.

## 6.5 Financing: swap points and the rollover seam

The spread is charged twice per round trip regardless of how long the position lives. Financing is charged every night the position survives, and it therefore overtakes the spread on any horizon longer than a few days. Chapter 3 established the retail formula and the back matter restates it; this section derives where the number comes from, which is the only way to know when a broker's published figure is doing something other than passing through the interest differential.

### Swap points from the interest differential

**Definition — Swap points**

*Swap points* are the price difference between the near and far legs of an FX swap, quoted in units of the pair's own price — conventionally in pips or in pipettes. For an overnight roll they are the one-day forward points: the amount by which the exchange rate for tomorrow's value date differs from today's. A retail broker's published "swap long" and "swap short" are these points with the broker's own financing markup applied, and they are published per standard lot per night in units that vary by broker (points of price, or a cash amount in the account currency). The units are provider-specific and must be checked; the mechanism below is not.

Chapter 5 derived covered interest parity in full, so take its one-period statement as given: with $S$ the spot rate in units of quote currency per unit of base, $r_b$ and $r_q$ the simple annualised base- and quote-currency interest rates, and $\tau$ the fraction of a year to the forward value date,

$$F = S\,\frac{1 + r_q\tau}{1 + r_b\tau} \qquad \Longrightarrow \qquad \boxed{F - S \;=\; S\,\frac{(r_q - r_b)\,\tau}{1 + r_b\tau}}$$

— the swap points are the interest differential expressed in price units, and nothing else. The sign follows immediately: when the quote currency yields more than the base ($r_q > r_b$), $F > S$ and the base currency trades at a forward premium; when the base yields more, the points are negative. To carry a position past its value date, the holder rolls it with a swap — selling the near leg and buying the far leg — so a *long* in the base pays $F - S$ per unit per period and a short receives it. Reading the boxed identity backwards: a long in the base earns $r_b$ and funds at $r_q$, so its carry is $S(r_b - r_q)\tau$ per unit, which is $-(F - S)$ up to the discount factor $1/(1 + r_b\tau)$. At one day that factor differs from unity in the fifth decimal place, and the discrepancy it induces — $1.2 \times 10^{-5}$

pips in the scenario below — is immaterial against any realistic markup. This is why the retail formula of Chapter 3 is stated undiscounted without apology.

**Result — Retail financing, restated from Chapter 3**

For $N$ units of base currency at spot $S$, annualised rates $r_b, r_q$, an annualised broker markup $\alpha > 0$ and a day-count basis $D$ (conventionally 365 or 360 by currency), the financing accrued per value-date day is

$$f_{\text{long}} = \frac{NS\big[(r_b - r_q) - \alpha\big]}{D}, \qquad f_{\text{short}} = \frac{NS\big[(r_q - r_b) - \alpha\big]}{D},$$

positive being a credit to the client. Two structural consequences follow directly. The markup enters with the *same* sign on both sides, so $f_{\text{long}} + f_{\text{short}} = -2\alpha NS/D$: holding both sides simultaneously is a guaranteed daily loss of twice the markup, which is what makes the financing book a spread business rather than a directional one. And a long earns a credit only when $r_b - r_q > \alpha$: the markup is a deadband around zero differential inside which *both* directions pay, typically covering the whole of a currency pair whose two policy rates are close.

**Worked Example 6.3 — Swap points and a week of financing, both directions**

A hypothetical AUDUSD scenario, with the reader's account denominated in AUD. Spot $S = 0.6510$, one standard lot $N = 100{,}000$ AUD, $D = 365$, and a broker markup of $\alpha = 1.00\%$ annualised. Rate levels are indicative of a mid-2020s environment and are used here only as arithmetic inputs; verify current specifications before pricing a live position. All figures computed in `verify_mk06.py`, §5.

**(a) The raw swap points.** Take $r_b = 4.35\%$ on AUD and $r_q = 4.90\%$ on USD, a differential of 0.55% against the long. For $\tau = 1/365$,

$$F - S = 0.6510 \times \frac{(0.0490 - 0.0435)/365}{1 + 0.0435/365} = 9.80843 \times 10^{-6} = 0.09808 \text{ pips per day.}$$

The undiscounted version, $S(r_q - r_b)\tau$, gives 0.09810 pips — a difference of $1.17 \times 10^{-5}$ pips, as promised. The notional in quote currency is $NS = \$65{,}100.00$, so the raw daily carry against a long is \$0.98096 per lot: a hundredth of a pip of price, on a position worth sixty-five thousand dollars.

**(b) With the markup.** Applying Result 6.5,

$$f_{\text{long}} = \frac{65{,}100\,\big[(-0.0055) - 0.0100\big]}{365} = -\$2.76452,$$

$$f_{\text{short}} = \frac{65{,}100\,\big[(+0.0055) - 0.0100\big]}{365} = -\$0.80260$$

per lot per day, or $-0.27645$ and $-0.08026$ pips respectively. Both sides pay, because the 1.00% markup exceeds the 0.55% differential; their sum, $-\$3.56712$, is exactly $-2\alpha NS/D$, as the result requires. Note the ratio: the long pays 3.44 times what the short pays, on the same instrument at the same instant, and the entire asymmetry is the differential.

**(c) A week.** A position opened Monday morning New York time and closed Friday afternoon before the 5pm cut straddles four rollovers: Monday, Tuesday, Wednesday and Thursday. Chapter 4's value-date arithmetic makes the Wednesday roll a triple, so the days financed are $1 + 1 + 3 + 1 = 6$. The week costs

$$\text{long:} \quad 6 \times (-2.76452) = -\$16.5871 = -\text{A}\$25.4795,$$

$$\text{short:} \quad 6 \times (-0.80260) = -\$4.8156 = -\text{A}\$7.3973,$$

converting to the AUD account at the same 0.6510. Against an indicative 1.4-pip round-trip spread — \$14.00, or A\$21.5054 — the long's week of financing is 1.1848 times the entire round-trip transaction cost. The crossover is

$$\frac{14.00}{2.76452} = 5.0642 \text{ financed days for the long,} \qquad \frac{14.00}{0.80260} = 17.4432 \text{ for the short.}$$

Past roughly five financed days — which, with the Wednesday triple, is barely three trading days of the week — financing is the larger of the two costs for a long, and it keeps growing while the spread does not. The annualised drag is $\$2.76452 \times 365/65{,}100 = 1.5500\%$ of notional, which is the differential plus the markup exactly, as it must be.

**(d) The other direction.** Now suppose instead $r_b = 5.60\%$ on AUD against $r_q = 4.10\%$ on USD, a 1.50% differential in the long's favour. The one-day points invert to $-0.26749$ pips (the base now at a forward discount), and

$$f_{\text{long}} = \frac{65{,}100\,\big[(+0.0150) - 0.0100\big]}{365} = +\$0.89178,$$
$$f_{\text{short}} = \frac{65{,}100\,\big[(-0.0150) - 0.0100\big]}{365} = -\$4.45890.$$

The long now receives A\$8.2192 over the six financed days and the short pays A\$41.0959 — a ratio of 5.0 to 1 in magnitude, against 3.44 to 1 in the opposite configuration. The lesson generalises: the markup makes the paying side always pay more than the receiving side receives, the asymmetry widens as the differential widens, and a carry position that looks attractive on the raw differential can lose most of its yield to $\alpha$ before it earns anything. On these numbers the long's gross carry of 1.50% is reduced to a net 0.50% — two thirds of the carry consumed by the markup alone, before any exchange-rate movement at all [Course Book §14.2].

### The rollover seam

Financing is not accrued continuously. It is applied in a discrete event at 5pm New York, and the discreteness has consequences the arithmetic above does not show.

Chapter 4 established the mechanics and the back matter tabulates them: the 5pm New York cut is the wholesale market's own value-date roll, spot's value date is T+2 business days, a value date cannot fall on a weekend, and therefore the Wednesday roll advances the value date from Friday to Monday and is charged triple. The weekly pattern $1, 1, 3, 1, 1$ sums to 7: the triple redistributes a week's financing across five events, it does not add to it. Two operational facts follow that matter to a systematic strategy. First, financing is charged to whatever is *open at the instant*, not to whatever traded that day: a position opened at 4:55pm New York and closed at 5:05pm pays a full day (or three, on a Wednesday) for ten minutes of exposure, while a position held from 9am to 4pm pays nothing at all. Second, the reader's own clock disagrees with the event's: 5pm New York is 22:00 UTC in winter and 21:00 UTC in summer, which on his AEST clock is 8am or 7am the following morning — so "Wednesday's triple swap" fires on his Thursday, and a calendar built in Sydney local time will place it on the wrong day.

The seam also has a price dimension, which is the part a cost model must carry. The rollover instant is one of §6.4's three recurring widenings, for supply-side reasons: LPs are rolling their own books, streams thin, the aggregation gain shrinks, and $\iota$ rises because the expected time to offset an unwanted position lengthens. So the same instant that applies the financing charge also applies the day's widest spread, and a strategy that trades across it pays both. The two costs are not independent draws; they are the same clock event seen from two different ledgers.

**Trap — Three ways the financing line is mis-modelled**

*Charging calendar days.* Days financed are *value-date* days, not wall-clock days. A position opened Monday morning and closed Friday afternoon is held for about four and a quarter days by the clock and financed for six, because its value date moves from Wednesday to the following Tuesday. The gap is not an overcharge; it is what T+2 means. Exercise 6.2 constructs a hold where wall clock and financed days differ by two.

*Averaging the two sides.* $f_{\text{long}}$ and $f_{\text{short}}$ are not symmetric about zero and never can be, because $\alpha$ enters both with the same sign. A backtest that models financing as $\pm$ one number by direction is wrong on one side by $2\alpha NS/D$ per day, and wrong in the optimistic direction on whichever side it happens to trade more.

*Assuming the rate is a market rate.* $\alpha$ is a provider-set commercial parameter, not an observable. It changes without any interest rate changing, it differs by instrument and account tier, and it is the reason two brokers quoting the same spot can publish materially different swaps. Any figure for it, including this section's 1.00%, is indicative and must be verified against the broker's current published schedule.

**Your project — The financing line the candles cannot see**

The platform's stored candles are mid-only, and a mid contains no spread and no swap. Financing is therefore not merely under-modelled in a price-driven backtest — it is structurally absent, and no amount of care with the price series recovers it, because the information was never in the series. Three requirements follow, and they are cheap next to what they prevent. Carry a *separate financing ledger*, keyed to value dates rather than bars, that applies $f_{\text{long}}$ or $f_{\text{short}}$ to every position open at each 5pm New York instant, with the day-count taken from Chapter 4's weekday table rather than from a bar count. Populate it from the broker's own published swap schedule where available, and from Result 6.5 with an estimated $\alpha$ where not — and treat $\alpha$ as point-in-time, because a schedule revised in 2025 does not describe 2023 [Course Book §18.2]. Then report net-of-cost profit and loss with the financing line *shown separately* from the spread line: on any horizon past a week they are different sizes and different shapes, and a single blended "costs" figure hides the one that compounds. Worked Example 6.3(c) is the whole argument in one number — past about five financed days the financing line is the larger of the two, and the platform's 50-bar label horizon on the higher timeframes crosses that threshold routinely.

## 6.6 Margin, leverage, and dying well

Everything so far has been about the price of trading. This section is about the mechanism that decides whether the account survives long enough to pay it.

**Definition — Margin, free margin, and margin level**

*Used margin* is the collateral the broker requires against currently open positions: notional divided by the applicable leverage ratio, expressed in the account currency. *Equity* is the account balance plus the unrealised profit and loss of all open positions — the mark-to-market value of the account, not the cash in it. *Free margin* is equity − used margin: what remains available to open further positions or to absorb further loss. *Margin level* is the ratio

$$\text{ML} = \frac{\text{equity}}{\text{used margin}},$$

conventionally quoted as a percentage, and it is the single number the broker's risk system monitors. A margin call fires when ML falls to a stated warning threshold; a *stop-out* fires when it falls to a lower threshold, at which the system begins closing positions — typically the largest loser first — without the client's instruction and without regard to his opinion of the trade.

Two features of these definitions do most of the damage in practice. Equity moves with the mark, so a position moving against the client erodes the numerator of ML immediately and continuously, not at some later settlement. And used margin is normally recomputed on the *current* notional rather than frozen at the opening one, so the denominator moves too — in the helpful direction for a long whose price is falling, which slightly delays the stop-out, and in the unhelpful direction for a short whose price is rising. Worked Example 6.4 quantifies the difference; it is not large, but modelling it the wrong way makes a stop-out calculation wrong by several pips, and the reader's own broker's convention should be checked rather than assumed.

### Leverage caps

The applicable leverage is not the broker's choice alone. Retail leverage on CFDs is capped by regulation in most jurisdictions the reader could trade from, tiered by instrument class, with major currency pairs receiving the most generous tier and single equities and crypto the least. Chapter 4 recorded the structural picture and the indicative levels; they are policy choices, they have moved materially within the last decade, and they differ by jurisdiction and by client classification — retail versus wholesale or professional — so any specific ratio quoted in a textbook is a vintage. Treat this section's 30:1 as an arithmetic input and verify current specifications, for the reader's own jurisdiction and account classification, before sizing anything.

The structural facts that do not drift are worth separating from the numbers that do. A cap constrains the *maximum* position an account of a given size can open; it does not constrain the position the client actually opens, and it does not cap loss. A client using a fraction of his available leverage is unaffected by the cap entirely. And — this is the part Exercise 6.3 makes precise — if the client's own sizing rule is expressed as a target margin level rather than a target position size, tightening the cap reduces the position and lengthens the distance to the stop-out, but leaves the *fraction of equity destroyed at the stop-out* almost unchanged. The cap changes how fast an account can die, not how much of it dies. Position sizing does that, and only position sizing does that [Course Book §14.3].

**Worked Example 6.4 — A margin call and a stop-out, end to end**

A hypothetical account, constructed for this chapter and computed in `verify_mk06.py`, §6. Equity A$10,000; one standard lot long EURUSD ($N = 100{,}000$ EUR) opened at $S_0 = 1.0850$; AUDUSD $= 0.6510$ and held fixed throughout so that a single exchange rate moves. Leverage cap 30:1 (indicative — verify current specifications). The broker warns at a margin level of 100% and stops out at 50% (both indicative and provider-set).

**(a) Opening state.** The cross rate is EURAUD $=$ EURUSD$/$AUDUSD $= 1.0850/0.6510 = 1.6666667$, so the notional is

$$100{,}000 \times 1.6666667 = \text{A\$}166{,}666.67, \qquad \text{used margin} = \frac{166{,}666.67}{30} = \text{A\$}5{,}555.56.$$

Free margin is A$4,444.44 and the margin level is $10{,}000/5{,}555.56 = 180.00\%$. The pip value is $0.0001 \times 100{,}000 =$ US$10.00, or A$15.3610 at the given AUDUSD — Chapter 4's third pip-value convention, since neither leg of EURUSD is the account currency.

**(b) The two functions of price.** Both sides of the margin-level ratio depend on $S$:

$$\text{equity}(S) = E_0 + \frac{N\,(S - S_0)}{A}, \qquad \text{used}(S) = \frac{N\,S}{A\,L},$$

with $A$ the AUDUSD rate and $L$ the leverage ratio. Setting $\text{equity}(S) = \theta \cdot \text{used}(S)$ for a threshold $\theta$ and solving:

$$\boxed{S^*(\theta) \;=\; \frac{N S_0 - A E_0}{N\,(1 - \theta/L)}}$$

The numerator, $100{,}000 \times 1.0850 - 0.6510 \times 10{,}000 = 101{,}990$, is the same for every threshold; only the denominator changes.

**(c) Margin call.** At $\theta = 1.00$,

$$S^* = \frac{101{,}990}{100{,}000\,(1 - 1/30)} = \frac{101{,}990}{96{,}666.67} = 1.055069,$$

a move of $-299.31$ pips, or $-2.7586\%$. Equity there is A$5,402.30, used margin is A$5,402.30 (equal, as ML $= 100\%$ requires), and the client has lost A$4,597.70 — 45.98% of the account — on a move of under three per cent in the underlying. That ratio, roughly seventeen to one, is what thirty-times leverage means in practice.

**(d) Stop-out.** At $\theta = 0.50$,

$$S^* = \frac{101{,}990}{100{,}000\,(1 - 0.5/30)} = \frac{101{,}990}{98{,}333.33} = 1.037186,$$

a move of $-478.14$ pips ($-4.4068\%$). Equity is A$2,655.37 against used margin of A$5,310.73, exactly half, and the realised loss is A$7,344.63 — 73.45% of the starting equity. The gap between the warning and the execution is only 178.83 pips: on an instrument that can move that far in an hour, the margin call is not a period of grace, it is a notification.

**(e) Floating versus frozen margin.** Had the broker frozen used margin at its opening A$5,555.56, the stop-out would have fired at 1.037983, some 7.97 pips *earlier*. The floating convention helps the long slightly, because the denominator shrinks as the price falls. The direction reverses for a short. The effect is small — eight pips of a four-hundred-pip move — but it is the kind of detail that makes a simulated stop-out and a real one disagree, and it costs nothing to get right.

**(f) What the stop-out is not.** It is not a guarantee of A$2,655.37. It is an instruction to the broker's system to begin closing at the next available price, and §6.6's closing subsection is about what happens when there is no next price near the level. Suppose the market gaps and the first tradeable price after the trigger is 1.0100, some 271.9 pips through the stop-out level. Equity realised is then

$$10{,}000 + \frac{100{,}000\,(1.0100 - 1.0850)}{0.6510} = -\text{A\$1,520.74},$$

a shortfall of A$4,176.10 against the stop-out arithmetic and a *negative* account balance: the client now owes the broker money he never deposited.

### Negative balance protection

Part (f) is the reason negative-balance protection exists. Where it is mandated — and it is mandated for retail clients in several major jurisdictions, including the reader's, as part of the same product-intervention regimes that imposed the leverage caps — the client's liability is capped at the funds

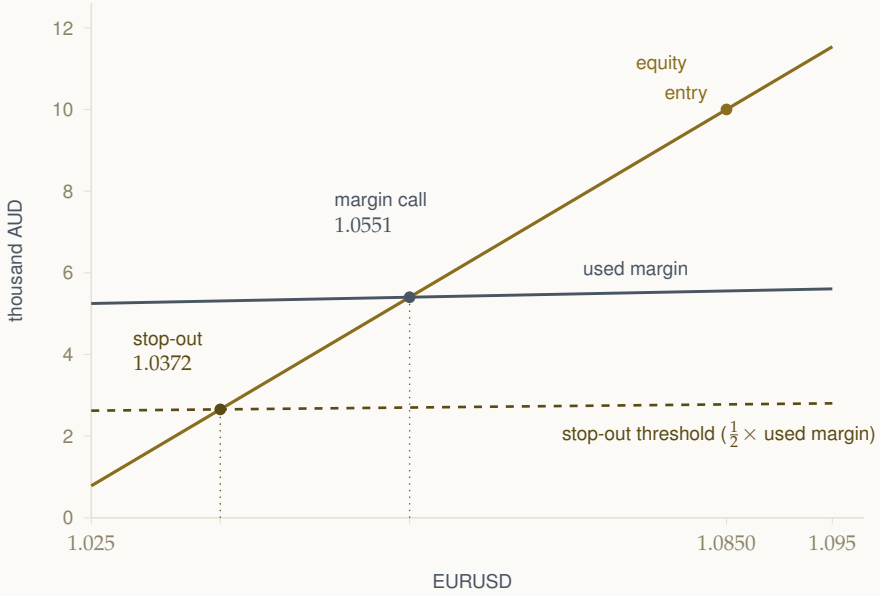



**Figure 6.5.** The two functions of price in Worked Example 6.4: equity falls steeply with EURUSD while used margin barely moves, so the margin level is driven almost entirely by the numerator. The margin call and the stop-out are the two crossings, 178.83 pips apart. Hypothetical account, computed in `verify_mk06.py`, §6.

in the account, and any deficit arising from a gap is borne by the broker rather than pursued from the client. The structural points are three. It is a *regulatory* protection attaching to a client classification, not a feature of the CFD contract, so it applies per jurisdiction and per classification and can be lost by opting up to professional status. It protects against *ruin beyond the account*, not against loss of the account: a client with negative-balance protection who gaps through a stop-out still loses everything he deposited. And it transfers the tail risk to the broker, which is one of the reasons brokers' own risk systems are as aggressive as they are about stop-outs — the firm is short the same gap the client is, without the option of ignoring it. Verify the current position for your own jurisdiction and account classification; do not infer it from a textbook.

### Gap risk, stated honestly

A stop order and a stop-out share one property that dominates every other: they are instructions to transact at the *next available price*, not at the level written on them. In continuous liquid trading the two are close and the distinction is pedantic. In a gap they are not close, and the distinction is the whole outcome. A weekend gap, a policy surprise, an abandoned peg or a liquidity vacuum can put the first tradeable price hundreds of pips through the level, and every protective mechanism the account has — the stop, the margin call, the stop-out — operates at that first tradeable price and not before it. The mechanisms are sequential in price but simultaneous in a gap: they all fire at once, at the same bad print, which is precisely when they are least useful.

This is not hypothetical, and it is not rare enough to ignore. Chapter 15 works the definitive case in full — the January 2015 Swiss franc episode, in which retail FX accounts using the reader's own instrument class, on the reader's own kind of leverage, discovered simultaneously that stops, margin calls, stop-outs and broker risk systems all fail together and for the same reason. That analysis, including what it did to the brokers themselves, belongs there and is not rebuilt here. What belongs here is the modelling consequence: a backtest that fills stops at their stated level is not modelling a stop, it is modelling a wish, and the correct treatment is a fill at the next traded price with the gap distribution taken seriously [Course Book §18.3]. The tail of that distribution is where the account dies, and it is exactly the region a Gaussian-tailed simulation has no mass in [Course Book §14.5].

## 6.7 What this means for the platform

Six sections of dealing-model mechanics reduce to a small number of obligations on the SST platform. They are stated here as obligations rather than suggestions because each one is forced by a mechanism this chapter derived, and the cost of ignoring any of them is quantified above.

**Your project — The dealing model, translated into platform requirements**

**One counterparty, stated plainly.** All twenty-one instruments are OANDA CFDs. Every price in the archive is OANDA's, every fill will be OANDA's, every financing charge is OANDA's schedule, and every stop-out will be OANDA's risk system. That is a concentration, and naming it is not criticism: it is the honest description of the object being modelled, and it means the platform's research question is "does this edge survive *this dealer's* cost and execution process", not "does this edge exist in EURUSD". Those are different questions with different answers, and conflating them is the most expensive error available in this project. The corollary is a diversification limit no model can remove: instrument diversification across twenty-one symbols does nothing about counterparty concentration across one.

**The cost surface is non-negotiable, and its estimator is now specified.** Mid-only candles cannot express the object §6.4 showed to be governing. The measured bid/ask surface must therefore be built, with per-cell *distributions* rather than means (Worked Example 6.2(b): the coefficient of variation exceeds one), *volume-weighted within the cell by the platform's own trade times* (part (c): time-weighting underprices by 1.567×), and applied point-in-time [Course Book §18.2]. The conditioning variables — session bucket, news proximity, volatility regime — are exactly the dealer's own state variables $(\mu, \Delta, \iota)$ in observable clothing, which is why they work.

**Sample the spread when the strategy trades, not when the clock ticks.** The single most important consequence of Worked Example 6.2(d). A spread series sampled on a uniform grid answers a question nobody is asking; the strategy's realised cost is the spread distribution conditional on *its own* entry and exit timestamps, and for a signal family that fires into volatility expansions those two numbers differ by a factor of 2.15 on this chapter's parameters. Build the sampler around the backtest's own trade log.

**Fill conservatively, and make the conservatism a parameter.** Chapter 3's discipline plus this chapter's mechanisms: enter at the ask and exit at the bid, never at the mid; add the measured adverse-slippage term of §6.3 on top of the quoted half-spread; model rejection as a non-fill with continued exposure rather than as a fill at a slightly worse price; and fill stops at the next traded price with the gap distribution intact [Course Book §18.3]. Each of these should be a named, tunable parameter, so that the sensitivity of a result to execution assumptions can be reported alongside the result itself.

**Carry the financing ledger separately.** Value-date day counts from Chapter 4's weekday table, $f_{\text{long}}$ and $f_{\text{short}}$ asymmetric by construction (Result 6.5), the Wednesday triple applied on the New York calendar rather than the Sydney one, and the financing line reported separately from the spread line. Past roughly five financed days the financing term dominates, and on the platform's higher timeframes the 50-bar label horizon crosses that threshold as a matter of course.

**Instrument the account, not just the model.** §6.3's three diagnostics — mean signed slippage, adverse-to-favourable ratio among slipped fills, and rejection rate conditional on session and news proximity — are computable from fill records the platform will generate anyway, and they are the only window it will ever have onto the routing and execution decisions of §6.2. Log requested price alongside filled price from the first live ticket, because the diagnostic is worthless retrospectively if the requested price was never stored.

**The Stage-0 gate, in dealing-model terms.** The programme's Stage-0 gate asks whether a strategy's edge survives realistic costs before any modelling sophistication is permitted [Course Book §14.7]. This chapter sharpens what "realistic" has to mean. It is not the average spread; it is the spread distribution conditional on the strategy's own trade times. It is not the quoted half-spread; it is the half-spread plus measured adverse slippage. It is not zero financing; it is the value-date-counted swap with the markup. And it is not a stop filled at its level; it is a stop filled at the next available price. A strategy with a high hit rate and negative realised expectancy — which is precisely the platform's own documented history — is the signature of a system whose edge is smaller than the gap between these four pairs, and every one of the four moves in the same direction [Course Book §1.6].

The uncomfortable summary is that the reader's counterparty knows more about the reader's expected cost than the reader does. It has his complete fill history, his holding-period distribution, his profitability by instrument and by hour, and it prices accordingly, continuously, under the routing logic of §6.2. He cannot see any of that. What he can do — and what this chapter has been building towards — is compute from his own records the same statistics his counterparty computes from theirs, and stop guessing about the ones he cannot. That is the entirety of the available response, and it is sufficient: nothing in this chapter requires knowing what the broker decided, only what the broker's decisions did to his fills.

## 6.8 Exercises

**Exercise 6.1 — The ledger continued, and the cost of hedging**

Resume Worked Example 6.1 where it stopped: the desk is short 400,000 EUR at a weighted-average price of 1.085429, the mid is 1.08570, and the desk rule is unchanged (0.1 pip of skew per 250,000 EUR, capped at 0.5; external hedge when $|I| \geq 1{,}000{,}000$). Three more tickets arrive, and the quoted width widens from 1.4 to 1.8 pips as the tape speeds up:

| # | Client | Size (EUR) | Mid | Quoted width |
|---|---|---|---|---|
| 7 | buys | 350,000 | 1.08572 | 1.4 pips |
| 8 | buys | 500,000 | 1.08591 | 1.8 pips |
| 9 | sells | 900,000 | 1.08585 | 1.8 pips |

When the hedge rule fires, the desk buys from an LP quoting a 0.6-pip width around the prevailing mid, and buys only enough to return to the 1,000,000 EUR limit. The mid after ticket 9 settles at 1.08560.

(a) Complete the ledger: dealt price, dealer inventory, weighted-average price and applied skew after each ticket, and state where the hedge fires and for how much.

(b) Compute spread capture over tickets 7 to 9, the cost of the hedge, and net spread capture.

(c) Compute the inventory carry over the same window, leg by leg, and verify that net spread capture plus inventory carry reproduces the tail's profit and loss.

(d) Give the total profit and loss over all nine tickets, and explain in one paragraph why the desk made money on the tail after losing on the first six.

▷ working space

**Solution — Exercise 6.1**

All figures computed in `verify_mk06.py`, §7.
**(a)** A client buy deals at mid plus the half-width, a client sell at mid minus it; the half-widths are 0.7, 0.9 and 0.9 pips respectively.

| # | Action | Dealt | Dealer $I$ | Avg price | Skew |
|---|---|---|---|---|---|
| 7 | dealer sells 350,000 | 1.08579 | −750,000 | 1.085597 | 0.30 |
| 8 | dealer sells 500,000 | 1.08600 | −1,250,000 | 1.085758 | 0.50 |
| | *hedge*: buys 250,000 from LP | 1.08594 | −1,000,000 | 1.085758 | 0.40 |
| 9 | dealer buys 900,000 | 1.08576 | −100,000 | 1.085758 | 0.04 |

The weighted average updates only on the two tickets that increase $|I|$: after ticket 7, $(400{,}000 \times 1.085429 + 350{,}000 \times 1.08579)/750{,}000 = 1.085597$, and after ticket 8, $(750{,}000 \times 1.085597 + 500{,}000 \times 1.08600)/1{,}250{,}000 = 1.085758$. The hedge and ticket 9 both reduce $|I|$ and leave the average alone. The hedge fires immediately after ticket 8, at which $|I| = 1{,}250{,}000 \geq 1{,}000{,}000$; the skew cap is reached exactly there, since $0.1 \times 1{,}250{,}000/250{,}000 = 0.5$ pips. The desk buys the excess 250,000 EUR at the LP's ask, $1.08591 + 0.3$ pips $= 1.08594$.
**(b)** Spread capture is size times half-width on each client ticket:

$$350{,}000 \times 0.00007 + 500{,}000 \times 0.00009 + 900{,}000 \times 0.00009 = 24.50 + 45.00 + 81.00 = \$150.50.$$

The hedge is not revenue but cost: the desk crosses the LP's half-spread of 0.3 pips on 250,000 EUR, $= \$7.50$. Net spread capture is $150.50 - 7.50 = \$143.00$. Note the ratio: 5.0% of the tail's gross spread revenue was spent removing a fifth of the inventory, which is exactly why Result 6.3 places hedging last in the ordering.
**(c)** Inventory carried across each move in the mid, starting from the −400,000 position at 1.08570:

| Position | Mid from | Mid to | Leg P&L |
|---|---|---|---|
| −400,000 | 1.08570 | 1.08572 | −\$8.00 |
| −750,000 | 1.08572 | 1.08591 | −\$142.50 |
| −1,000,000 | 1.08591 | 1.08585 | +\$60.00 |
| −100,000 | 1.08585 | 1.08560 | +\$25.00 |
| | | Total | −\$65.50 |

Then $143.00 - 65.50 = \$77.50$, which is exactly the tail's contribution: the nine-ticket total is −\$39.00 against the six-ticket total of −\$116.50, a difference of +\$77.50. The decomposition reconciles.
**(d)** Total over nine tickets: cash of \$108,521.00 against a closing position of −100,000 EUR marked at 1.08560 for −\$108,560.00, giving −\$39.00. The tail was profitable for three reasons, none of which is that the desk got better at forecasting. It widened — the 1.8-pip width earned 0.9 pips per unit instead of 0.7, worth an extra \$28.00 on the 1.4 million traded at the wider quote. It got the flow it had been skewing for: ticket 9's 900,000 EUR client sale cut the position by ninety per cent in one ticket, which is what the skew ladder existed to recruit. And the mid came back, so the last leg of inventory carry was positive rather than negative. Reversing the causality is the trap here: the desk did not profit because it predicted the reversal, it profited because it was still short when the reversal arrived, having been unable to get flat any faster. On a batch where the mid had continued upward the same three tickets would have produced a larger loss than the first six did. This is the whole risk of warehousing in one paragraph — the revenue is reliable and the inventory term is not, and the desk's control over the second is limited to three imperfect levers.

**Exercise 6.2 — Swap financing across a week, including the triple roll**

A hypothetical GBPUSD scenario. Spot $S = 1.2740$; position $N = 200{,}000$ GBP; annualised rates $r_b = 4.60\%$ on GBP and $r_q = 4.90\%$ on USD; broker markup $\alpha = 0.90\%$; day-count basis $D = 365$. The account is denominated in AUD and AUDUSD is 0.6510. All rate and spread levels are indicative; verify current specifications before pricing anything live.

(a) Compute the one-day swap points from covered interest parity, then the broker's $f_{\text{long}}$ and $f_{\text{short}}$ per day, in USD and in pips.

(b) A short position is opened at 10:00 New York on a Tuesday and closed at 11:00 New York on the following Monday. Which rollovers fire, how many days does each finance, and what is the total? Verify the total independently using the value dates at open and at close.

(c) Compute the total financing in USD and in AUD, for the short and for a long of the same size. Express each as a percentage of one round trip at an indicative 1.6-pip quoted width, and find the number of financed days at which each side's financing equals that round trip.

(d) A different position is held from 09:00 New York on a Monday until 14:00 New York on the Friday of the *following* week. Count the financed days and compare with the wall-clock days held. Explain the discrepancy.

▷ working space

**Solution — Exercise 6.2**

All figures computed in `verify_mk06.py`, §8.

**(a)** With $\tau = 1/365$ and $r_q > r_b$ by 0.30%, the swap points are

$$F - S = 1.2740 \times \frac{(0.0490 - 0.0460)/365}{1 + 0.0460/365} = 1.0470 \times 10^{-5} = 0.10470 \text{ pips},$$

so sterling is at a small forward premium against the dollar. Applying Result 6.5 with $NS = \$254{,}800.00$:

$$f_{\text{short}} = \frac{254{,}800\,[(+0.0030) - 0.0090]}{365} = -\$4.18849 \;=\; -0.20942 \text{ pips},$$
$$f_{\text{long}} = \frac{254{,}800\,[(-0.0030) - 0.0090]}{365} = -\$8.37699 \;=\; -0.41885 \text{ pips}.$$

Both sides pay, since $\alpha = 0.90\%$ comfortably exceeds the 0.30% differential, and the long pays exactly twice the short — a coincidence of these particular numbers ($0.0120 = 2 \times 0.0060$), not a general rule.

**(b)** The rollovers that fire while the position is open are Tuesday, Wednesday, Thursday and Friday; Monday's 17:00 cut is not reached, since the position closed at 11:00. Using Chapter 4's weekday day-counts,

$$\text{Tue}(1) + \text{Wed}(3) + \text{Thu}(1) + \text{Fri}(1) = 6 \text{ days}.$$

Independent check by value dates: a trade dealt on Tuesday has value date Thursday of the same week (T+2 business days); a trade dealt on the following Monday has value date Wednesday of that week. Thursday of week one to Wednesday of week two is six calendar days. The two methods agree, as they must — days financed is a value-date span, and the weekday table is only a convenient way of accumulating it.

**(c)** Six days at each rate:

$$\text{short:}\quad 6 \times (-4.18849) = -\$25.1310 = -\text{A}\$38.6036,$$
$$\text{long:}\quad 6 \times (-8.37699) = -\$50.2619 = -\text{A}\$77.2072,$$

converting at AUDUSD 0.6510. One round trip at 1.6 pips on 200,000 GBP costs $0.00016 \times 200{,}000 = \$32.00$, or A\$49.1551. So the short's week of financing is 78.53% of a round trip and the long's is 157.07% — the long has already paid more in financing than in spread before the week is out. The crossovers are

$$\frac{32.00}{4.18849} = 7.6400 \text{ financed days (short)}, \qquad \frac{32.00}{8.37699} = 3.8200 \text{ financed days (long)},$$

the second being exactly half the first because $f_{\text{long}} = 2 f_{\text{short}}$ here. Neither is a long time: with a triple Wednesday inside any full week, 3.82 financed days is reached on the Wednesday night of the first week for a position opened Monday.

**(d)** Nine rollovers fire: Monday to Friday of week one, then Monday to Thursday of week two (Friday's 17:00 cut in week two is not reached). The day-counts are $(1{+}1{+}3{+}1{+}1) + (1{+}1{+}3{+}1) = 7 + 6 = 13$ financed days. The wall-clock holding period is Monday to the Friday eleven days later: 11 days. The two differ by two, and the reason is that financing is charged on *value dates*, not on elapsed time. The position's value date at open is Wednesday of week one (Monday +2 business days); at close it is Tuesday of week three (Friday of week two +2 business days). Wednesday of week one to Tuesday of week three is thirteen days — exactly the day-count total, and two more than the clock, because the T+2 convention at

the close reaches two business days further forward than the T+2 convention at the open did, and one of those spans includes a weekend. A backtest that finances by bar count or by elapsed hours will be short two days of charge on this hold, which is $8.38 on the short side and $16.75 on the long — around a quarter and a half of a round trip respectively, arising purely from a calendar convention.

**Exercise 6.3 — Margin and stop-out at two leverage caps**

An account holds equity of A$10,000. EURUSD is 1.0850 and AUDUSD is 0.6510, both held fixed except where EURUSD is the variable being solved for. The account's own rule is never to open a position that would put the margin level below 200%; the broker stops out at 50%; used margin floats with the current notional. Consider two regulatory leverage caps on major pairs: 30:1 and 10:1 (both indicative — verify current specifications).

(a) Under each cap, find the largest position the house rule permits, in EUR and in standard lots, and its pip value in AUD.

(b) Find the EURUSD level at which each position stops out, and the adverse move in pips.

(c) Find the equity remaining and the loss incurred at each stop-out, in AUD and as a percentage of starting equity.

(d) Interpret. What does halving the leverage cap actually protect, and what does it not protect? Confirm your answer by asking what happens if the client attempts the 30:1 position size under the 10:1 cap.

▷ working space

**Solution — Exercise 6.3**

All figures computed in `verify_mk06.py`, §9. Throughout, EURAUD $= 1.0850/0.6510 = 1.6666667$.

**(a)** The house rule fixes used margin at open: ML $= 200\%$ means used $= E_0/2 =$ A\$5,000 in both cases. Affordable notional is used margin times the cap, and position size is notional divided by EURAUD:

| Cap | Used margin | Notional | Position | Pip value |
|---|---|---|---|---|
| 30:1 | A\$5,000 | A\$150,000 | 90,000 EUR (0.90 lots) | A\$13.8249 |
| 10:1 | A\$5,000 | A\$50,000 | 30,000 EUR (0.30 lots) | A\$4.6083 |

The position ratio is exactly 3.0, the ratio of the caps.

**(b)** Using the stop-out solution of Worked Example 6.4(b), $S^* = (NS_0 - AE_0)/\big(N(1-\theta/L)\big)$ with $\theta = 0.50$:

$$30{:}1: \quad S^* = \frac{90{,}000 \times 1.0850 - 0.6510 \times 10{,}000}{90{,}000\,(1 - 0.5/30)} = \frac{91{,}140}{88{,}500} = 1.029831,$$

$$10{:}1: \quad S^* = \frac{30{,}000 \times 1.0850 - 0.6510 \times 10{,}000}{30{,}000\,(1 - 0.5/10)} = \frac{26{,}040}{28{,}500} = 0.913684.$$

The adverse moves are $-551.69$ pips and $-1{,}713.16$ pips respectively — a ratio of 3.1053, slightly more than the 3.0 ratio of position sizes because the floating-margin denominator shrinks faster over the longer move.

**(c)** Equity at the stop-out is half the used margin there:

| Cap | Used margin at stop-out | Equity | Loss | Loss % |
|---|---|---|---|---|
| 30:1 | A\$4,745.7627 | A\$2,372.8814 | A\$7,627.1186 | 76.27% |
| 10:1 | A\$4,210.5263 | A\$2,105.2632 | A\$7,894.7368 | 78.95% |

A cross-check on the first row: $7{,}627.1186/13.8249 = 551.69$ pips, matching (b) exactly.

**(d)** The two accounts lose A\$7,627 and A\$7,895 — a difference of A\$267.62, or 2.68 percentage points of starting equity. Halving the leverage cap changed the loss at stop-out by less than three per cent of the account. What it changed by a factor of three is the *distance* the market must travel to get there: 552 pips against 1,713. So the cap buys survival probability, not loss containment. Conditional on being stopped out at all, both accounts are destroyed to roughly the same degree, because the stop-out rule is expressed as a ratio of equity to margin and the client's sizing rule was also expressed as a ratio — so the leverage cancels out of the terminal state and survives only in the path. If the client wants to lose less *when* he is wrong, the instrument is position size, not the regulator's ceiling [Course Book §14.3].

The confirmation is the failed attempt: the 90,000 EUR position under a 10:1 cap requires $150{,}000/10 =$ A\$15,000 of margin against A\$10,000 of equity, so it cannot be opened at all. That is the cap doing the only thing a cap can do — refusing a size — and it is worth noticing that it binds only on a client who was going to use the full allowance anyway. A client sizing at a quarter of his available leverage would not notice a change in the cap in either direction, which is the sense in which the regulation protects the least disciplined clients and is invisible to the rest.

**Exercise 6.4 — Essay: which dealing model would you rather face?**

Argue, as a *client* rather than as a regulator or a broker, which of the three handling models of §6.2 — pure A-book, pure B-book, or hybrid routing — you would rather be facing, and

why. Your answer must do three things. State the criterion you are optimising: it is not obviously "the model with no conflict of interest", and you should say what it actually is for a systematic trader with a measurable edge. Identify the evidence available in your own fill and statement data that discriminates between the models, using the diagnostics of §6.3 and the spread mechanics of §6.4, and say what each piece of evidence would have to look like to move you. And state honestly what your data cannot tell you. Aim for a page.

▷ working space

**Solution — Exercise 6.4**

A model answer; other defensible positions exist, and the criterion chosen determines the conclusion, so the argument matters more than the verdict.

**The criterion.** "No conflict of interest" is the wrong objective, and noticing that is most of the work. What a systematic trader wants is the lowest and most *predictable* all-in cost of executing his strategy, since that is the quantity his expectancy is computed against [Course Book §14.1]. A conflict of interest matters only to the extent that it changes that number or its variance. Under B-book handling the conflict is direct but the execution is typically clean and internal: no LP rejection, no last look in the client's path, and fills at the quoted price because the broker is the price. Under A-book handling there is no directional conflict, but the client inherits the wholesale layer's frictions — rejections, requotes and asymmetric slippage — and typically pays a wider all-in cost, because the broker's markup now sits on top of an LP spread it must actually cross. The trader who has never had an edge worth warehousing is, on cost grounds alone, better served by the B-book. This is an uncomfortable conclusion and it is the correct one.

**The asymmetry that reverses it.** The reason not to want a pure B-book is not the conflict but the *option*. A pure B-book firm carries the client's tail: if the client is right about a gap, the firm owes him money it may not have, and a firm whose risk system is warehousing correlated retail positions into a dislocation is exactly the counterparty most likely to fail when the client's position finally pays. Chapter 15's January 2015 case is the canonical demonstration. A pure A-book firm has no such exposure; a hybrid firm has whatever exposure its classifier left it with. So the criterion has two terms — expected cost, which favours the B-book, and counterparty survival in the states where the client's biggest gains occur, which favours the A-book — and the hybrid is a defensible optimum precisely because it routes the dangerous flow out and warehouses the harmless flow, which is the same trade-off the client would make on the firm's behalf if he could. The honest answer for a trader who intends to become profitable is the hybrid, with the expectation of being reclassified once he does, and the recognition that the reclassification will itself widen his costs.

**The evidence.** Four statistics discriminate, all computable from the client's own records. *Mean signed slippage and the adverse-to-favourable ratio* among slipped fills (§6.3): near-symmetric slippage and near-zero mean is consistent with internalisation, where the broker is the price and has no wholesale fill to pass through; strongly asymmetric slippage with a $z$ statistic of the order computed in that section is evidence of a wholesale path with an option in it. *Rejection and requote rate, conditional on session bucket and news proximity*: a B-booked account should see near-zero rejections at all times, because there is nothing to reject; a rising conditional rejection rate into releases is evidence of pass-through. *Spread behaviour versus the dealer-side mechanics of §6.4*: widening that tracks the three structural causes (news, rollover, thin hours) is what both models predict and discriminates between neither; widening that appears in one's own instruments and not in others, or at times unrelated to the structural causes, is evidence of something account-specific. *A change in all of the above coinciding with a change in one's own profitability*: this is the single strongest signal available, because a routing reclassification is the only mechanism that predicts a simultaneous shift in slippage symmetry, rejection rate and effective spread with no corresponding change in market conditions. Establishing a baseline now, before profitability changes, is what makes that test possible later — and it is not reconstructible after the fact if requested prices were never logged.

**What the data cannot say.** It cannot identify the routing decision, because the client's contract is identical either way and no field on a confirmation encodes it. It cannot separate a broker's markup change from a wholesale liquidity change, since the client sees only the sum (§6.1). It cannot detect targeting from a single account's history, since the discriminating comparison — his fills against other clients' fills at the same instant — is not available to him. And it has

essentially no power in the tail: the states that determine whether the counterparty survives are, by construction, ones that have not yet occurred in the sample. The correct posture is therefore neither trust nor accusation but instrumentation: measure what is measurable, state plainly what is not, and size the position against the possibility that the unmeasured part is worse than assumed.

**Key idea — What to remember**

A retail FX CFD is a bilateral contract with one company, priced off that company's own manufactured quote — liquidity-provider streams, aggregated to a composite tighter than any single stream, then marked up to a retail width several times wider — so the reader's exposure is to EURUSD and to his broker jointly, and the second exposure is not optional. Whether that broker hedges the trade or warehouses it is an internal, invisible, revocable decision, profitable in the warehousing case because retail flow is low-$\mu$ flow in the Glosten–Milgrom sense and because netting across many clients leaves a small directional residual — profitable in expectation, never riskless, as the inventory ledger's losing batch shows. The spread the reader pays is the dealer's breakeven width $\mu\Delta + 2c + 2\iota$ plus competitive rent, and because all three components rise together at news, at rollover and in thin hours, the spread's distribution is right-skewed enough that its standard deviation exceeds its mean and its *correct* average depends entirely on whether it is weighted by the clock, by market volume, or by the strategy's own trade times — three answers differing by a factor of two. Financing is the interest differential in price units plus a markup that makes both directions pay, charged on value dates rather than elapsed time, tripled every Wednesday in New York, and larger than the round-trip spread past about five financed days; margin is a ratio whose numerator moves with every tick, so a leverage cap buys distance to the stop-out rather than protection from loss, and every protective mechanism the account has fires at the same bad price in a gap. Chapter 7 turns from the dealer's own book to the market's aggregate one: what positioning data can and cannot reveal about where the crowd is, and how a report built for regulators becomes a covariate.

# 7

# Positioning, Flow, and Sentiment Data

*Every object in this chapter answers the question "what are other participants doing", never "what will price do next" — and the distance between those two questions is exactly what this book's method exists to police. The Commitments of Traders report is the spine: what the CFTC actually collects, the three report families the platform's own point-in-time build already names by their exact CFTC titles, and the reference-Tuesday/release-Friday lag that produces the single most frequently repeated leakage warning in this library. From there to net positioning and its z-score, to signed order flow and its retail shadow, to risk-on/risk-off as a cross-asset regime description rather than a trade: each object is introduced as a market fact with a citable structure, then handed immediately to the same two questions — what does it plausibly mean, and has anyone shown it survives contact with costs and multiple testing. Nothing here is asserted as a working signal; the deflation machinery [Course Book §15.4] is the only authority permitted to say so.*

## 7.1 The Commitments of Traders report

Positioning data records not what a price did but what identifiable groups of participants *hold*: a stock of exposure at a point in time, rather than a flow of trades. It is a different information source from everything Chapter 4 discussed, arrives on its own institutional schedule, and carries its own honest limits. The Commodity Futures Trading Commission's (CFTC) weekly *Commitments of Traders* (COT) report is the most complete and longest-running public positioning dataset in existence, and every positioning feed this platform reads — FX, metals, the two index CFDs, and, thinly, the two crypto instruments — ultimately traces back to it, which earns it the detailed, structural treatment below.

**What the CFTC collects.** The CFTC requires every futures commission merchant, clearing member and foreign broker to report to it the futures and options-on-futures positions of any trader whose position in a given contract exceeds a reporting threshold that the Commission sets, and periodically revises, per contract (CFTC). A reportable trader further declares, at account opening, whether its trading in that contract is *commercial* — hedging a bona fide business exposure to the underlying — or not; the Commission aggregates by contract and category and publishes weekly. Neither the threshold levels nor the roster of covered contracts is reproduced here: both drift over time and belong to the "verify current specifications" class this book flags rather than states as permanent. Positions are recorded as of *Tuesday's* close of business and released the following *Friday* afternoon, US Eastern time; the Commission publishes two parallel variants for most contracts, *Futures Only* and *Futures and Options Combined* (the latter folding each option's futures-equivalent delta into the same category totals) (CFTC).

The report comes in three family variants, distinguished by how finely each splits the non-hedging side of the book. Every covered contract carries a Legacy report; a contract then carries *either* a Disaggregated report, for physical commodities, *or* a Traders in Financial Futures report, for financial contracts including every FX future — never both.

**Definition — The Legacy report**

The original, coarsest split, published for every covered contract:

| Category | Definition |
|---|---|
| Commercial | Declared hedgers: entities using the contract against a bona fide exposure to the underlying (a producer, processor, merchandiser, or end user). |
| Non-Commercial | Large traders holding no such declared hedging exposure — the reportable speculative bucket: CTAs, macro funds, and other large discretionary or systematic traders. |
| Non-Reportable | The residual: open interest minus the sum of all reportable long (or short) positions. Small traders below threshold, never observed individually. |

Legacy's Commercial bucket mixes two economically distinct motives: a wheat farmer hedging next season's crop and a swap dealer warehousing the risk of an OTC commodity-index client's exposure are both "commercial" under Legacy, despite one hedging physical production and the other hedging a derivatives book. The CFTC's response, for physical-commodity markets, was the Disaggregated report, introduced in 2009 (CFTC):

**Definition — The Disaggregated report**

Physical-commodity markets only (metals and WTI, among this platform's instruments):

| Category | Definition |
|---|---|
| Producer/ Merchant/ Processor/User | Entities engaged in producing, processing or merchandising the physical commodity, using futures predominantly to hedge that activity. |
| Swap Dealers | Entities that facilitate commodity-linked swaps and hedge *that* swap book with futures — a genuinely different economic motive from physical hedging, despite Legacy folding both into Commercial. |
| Managed Money | Registered commodity trading advisors, commodity pool operators, and other money managers trading on clients' behalf — the bucket most often read as "speculative" or "trend" money. |
| Other Reportables | Every other large trader not fitting the first three categories. |

The same critique applies with even more force to financial futures, where "hedging" and "speculating" are blurrier still: a dealer bank's FX futures position may hedge its own market-making book, and a pension fund's position is neither hedging physical production nor purely speculative in the retail sense. The Traders in Financial Futures (TFF) report answers this for interest-rate, FX, equity-index and credit futures (CFTC) — the family that matters most for this platform, since every FX instrument becomes a financial future once read through a CME ticker:

**Definition — The Traders in Financial Futures report**

Financial futures, including FX, rates and equity indices (this platform's FX pairs and its two index CFDs):

| Category | Definition |
|---|---|
| Dealer/ Intermediary | The sell side: banks and dealers intermediating derivatives, typically hedging their own market-making or client-facilitation book. |
| Asset Manager/ Institutional | Real-money institutional investors — pension funds, insurers, mutual and other portfolio managers — generally not hedging a directly analogous "production" exposure. |
| Leveraged Funds | Hedge funds, CTAs and similar — the speculative bucket most FX and rates COT commentary means by "speculators"; TFF's analogue of Disaggregated's Managed Money. |
| Other Reportables | Every other large trader not fitting the first three categories. |

Leveraged Funds is the category most FX commentary means when it says "speculators are net long"; Dealer/Intermediary is the sell-side bucket a Legacy report would have folded into Commercial alongside genuine end-user hedgers.

**The reference-Tuesday/release-Friday lag.** Every COT report describes positions as they stood at Tuesday's close and reaches the public the following Friday afternoon — a reference date of 11 August 2026 (a Tuesday) is published 14 August 2026 (a Friday), three calendar days later (`verify_mk07.py` §1) — and if that Friday is a US federal holiday the release slips to the next business day (CFTC). This is the single fact about COT this library states most often, because it is the single easiest fact to get wrong in a backtest: joining a bar to the report by its *reference* date rather than the date it actually became public smuggles three days of look-ahead into every observation, more across a long weekend or a holiday-shifted release. [Course Book §18.2] states the general fix — join on when a value became knowable, never on what it describes — and the Playbook's Chapter 3 build carries it out end to end for this platform's five point-in-time feeds: the exact release calendar, the holiday-shift rule, and the rollover from a Friday-afternoon release to the next *tradeable* session, computed to the timestamp rather than the calendar day as this book does. Section 7.5 returns to what that discipline buys.

**Trap — Report date is not release date**

Take reference Tuesday 11 August 2026, released Friday 14 August 2026. A model that joins on the reference date reads Tuesday's positioning as if it were known on the Tuesday it describes — Wednesday and Thursday's bars would then trade on a number that will not exist publicly for one to three more days. The fix is not a correction factor added after the fact; it is joining on the *release* timestamp from the start, so the defective join is structurally impossible rather than merely discouraged.

### 7.1.1 Which markets have a COT report, and the futures-as-proxy problem

COT exists only for contracts traded on a US, CFTC-regulated futures exchange (CME, ICE Futures U.S., COMEX, NYMEX and similar) — a fact with sharp consequences for a platform whose actual exposure is mostly OTC. Spot and forward FX, the market Chapter 4 described as the dominant venue by turnover, has no COT report of its own: this platform's "FX COT" feed reads large-trader positioning in CME's currency *futures* contract, a centrally-cleared, exchange-traded instrument that coexists alongside, but is materially smaller than, and populated by a different participant mix from, the OTC spot/forward/swap market that OANDA's CFD prices actually reference. Futures participants skew toward CTAs, macro funds and asset managers who prefer exchange-cleared FX exposure; the interbank dealers and corporate and real-money clients who dominate OTC turnover (Chapter 4) are barely represented on a CME ticker at all. Reading CME EUR/USD futures positioning as a signal about the OTC EUR/USD market this platform trades is therefore reading a *proxy for a subset of the participant base*, not the market itself — an honest caveat stated once, here, and carried silently into every later use of a COT feed.

The proxy problem compounds for this platform's non-USD crosses. Of the fourteen FX pairs in the SST book, only the USD-leg majors — each currency directly against the US dollar — map onto a single CME contract; a cross such as EUR/GBP or GBP/JPY has no CME contract of its own, and the operational fix, synthesising a net position as one USD-leg's Legacy Non-Commercial net minus the other's, is an approximation stacked on an approximation, compounding rather than resolving the proxy's error (the Playbook's Chapter 3 build states this synthesis rule and flags it explicitly). Two further gaps are worth naming plainly. First, this platform's crypto instruments: CME lists Bitcoin and Ether futures, but the resulting COT series is short — Bitcoin's begins in 2018, Ether's in 2021 — and thin, undefined before contract inception and never backfilled (CFTC). Second, and sharper still: an equity index CFD referencing a non-US benchmark inherits *no* COT proxy at all, however approximate, because the underlying index future does not trade on a CFTC-regulated exchange in the first place; only a US-listed equivalent, an E-mini-style contract, generates a reading, and that reading describes a related but distinct market. Exercise 7.4 returns to the inferential consequences of building an FX signal on a proxy this indirect.

## 7.2 Reading positioning: net, z-scores, and the crowding narrative

Every category in every COT family reports two raw numbers, long contracts and short contracts; the object anyone actually reads is their difference.

**Definition — Net position and its standardised (z-scored) measure**

For a trader category's weekly reports $k = 1, 2, \ldots$ in ascending reference-period order, with reported long and short contracts $L_k, S_k$:

$$\text{net}_k = L_k - S_k.$$

Raw net drifts with open interest itself — more participants over a decade means larger raw nets with no change in crowding — so comparability requires standardising against the category's *own* recent history, using only reports strictly before $k$:

$$\hat{\mu}_k = \frac{1}{W}\sum_{i=k-W}^{k-1} \text{net}_i, \qquad \hat{\sigma}_k = \sqrt{\frac{1}{W-1}\sum_{i=k-W}^{k-1} (\text{net}_i - \hat{\mu}_k)^2}, \qquad z_k = \frac{\text{net}_k - \hat{\mu}_k}{\hat{\sigma}_k},$$

undefined for $k \leq W$. A rise in $|z_k|$ signals that the current report sits further from its own recent baseline than usual — the operational definition of "crowded" this book uses throughout. Common choices of window run from $W = 52$ (one year of weekly reports) to $W = 156$ (three years, the value the Playbook's Chapter 3 build recommends for this platform): short enough to track a genuine shift in the market's participant base, long enough to span more than one positioning cycle before recalibrating — the same adaptivity-versus-stability trade-off every backward-only standardisation in this programme faces [Course Book §13.8].

**Worked Example 7.1 — A hypothetical net-position path and its trailing $z$-score**

The following eight weekly Non-Commercial reports for a hypothetical CME FX future are an illustrative construction, not observed CFTC data:

| Week | Long | Short | Net |
|---|---|---|---|
| 1 | 120,000 | 95,000 | +25,000 |
| 2 | 128,000 | 90,000 | +38,000 |
| 3 | 135,000 | 88,000 | +47,000 |
| 4 | 142,000 | 81,000 | +61,000 |
| 5 | 150,000 | 77,000 | +73,000 |
| 6 | 158,000 | 70,000 | +88,000 |
| 7 | 165,000 | 64,000 | +101,000 |
| 8 | 179,000 | 52,000 | +127,000 |

With only eight reports available, a genuine $W = 52$ or $W = 156$ window cannot be shown by hand; the arithmetic below uses $W = 4$ purely to keep every step checkable, computed and verified in `verify_mk07.py` §2 — the mechanics are identical at any window length, and the platform runs the same formula at its full three-year window.

| $k$ | Window net (weeks $k-4, \ldots, k-1$) | $\hat{\mu}_k$ | $\hat{\sigma}_k$ | $z_k$ |
|---|---|---|---|---|
| 5 | 25,000; 38,000; 47,000; 61,000 | 42,750.0000 | 15,152.0076 | 1.9964 |
| 6 | 38,000; 47,000; 61,000; 73,000 | 54,750.0000 | 15,413.7385 | 2.1572 |
| 7 | 47,000; 61,000; 73,000; 88,000 | 67,250.0000 | 17,442.7635 | 1.9349 |
| 8 | 61,000; 73,000; 88,000; 101,000 | 80,750.0000 | 17,442.7635 | 2.6515 |

Net long positioning has risen in every one of the seven week-on-week changes above, and by week 8 sits $z_8 = 2.6515$ above its own trailing baseline: a textbook "extreme" by construction.

### 7.2.1 Extremes, and two ways to read them

An extreme $z$-score is read as *crowding*: a large share of the reporting speculative community leaning the same way. What that extreme means is genuinely contested, and this book states both readings without adjudicating between them.

**Trend confirmation.** Positioning accumulates because a trend is genuinely under way and participants have correctly read it; an extreme long $z$-score simply co-occurs with, and may partly constitute, an ongoing move. On this reading, $z_8 = 2.6515$ above is evidence the trend has further to run, not a warning.

**Contrarian at extremes.** A crowded long position means most of the willing marginal buyer base has already bought: there is limited fresh demand left to extend the move, and if a shock arrives the unwind is one-directional, since everyone already long has no one left to sell to but other existing longs. Brunnermeier, Nagel, Pedersen (2008) formalise exactly this mechanism for carry trades — crowded positioning unwinds abruptly rather than gradually, "up the stairs, down the lift" — and Brunnermeier, Pedersen (2009)'s funding-liquidity spiral describes the general case: a forced unwind by leveraged holders tightens the funding conditions that force further unwinds. On this reading, $z_8 = 2.6515$ above is a warning, not a confirmation.

The honest empirical status is mixed: published support exists for each reading, on different samples, horizons, and definitions of "extreme," and neither dominates the literature cleanly. [Course Book §15.4] states the standard this book actually applies: both readings are hypotheses, and neither is believed until it clears a deflation-adjusted, cost-aware, purged evaluation — a covariate that *sounds* fundamental earns no exemption from that gate.

### 7.2.2 Managed money against commercial: the divergence reading

A second, related reading compares categories against each other rather than one category against its own history: when the speculative bucket (Managed Money in Disaggregated, Leveraged

Funds in TFF) is heavily long while the hedging bucket (Producer/Merchant, or Legacy Commercial) is heavily short, the two are said to be *diverging*. Keynes (1930)'s theory of normal backwardation supplies the oldest version of the underlying story: hedgers who need to lay off price risk pay speculators an expected premium to bear it, so the two sides of the book are not symmetric participants making unbiased bets but a risk transfer struck at a price. Applied to reading COT categories, that asymmetry motivates two different conclusions depending on which reading in Section 7.2.1 is held: a trend-confirmation reading treats a large hedger short position as corroboration — why else would the participants closest to the underlying market still choose to sell — while a contrarian reading treats the mirror-image large speculative long as the crowded, vulnerable side of exactly the same fact. Exercise 7.1's Managed-Money-against-Producer/Merchant snapshot is constructed to admit both readings at once, by design: this book does not resolve the ambiguity, because it is not resolved.

## 7.3 Order flow: what actually moves price

Positioning is a stock; order flow is the trade-by-trade process that built it, and Chapter 4 already introduced the empirical claim that flow, not public news arriving between trades, is where FX information actually surfaces. This section recaps that evidence in the vocabulary this chapter needs and states plainly what a retail participant can and cannot see of it.

**Definition — Signed order flow**

Over an interval containing trades $i = 1, \ldots, n$,

$$\mathrm{OF} = \sum_{i=1}^{n} \mathrm{sgn}(\mathrm{trade}_i) \cdot \mathrm{size}_i, \qquad \mathrm{sgn}(\mathrm{trade}_i) = \begin{cases} +1 & \text{buyer-initiated (aggresses the ask)} \\ -1 & \text{seller-initiated (aggresses the bid).} \end{cases}$$

Signed order flow measures net aggressive buying against net aggressive selling pressure, and is distinct from raw volume: $|\mathrm{OF}| \leq \sum_i \mathrm{size}_i$, with equality only when every trade in the interval runs the same direction.

**Result — Order flow explains price — the Evans–Lyons finding, recapped**

Chapter 4 introduced Evans and Lyons (2002)'s central empirical result on interdealer FX data: signed order flow explains a large share of daily exchange-rate variation, a finding that overturned the prevailing view of flow as noise around a fundamentals-driven price and gave the microstructure approach to exchange rates (Lyons, 2001) its central empirical anchor. The result is consistent with the informed-trading theory behind it generally: order flow aggregates private information no single dealer's quote can see directly (Kyle, 1985; Glosten, Milgrom, 1985), so a persistent imbalance reveals something dealers update on, and price moves because dealers reprice against it, not because a headline crossed the wire at that instant.

None of the above is available to a retail account. Signed interdealer flow of the kind Evans and Lyons used is proprietary to the venues and banks that internalise it — EBS and Refinitiv's matching engines, and the handful of dealers who see the bulk of client flow; at best, a retail trader sees one broker's own book. Some brokers publish an aggregate ratio: the share of that broker's own open retail positions held long versus short in a given instrument, sometimes updated through the trading day. Read this structurally, not as a market fact — it describes one broker's clientele, not the market, and its composition (predominantly trend-following retail, or predominantly counter-trend scalpers, or whatever mix that broker happens to attract) shifts the ratio in ways unrelated to the wider market's actual positioning. If OANDA, or any other broker, publishes such a ratio, it is indicative and broker-specific by construction, never a market-wide

flow measure, and never stated as one here.

**Trap — Sentiment indicators of unknown construction**

A vendor selling a proprietary "sentiment" index is selling a black box: without knowing exactly what population it samples — positions, orders, and survey opinions are three entirely different objects — what window it aggregates over, and how it weights or filters its inputs, there is no way to state *why* the resulting number should predict anything. The deflation machinery [Course Book §15.4] needs a stated hypothesis to test against costs and multiple comparisons; an indicator whose construction is undisclosed forecloses that step before it starts. Treat any such feed as untested by default, whatever track record its marketing claims, until its construction is known well enough to state what would make it work.

The pattern across this section is one asymmetry, not several: the flow that Evans and Lyons showed actually moves price sits with the institutions positioned to see the whole book, and what filters down to a retail account — one broker's own long/short split, or a vendor's undisclosed sentiment score — is a weak, selection-biased shadow of it. That is not a data gap better research closes; it is a structural fact about who gets to see what, in an OTC dealer market built exactly as Chapter 4 described.

## 7.4 Risk-on/risk-off as a market regime

Risk-on/risk-off (RORO) is not an asset, a trade, or even a single series: it is a label for a co-movement pattern across many assets, driven by a common latent factor most naturally read as global risk appetite or, equivalently, the tightness of global funding conditions. Rey (2015)'s global financial cycle names the same latent factor at the macro level — capital flows, leverage and asset prices across borders co-move with a common driver substantially independent of any one country's own policy stance, so that a shift abroad, not in this platform's own FX pairs, can move every leg of a RORO basket at once. That common-factor structure is what makes the co-movement below a coherent regime description rather than a coincidence of correlated headlines.

A canonical risk-off episode co-moves as follows. The yen and Swiss franc, both long-standing funding and haven currencies, strengthen; a broad US dollar index typically strengthens too, though this is the least reliable of the four signs — the dollar is sometimes the haven bought on a flight to quality and sometimes the currency sold as the Federal Reserve eases hardest into the same shock, so treat its sign as regime-dependent rather than fixed. High-beta, growth-sensitive currencies — the Australian and New Zealand dollars prominent among them — and emerging-market currencies weaken against the majors. Global equity indices fall together, largely irrespective of the idiosyncratic story behind any one constituent. Gold is *conditional*: safe-haven demand pushes it up, but the same episode often carries higher real yields or a stronger US dollar, both of which push it down, and both forces are frequently present at once, so gold carries no clean a priori sign the way the other four legs do.

**Worked Example 7.2 — A hypothetical risk-off week and a two-basket composite gauge**

The following weekly returns are an illustrative, hypothetical scenario, not observed market data, constructed to match the theoretical co-movement above exactly:

| Instrument | Raw return | Oriented value | Basket |
|---|---|---|---|
| USD index (DXY-style) | +1.2% | +1.2 | Haven |
| USD/JPY (JPY strengthens) | −1.8% | +1.8 | Haven |
| USD/CHF (CHF strengthens) | −1.1% | +1.1 | Haven |
| AUD/USD | −2.4% | +2.4 | Growth |
| NZD/USD | −2.1% | +2.1 | Growth |
| EM FX basket (vs USD) | −1.6% | +1.6 | Growth |
| S&P 500 | −3.5% | +3.5 | Growth |
| ASX 200 | −2.8% | +2.8 | Growth |
| Gold (XAU/USD) | +0.6% | — | excluded, conditional |

Reorienting each series so a positive value means "moved the way the risk-off theory predicts" — sign-flipping only where the natural USD-quote convention would otherwise obscure the direction (Chapter 4's quoting conventions apply directly) — and averaging within, then across, the haven and growth baskets (`verify_mk07.py` §4):

| Composite | Value |
|---|---|
| Haven-basket average ($n = 3$) | +1.3667 |
| Growth-basket average ($n = 5$) | +2.4800 |
| Overall gauge (all 8, oriented, equal-weighted) | +2.0625 |

All eight oriented signs agree, so the composite reads unambiguously positive: a risk-off week by construction. Gold's raw +0.6% happens to carry the same sign as the haven basket's average here — consistent with, but not proof of, the safe-haven story — which is exactly why it is reported separately rather than folded into the mechanical average: in a different episode dominated by a real-yield shock, gold's sign could as easily have disagreed with every other leg.

A RORO *gauge* generalises the worked composite above into a standing, continuously updated series. Take a basket of cross-asset "barometer" series — typically some blend of an equity-index or implied-volatility measure, a credit-spread or funding-stress proxy, one or two haven-versus-growth FX crosses, and, more cautiously, gold; orient each so a rise means more risk-off; standardise each causally against its own backward-only window, using precisely the $\hat{\mu}_k, \hat{\sigma}_k$ machinery Section 7.2's COT $z$-score already introduced; and combine, by an average or a more deliberate weighting, into one scalar reading per bar. This platform's own gauge is a frozen engineering artefact: its exact input list and combining formula live in running code, documented and version-controlled as the Playbook's Chapter 3 build requires, and are not reproduced here. The family of construction, however, is exactly this: a backward-window composite of cross-asset signals, not a proprietary secret sauce.

[Course Book §13.8]'s Stage-2 router is this gauge's consumer: it reads the composite to classify *what kind of week this is*, conditioning how other stages behave, and issues no trade of its own. Treating a RORO reading as a signal — buy when the gauge says risk-on, sell when it says risk-off — collapses a regime description into a trading rule it was never built to be, a category error the platform's staged architecture is deliberately built to avoid: the router conditions; it does not decide.

Every sign stated in this section is a historical regularity, not a physical constant. Correlations between growth-sensitive and haven assets shift across decades and can invert within a single crisis: BIS (2020)'s bulletins on the March 2020 turmoil document a "dash for cash" in which even conventional havens sold off alongside risk assets for several sessions, as investors of every stripe raised US-dollar liquidity indiscriminately rather than rotating into the usual safe harbours.

Menkhoff, Sarno, Schmeling, Schrimpf (2012)'s finding that carry returns turn sharply negative exactly in high-volatility regimes is the FX-specific version of the same instability: the correlation a carry book relies on in calm markets is precisely the one that breaks when it matters most. Chapter 14 returns to this instability directly and formally; here it is enough to flag that a RORO gauge built today on today's correlation structure is not guaranteed to describe tomorrow's crisis.

## 7.5 Using positioning data honestly

Section 7.1's reference-Tuesday/release-Friday lag is one instance of a general rule [Course Book §18.2]: a series is usable at time $t$ if and only if it was already knowable at $t$, which means joining on *when* a value was released, never on the period it describes, and, for series subject to revision, on which *vintage* of the value existed at $t$ rather than whichever value happens to be on file today. COT itself is not usually revised — a published report is not corrected after the fact the way a first-estimate payrolls or GDP print is — so its only defect is release-lag; the other four point-in-time feeds this platform maintains (FRED-sourced macro levels, the RORO gauge, the VolRegime gauge, and every yield differential) inherit release-lag *and* revision risk, since several sit on top of macro series that genuinely restate. The Playbook's Chapter 3 build carries out the fix for all five, end to end: one schema (series, reference period, release timestamp, value, ingestion vintage), one as-of query against the release timestamp, a documented holiday-shift calendar for COT specifically, and a standing audit asserting, over the full backfilled history, that the query never returns a value dated after the query time. Nothing in this chapter substitutes for that build; this chapter states why it matters and reads its output.

That output feeds directly into how the platform's forecasting model is allowed to treat each series. [Course Book §8.4]'s covariate taxonomy for the Temporal Fusion Transformer distinguishes *known-future* inputs — the calendar date, a scheduled release time, anything fixed in advance — from *observed-past* inputs, values known up to and including the current bar but not beyond. A correctly point-in-time-joined COT $z$-score, RORO reading, or yield differential is a canonical observed-past covariate: it may condition the model's state at bar $t$, but the model must never be handed a version of it that a real-time system could not yet have queried at $t$. Get the join wrong — read a value one day before it was actually eligible — and an observed-past covariate is silently promoted to something no real trading system could ever have known: exactly the failure the point-in-time build exists to make structurally impossible rather than merely policed after the fact.

**Your project — The Stage-2 gate decides, not this chapter**

Everything in this chapter — net positioning and its $z$-score, the managed-money-versus-commercial divergence reading, order flow's retail shadow, the RORO composite — is a candidate covariate, and candidacy is all this chapter grants it. [Course Book §13.8]'s Stage-2 regime router is where the platform actually tests whether any of it carries conditional information: a Chronos-2 covariate probe run alongside the live TimesFM envelope, evaluated through the same purged, embargoed, cost-aware, multiple-testing-corrected machinery as every other candidate signal in the programme, never exempted because the input happens to describe what other traders are doing rather than price itself. The FX crosses without a direct CME contract — Section 7.1.1's synthesised nets — are a natural stress case for that gate: whatever signal COT-for-FX carries on the USD-leg majors, the added approximation error of a leg-minus-leg synthesis should show up as weaker, noisier evidence on the crosses, and a probe that fails to show that gap is itself worth investigating before the covariate is trusted anywhere. This is v1's territory — the primary FX book, where COT and RORO are tested first — and nothing here is promoted to v2's options overlays, v3's pairs, or v4's commodities without clearing the same gate again on that book's own instruments and costs.

## 7.6 Exercises

The four exercises below span the chapter: reading a live-looking snapshot, the release-calendar arithmetic underneath every $z$-score, designing a gauge from scratch, and confronting the proxy problem in prose rather than arithmetic.

**Exercise 7.1 — Reading a disaggregated snapshot**

A hypothetical Disaggregated report for a crude-oil future shows the following long and short contracts by category (illustrative, not real CFTC data):

| Category | Long | Short |
|---|---|---|
| Producer/Merchant/Processor/User | 400,000 | 600,000 |
| Swap Dealers | 320,000 | 280,000 |
| Managed Money | 300,000 | 100,000 |
| Other Reportables | 80,000 | 120,000 |

(a) Compute each category's net position and the percentage of its own gross position held long.

(b) Which category is most crowded, by the definition of crowding in Section 7.2? Which is most net short?

(c) Producer/Merchant/Processor/User and Managed Money sit on opposite sides here. State, in one sentence each, what the trend-confirmation reading and the contrarian reading (Section 7.2.1) would conclude about Managed Money's position.

▷ working space

**Solution — Exercise 7.1**

**(a)** (`verify_mk07.py` §3):

| Category | Net | % long |
|---|---|---|
| Producer/Merchant/Processor/User | −200,000 | 40.0% |
| Swap Dealers | +40,000 | 53.3% |
| Managed Money | +200,000 | 75.0% |
| Other Reportables | −40,000 | 40.0% |

**(b)** Managed Money is most crowded: 75.0% of its own gross position is long, the largest departure from an even split of any category, and it carries the largest net long, +200,000. Producer/Merchant/Processor/User is the most net short, at −200,000 — the mirror image.
**(c)** Trend confirmation: Managed Money's extreme long reflects genuinely informed bullish positioning, and Producer/Merchant/Processor/User's matching short is hedgers confirming the same view from the physical side — the move has further to run. Contrarian: Managed Money's 75.0% long skew leaves little fresh buying power in reserve; if a shock arrives, this is the crowded, vulnerable side of the book, not the informed one. Section 7.2.1 does not choose between these, and neither does this exercise.

**Exercise 7.2 — The $z$-score meets the release calendar**

A hypothetical GBP-flavoured CME future reports the following five weekly Non-Commercial positions, with reference Tuesdays chosen to run into a US federal holiday (illustrative, not real CFTC data):

| Week | Reference Tuesday | Long | Short |
|---|---|---|---|
| 1 | 2026-11-24 | 85,000 | 102,000 |
| 2 | 2026-12-01 | 90,000 | 95,000 |
| 3 | 2026-12-08 | 97,000 | 88,000 |
| 4 | 2026-12-15 | 101,000 | 80,000 |
| 5 | 2026-12-22 | 108,000 | 70,000 |

(a) Compute each week's net position, and the $z$-score of week 5's net using the preceding four weeks as the window ($W = 4$).

(b) State the scheduled release date for each week under the reference-Tuesday-plus-three-days rule; identify which one falls on a US federal holiday, and the date the holiday-shift rule (Section 7.1) moves it to.

(c) Using [Course Book §18.2]'s eligibility rule — usable from the next weekday after the (possibly shifted) release date — state the most recent week whose $z$-score is legitimately usable on each of: (i) Saturday 19 December 2026; (ii) Friday 25 December 2026; (iii) Monday 28 December 2026; (iv) Tuesday 29 December 2026.

▷ working space

**Solution — Exercise 7.2**

**(a)** Net positions (`verify_mk07.py` §5): week 1 −17,000; week 2 −5,000; week 3 +9,000; week 4 +21,000; week 5 +38,000. Window (weeks 1–4): $\hat{\mu}_5 = 2,000.0000$, $\hat{\sigma}_5 = 16,532.7957$, so

$$z_5 = \frac{38,000 - 2,000.0000}{16,532.7957} = 2.1775.$$

**(b)** Scheduled release Fridays: week 1, 27 Nov 2026; week 2, 4 Dec 2026; week 3, 11 Dec 2026; week 4, 18 Dec 2026; week 5, 25 Dec 2026. Week 5's scheduled release, 25 December 2026, is Christmas Day, a US federal holiday; the holiday-shift rule walks forward skipping the weekend that immediately follows, landing on Monday 28 December 2026.
**(c)** Eligible-from dates: week 1, Mon 30 Nov; week 2, Mon 7 Dec; week 3, Mon 14 Dec; week 4, Mon 21 Dec; week 5, Tue 29 Dec.

(i) Sat 19 Dec 2026: latest usable is **week 3** — week 4's eligibility (21 Dec) is still two days away.

(ii) Fri 25 Dec 2026: latest usable is **week 4** — week 5 has not even been released yet, since the CFTC does not publish on Christmas Day.

(iii) Mon 28 Dec 2026: still **week 4** — week 5's actual (shifted) release lands today, but eligibility requires the next trading day *after* the release, not the release day itself.

(iv) Tue 29 Dec 2026: **week 5** is now eligible; this is the first date its $z$-score, 2.1775, may legitimately be used.

Queries (ii) and (iii) both resolve to week 4 despite falling on different calendar days three days apart, illustrating exactly how much extra delay one holiday shift adds on top of the ordinary Friday-release-to-Monday-eligibility lag that queries against an un-shifted week would show.

**Exercise 7.3 — Designing a RORO gauge from five series**

You are given five daily series: (i) an equity-index implied-volatility level; (ii) AUD/JPY spot; (iii) a broad USD index; (iv) a credit-spread proxy (e.g. a high-yield option-adjusted spread); (v) gold (XAU/USD). For each, state a sign convention (does a rise mean more risk-on or more risk-off), a standardisation window with a one-line justification, and any causality or circularity concern that should shape how it is used. No computation is required; answer in a short table plus supporting prose.

▷ working space

**Solution — Exercise 7.3**

| Series | Rise means | Window | Flag |
|---|---|---|---|
| Implied vol | Risk-off | Short (days–weeks): vol clusters and mean-reverts quickly; a long window would smear the state-dependence the gauge exists to capture. | None material. |
| AUD/JPY | Risk-*on* (sign-flip needed) | Same backward-only causal window as the COT $z$-score (Section 7.2). | If the platform includes AUD/JPY among its fourteen FX pairs (plausible: a liquid, commonly quoted cross), the router's state is partly a transform of a price series a downstream stage is simultaneously forecasting; backward-only standardisation contains, but does not eliminate, the concern. |
| USD index | Ambiguous in general; more often broad strength in an acute risk-off episode. | Medium (weeks): separates the fast risk component from the slower differential-driven drift. | The standard basket (EUR, JPY, GBP, CAD, SEK, CHF) is a weighted transform of several of the platform's own FX legs — a near-certain structural overlap, sharper than AUD/JPY's; cross-check the gauge with this input removed. |
| Credit spread | Risk-off | Longer than the vol index's: spreads move more slowly. | No FX-instrument overlap, but typically vendor-sourced and revisable — needs the same release-timestamp point-in-time treatment as any macro input [Course Book §18.2]. |
| Gold | No fixed convention | N/A as a hard rule | Section 7.4's conditional discussion applies directly: exclude from a mechanical composite, as the worked example did, or include at reduced weight with its sign left to vary. |

**Exercise 7.4 — Essay: COT-for-FX is a proxy — what that does to inference**

In no more than one page, explain why treating CME currency-futures positioning as "FX positioning" is a proxy relationship rather than a direct measurement, and state what that

does to any inference drawn from a COT-based FX covariate — including how the proxy problem differs between a USD-leg major and a synthesised cross.

▷ working space

**Solution — Exercise 7.4**

Two mismatches, stacked. First, a venue and population mismatch: Chapter 4 established that spot, forward and swap OTC trading dominates FX turnover and that interbank dealers and their corporate and real-money clients drive it, while CME's currency futures are a materially smaller, separately populated market skewed toward CTAs, macro funds and asset managers who prefer exchange-cleared exposure. A COT reading on that contract describes this second population, not the first. Second, a coverage mismatch: COT exists only because CME is a CFTC-regulated US exchange; nothing about the OTC market itself is reported to, or reportable by, the CFTC, so even a perfect futures-market reading is necessarily a reading of a different, if correlated, market rather than an imperfect reading of the same one.

For a USD-leg major, that is the whole proxy problem: one contract, one population mismatch, stated once and carried forward. For a synthesised cross — $\text{net}_{\text{leg1}}$ minus $\text{net}_{\text{leg2}}$, both Legacy Non-Commercial versus USD — a second layer stacks on top. The subtraction implicitly assumes the two legs' futures books are additive stand-ins for the cross's own OTC positioning: a trader long EUR/USD futures and long GBP/USD futures need not be running a EUR/GBP view at all, and the trader base active in each USD-leg future need not overlap with whoever actually sets the cross's OTC price. The synthesis is an approximation of an approximation, and its error does not merely add to the major's proxy error, it compounds with it.

The consequence for inference is the same at both levels, only sharper for the cross: any observed relationship between a COT-based $z$-score and subsequent price moves is evidence about the behaviour of CME futures participants, at best correlated with the OTC market's own positioning, never a direct measurement of it. Practically, that means (i) a COT covariate earns no exemption from the deflation machinery [Course Book §15.4] on the strength of sounding fundamental; (ii) a result that clears that gate on a USD-leg major must not be assumed to transfer to a synthesised cross without testing the cross separately, since the added approximation error could dilute a genuine signal into noise there, or, less plausibly, manufacture a spurious one; and (iii) triangulating a COT reading against an independent, differently-biased flow proxy — Section 7.3's broker sentiment ratio, despite its own broker-specific caveats — is more informative than trusting either proxy alone, precisely because their errors do not share a common source.

**Key idea — What to remember**

Positioning, flow, and regime data describe what other participants are doing, or a proxy for it — never what price will do next — and every series in this chapter reaches a model only after being dated by when it became knowable rather than what it describes. Net positioning and its $z$-score, signed order flow, and RORO composites are all read through the same two competing lenses, trend confirmation against contrarian crowding, and neither lens is asserted true here: the honest empirical status is mixed, on purpose, because a stated position would be an assertion this book's method forbids. The reason to build all of this carefully is not that positioning data is special; it is that it is data, subject to exactly the point-in-time discipline, leakage taxonomy, and multiple-testing correction that governs every other covariate in the programme, and it earns a place in a live decision only by clearing Stage-2's gate, never by intuitive appeal. Chapter 8 leaves this FX-specific covariate world for equities and indices, where a different positioning problem — short interest, borrow cost, and index-rebalancing flow — reopens the same questions on a different instrument class.

PART III

# Equities and Indices



LUNA LABS

# 8

# Stock Markets

*Equities are the pillar of this book with the most institutional furniture: listed venues, index committees, clearing houses, disclosure regimes, and a securities-lending market that makes a short position economically different from a long one rather than its mirror image. This chapter builds that furniture from the order-book primitives of Chapter 2 upward — continuous trading bracketed by opening and closing auctions, at both the deep US market and the reader's own local venue, the ASX — and then spends most of its length on the arithmetic a systematic desk cannot skip: how an index level and its divisor are actually computed, why a dividend, a split, a buyback or a rights issue each rewrites a raw price series in a different and mechanically predictable way, and what it costs, in hard percentage points, to hold a share short. None of this is optional colour. Every one of these mechanics is a way an unadjusted or naively modelled price series lies to a backtest, and the chapter's running trap — corporate actions as a backtest hazard, not a footnote — is the thread that connects the index-construction arithmetic of Section 8.3 to the leakage taxonomy already built in the Course Book and to the adjusted-series discipline the platform's v3 equity-pairs work depends on.*

## 8.1 What a Share Is and How It Lists

**Definition — Ordinary share**

An ordinary (common) share is a residual claim on a company's assets and earnings: its holder is paid, in liquidation, only after every creditor and every preference shareholder has been paid in full, and receives a dividend only if and when the board declares one. In exchange for standing last in the queue, the ordinary shareholder typically carries the company's only unlimited upside and, under limited liability, a strictly bounded downside — the most that can be lost is the capital paid for the shares, never more, regardless of the size of the company's subsequent liabilities. One ordinary share ordinarily carries one vote at general meetings, the mechanism by which the class as a whole exercises control over the board, subject to the share-class caveat below.

A market in these claims separates cleanly into two layers that this book keeps distinct throughout. In the *primary market*, the company itself is a party to the trade: it issues new shares directly to buyers and receives the proceeds, whether at an initial public offering (IPO), a follow-on placement, or a rights issue (Section 8.4). In the *secondary market*, existing shares change hands between investors on an exchange or over the counter; the company is not a party, receives no proceeds, and its balance sheet is unaffected by the trade's price. Everything in Section 8.2 onward is secondary-market machinery — it is where liquidity, price discovery, and (for this book's purposes) index and factor structure live — but the primary market is where the security is created, and its two central frictions are worth naming once. An IPO is underwritten: one or more investment banks commit, formally or by best-efforts undertaking, to place the new shares, running a book-build in which institutional investors submit indicative demand across a price range so the underwriters can set an offer price intended to clear the book without leaving an

excessive amount on the table; the company and its underwriters then carry the (well-documented, cross-market) risk of first-day mispricing in either direction. Retail participation in the primary leg itself is typically a minority allocation; most retail exposure to a newly listed company is secondary-market exposure, bought on-exchange from the first day of trading onward exactly like any other listed name.

Not every ordinary share in a company is identical. Preference shares sit between debt and ordinary equity: they typically carry a fixed or formula-based preferential dividend that must be paid before any ordinary dividend, rank ahead of ordinary shares (though behind all creditors) in a wind-up, and usually carry no vote — a hybrid instrument, priced more like a bond with equity-linked features than like the ordinary shares of the same issuer. Separately, a company can list more than one class of *ordinary* share with identical economic (dividend and liquidation) rights but different voting weight per share — a dual-class structure, most visible among founder-controlled companies listed on US exchanges, that concentrates control in the class typically held by founders/insiders while the publicly traded class carries reduced or no votes; ASX listing practice has historically been markedly less receptive to such structures than the major US exchanges, a stance itself subject to periodic review and best treated as indicative of current practice rather than a permanent rule. Finally, one line on cross-border access: an American Depositary Receipt (ADR) is a US-dollar certificate, issued by a US depositary bank and traded on a US exchange or OTC, representing a fixed number of shares of a non-US company held in local custody — it lets a US investor gain economic exposure to (say) an ASX- or LSE-listed company through an ordinary US brokerage trade, without touching the foreign settlement system directly.

## 8.2 The Venues

Chapter 2 derived the two primitives every listed equity venue in this chapter assembles: the continuous double auction, in which resting limit orders form a book and an arriving marketable order walks that book level by level, and the single-price call auction, in which orders accumulated over a window are uncrossed at the one price that maximises matched volume. A listed equity's trading day layers these primitives in a fixed sequence — an opening call auction that discovers the first tradeable price from overnight and pre-open order accumulation, a continuous session in which the book trades as Chapter 2 described, and a closing call auction — and the reason closing auctions are disproportionately deep, worked there as a stand-alone uncrossing, recurs here on a larger scale in Section 8.3: index-tracking funds are structurally required to transact at (or as close as possible to) the closing print, so the close concentrates a share of the day's volume far above its share of the trading day's length.

**US structure.** Listed US equities trade across more than a dozen registered national exchanges operating in parallel under a National Best Bid and Offer (NBBO) linkage regime, rather than on one consolidated venue — among them exchanges descended from NYSE's floor-based, designated-market-maker model and from Nasdaq's all-electronic, competing-dealer model, joined since by several purely electronic entrants. Layered on top of the lit exchanges is a substantial volume of trading that never displays a quote on any of them: alternative trading systems (ATSs, including so-called dark pools) that match orders without pre-trade display, and wholesaler internalisation, in which a large share of retail marketable order flow is filled by a wholesaling dealer off-exchange rather than routed to a lit book. Both are described here purely structurally; the fraction of consolidated volume that trades off-exchange is a figure that moves with rule changes, dealer economics and market conditions, is commonly cited as running above two-fifths of consolidated volume through the mid-2020s, and should be treated as indicative and verified current rather than quoted as a fixed constant. The same caution applies to settlement: US equities moved to a T+1 settlement cycle in 2024, shortening the T+2 convention that had held for the prior several years — a concrete, dateable instance of exactly the kind of "stable"

institutional parameter this book's market-facts discipline insists on re-verifying, since it changed live, market-wide, within the recent past.

**ASX structure — the reader's local venue.** The ASX runs the same three-phase day on a single national market: a staggered pre-open auction phase (orders accumulate from around 7:00am AEST, with different security groups opening in a fixed sequence to spread the uncrossing load), continuous trading from approximately 10:00am to 4:00pm AEST, and a single-price closing auction a few minutes after the continuous session ends (commonly around 4:10–4:12pm AEST). Treat every clock time in this paragraph as indicative and confirm against the current ASX Operating Rules before using it operationally; auction windows, randomised uncrossing delays and phase boundaries are exactly the sort of parameter an exchange amends periodically. Settlement and registry are structurally distinctive relative to the US: CHESS (the Clearing House Electronic Subregister System) both settles ASX trades and maintains the official electronic sub-register of legal title for most ASX-quoted securities, a combined settlement-and-registry role with no exact US analogue. ASX settlement has run on a T+2 cycle for most of the 2020s, but — as the US T+1 move above illustrates for a comparable market — exchanges do periodically shorten settlement cycles, ASX has at various points targeted replacing CHESS's underlying technology, and both the settlement cycle and the replacement programme's status should be verified current rather than assumed static.

**Trading halts and circuit breakers.** Both markets pause trading, by rule, when price moves too far too fast, at two granularities. At the market-wide level, US exchanges apply tiered circuit breakers keyed to a decline in a broad benchmark (historically the S&P 500) from the prior close, triggering a coordinated trading halt across listed venues once a threshold is crossed — the direct regulatory descendant of the 1987 crash, whose one-day decline of that magnitude the mechanism exists specifically to interrupt (Brady Commission, 1988). At the single-stock level, US markets run a limit-up/limit-down (LULD) band around a reference price for each name individually, pausing trading in that name alone when its price would otherwise move outside the band — a mechanism introduced and repeatedly refined after the 6 May 2010 Flash Crash exposed how far and how fast a single name's price could dislocate in a fragmented, high-speed market before any market-wide breaker engaged (CFTC-SEC, 2010; Kirilenko, Kyle, Samadi, Tuzun, 2017). The ASX runs structurally analogous protections at the single-security level — a company-requested trading halt pending release of material, price-sensitive information (the standard mechanism a listed company itself invokes ahead of an announcement, revisited in Section 8.5) and an automated pause keyed to an extreme, fast intraday price move in a single security — with the exact trigger thresholds, halt durations and eligible-order rules in both jurisdictions treated here as structural facts with indicative, verify-current parameters, not as fixed numbers to memorise.

## 8.3 Indices

An index is not a market fact to be observed; it is a formula, published and administered by an index provider, applied to a fixed rule for selecting and weighting constituents. Three weighting schemes cover essentially every index this book's reader will encounter.

**Definition — Index construction: three weighting schemes**

For constituents $i = 1, \dots, n$ with price $P_{i,t}$, total shares outstanding $Q_i$, and a free-float factor $f_i \in (0, 1]$ (the fraction of $Q_i$ available to ordinary investors, excluding closely held strategic, founder or government stakes), an index level at time $t$ is

$$\textbf{price-weighted:} \quad I_t = \frac{\sum_{i=1}^{n} P_{i,t}}{D_t},$$

$$\textbf{cap-weighted:}\quad I_t = \frac{\sum_{i=1}^{n} P_{i,t} Q_i}{D_t},$$

$$\textbf{float-adjusted:}\quad I_t = \frac{\sum_{i=1}^{n} P_{i,t} Q_i f_i}{D_t},$$

where the divisor $D_t$ is fixed at the index's inception so $I_0$ equals a chosen base value, and held fixed thereafter except at a corporate action or constituent change, when it is reset (Result 8.3) so the index level is continuous through the event.

Price-weighting (the Dow Jones Industrial Average is the canonical survivor of this convention) is purely a legacy of pre-computer index arithmetic: a $1 move in any constituent moves the index by the same number of points regardless of that company's economic size, so a high-priced but small-capitalisation name can dominate the index purely because of where its price happens to sit, with no correction for shares outstanding at all. Cap-weighting fixes this by weighting every constituent by its total market value; float-adjustment refines it further by excluding shares an ordinary investor cannot actually buy (a founder's control block, a government's strategic holding), so the index tracks the capitalisation an investor can actually acquire, and rebalancing flows (below) trade against float, not total shares. The S&P/ASX 200, the S&P 500 and the Nasdaq-100 are all float-adjusted cap-weighted; this book's arithmetic below works with plain cap-weighting for clarity and notes explicitly where float-adjustment would change a number.

**Result — A cap-weighted index return is a weighted sum of constituent returns**

Let $w_{i,t-1} = P_{i,t-1}Q_i / \sum_j P_{j,t-1}Q_j$ be constituent $i$'s cap-weight and $r_{i,t} = P_{i,t}/P_{i,t-1} - 1$ its simple return, holding $Q$ and $D$ fixed over $(t-1, t)$ (no corporate action in the interval). Then the index's own return decomposes exactly as

$$R_t^I = \frac{I_t}{I_{t-1}} - 1 = \sum_{i=1}^{n} w_{i,t-1}\, r_{i,t}.$$

*Derivation.* With $D$ fixed, $I_t / I_{t-1} = \left(\sum_i P_{i,t}Q_i\right) / \left(\sum_i P_{i,t-1}Q_i\right)$, so

$$R_t^I = \frac{\sum_i P_{i,t}Q_i - \sum_i P_{i,t-1}Q_i}{\sum_j P_{j,t-1}Q_j} = \sum_i \underbrace{\frac{P_{i,t-1}Q_i}{\sum_j P_{j,t-1}Q_j}}_{w_{i,t-1}} \cdot \underbrace{\frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}}}_{r_{i,t}} = \sum_i w_{i,t-1} r_{i,t}. \qquad \blacksquare$$

No such identity holds for a price-weighted index in terms of economically meaningful weights: the quantity that plays the role of $w_{i,t-1}$ there is $P_{i,t-1} / \sum_j P_{j,t-1}$, a share of the index's arithmetic determined by where the price happens to sit, not by market value.

**Worked Example 8.1 — A four-stock cap-weighted index, and one constituent's 10% move**

Four hypothetical companies — Alpha, Beta, Gamma, Delta — have the prices and shares outstanding below (illustrative scenario, not market data).

| Stock | Price ($) | Shares (m) | Market cap ($m) | Cap-weight |
|---|---|---|---|---|
| Alpha | 40.00 | 500 | 20,000 | 31.25% |
| Beta | 25.00 | 800 | 20,000 | 31.25% |
| Gamma | 10.00 | 1,200 | 12,000 | 18.75% |
| Delta | 60.00 | 200 | 12,000 | 18.75% |
| Total | – | – | 64,000 | 100.00% |

Choosing a base index value of 1,000 fixes the divisor at $D_0 = 64{,}000/1{,}000 = 64$, so $I_0 = 64{,}000/64 = 1{,}000.00$. Before moving on, the price-weighted contrast: summing raw prices instead gives $\sum_i P_i = \$135$, in which Delta — tied for the *smallest* market capitalisation in the table — carries a price-weight of $60/135 = 44.44\%$, well over double its true 18.75% economic share; this is the distortion price-weighting builds in by construction, not an edge case.

Now suppose Alpha rises 10%, to \$44.00, with the other three prices unchanged. Alpha's new market cap is $44.00 \times 500 = \$22{,}000$m, so the new total is $22{,}000 + 20{,}000 + 12{,}000 + 12{,}000 = \$66{,}000$m and the new index level is $66{,}000/64 = 1{,}031.25$, an index return of $1{,}031.25/1{,}000 - 1 = 3.125\%$. Result 8.3 gives the same number directly, with no need to touch the other three prices: only Alpha moved, so $R^I = w_{\text{Alpha}} \times r_{\text{Alpha}} = 31.25\% \times 10\% = 3.125\%$. A cap-weighted index's response to any one constituent's move is exactly that constituent's weight times its return — the identity that makes "the index is up because X did Y" a computable, checkable sentence rather than a narrative one.

**Result — The divisor-update formula**

Write $A_t$ for whichever numerator aggregate the index uses ($\sum_i P_{i,t}$ for a price-weighted index, $\sum_i P_{i,t}Q_i$ or $\sum_i P_{i,t}Q_i f_i$ for a cap- or float-weighted one). At a corporate action or constituent change occurring at time $t$, the divisor is reset from $D_{t^-}$ (just before) to $D_{t^+}$ (just after) so the index level is unchanged across the instant, $A_{t^-}/D_{t^-} = A_{t^+}/D_{t^+}$, giving

$$D_{t^+} = D_{t^-} \times \frac{A_{t^+}}{A_{t^-}}.$$

Whenever the event leaves $A$ itself unchanged ($A_{t^+} = A_{t^-}$), the formula returns $D_{t^+} = D_{t^-}$: no adjustment is needed, and none is made.

**Worked Example 8.2 — Divisor mechanics: a split, a buyback, and the Dow-style contrast**

Return to Worked Example 8.1's base index ($D_0 = 64$, $I_0 = 1{,}000.00$).

**(a) A split is divisor-neutral for a cap-weighted index.** Gamma executes a 3-for-1 split: its price becomes $\$10.00/3 = \$3.3333$ and its shares outstanding become $1{,}200 \times 3 = 3{,}600$m. Its market cap is $\$3.3333 \times 3{,}600\text{m} = \$12{,}000$m — identical to before, because a split multiplies $Q_i$ by exactly the factor it divides $P_i$ by, leaving $P_iQ_i$ untouched. The index total is therefore still \$64,000m and, by Result 8.3 with $A_{t^+} = A_{t^-}$, the divisor stays at 64 and $I = 64{,}000/64 = 1{,}000.00$: unchanged, with no adjustment made because none was required. This is the single most common point of confusion this book will correct: for a market-cap index, "the divisor adjusts for splits" is true only in the trivial sense that the adjustment it computes is zero.

**(b) A share-count change unrelated to price discovery does move the divisor.** Instead, Beta conducts an on-market buyback, retiring 40m shares at its prevailing \$25.00 price: its shares fall to 760m and its market cap to $25.00 \times 760 = \$19{,}000$m, a \$1,000m reduction with no change in Beta's quoted price. The index total falls to $20{,}000 + 19{,}000 + 12{,}000 + 12{,}000 = \$63{,}000$m. Left at the old divisor of 64, the index would read $63{,}000/64 = 984.375$ — a fictitious 1.5625% drop attributable entirely to a change in share count, not to a single price moving. Result 8.3 resets the divisor at the moment of the buyback to $D_1 = 64 \times (63{,}000/64{,}000) = 63.00$, restoring $I = 63{,}000/63 = 1{,}000.00$: continuous across the event, with the smaller divisor now the base for future price moves.

**(c) The Dow-style contrast.** Take the same four raw prices as a price-weighted index with base divisor $D'_0 = 4$ (the constituent count), giving $I'_0 = 135/4 = 33.75$. Now Delta — the highest-priced name, and hence the largest price-weight, from the contrast in Worked

Example 8.1 — executes an ordinary 2-for-1 split: its price falls to \$30.00 and the new price sum is $40 + 25 + 10 + 30 = \$105$. A price-weighted index has no shares term to offset the price change, so left at the old divisor the index would read $105/4 = 26.25$, a fictitious drop of 7.50 points, or 22.22% of the index's level, on a corporate action that transferred no wealth to or from anyone. Result 8.3 applies exactly as in (b), now with $A = \sum P_i$: $D_1' = 4 \times (105/135) = 3.1111$ (exactly 28/9), restoring $I' = 105/3.1111 = 33.75$. Every real-world Dow divisor update, across a century of splits, is this same one-line calculation — which is also why the DJIA's current divisor is a small fraction far below its original value of 1.

**Trap — The divisor's job depends on the weighting scheme**

Part (a) above is the trap in isolation: a systematic desk that treats "adjust the divisor for splits" as a universal rule, and applies it to a cap-weighted benchmark, will introduce exactly the error the divisor exists to prevent. The rule that is actually universal is Result 8.3 — reset the divisor whenever, and only when, the aggregate $A$ changes for a reason other than a constituent's own price discovery. What triggers that for a cap-weighted index (share issuance, buyback, constituent substitution, a scheduled float-factor update) is precisely what does *not* trigger it for a price-weighted one (which has no shares term to begin with, and for which an ordinary split is the main trigger). Knowing which family a published index belongs to is a precondition for reasoning about its divisor at all.

Three families the reader will meet by name, all float-adjusted cap-weighted and all structurally stable facts, cited here without a numeric constituent count that would itself drift: the **S&P/ASX 200** is S&P's float-adjusted, market-capitalisation-weighted index of the largest eligible ASX-listed companies, reconstituted on a quarterly review cycle by an index committee applying liquidity and eligibility screens, not a purely mechanical top-$n$-by-size rule; the **S&P 500** is the analogous large-capitalisation US benchmark, likewise committee-selected against liquidity, profitability and domicile criteria rather than assembled by a mechanical size cutoff alone; and the **Nasdaq-100** is a float-adjusted cap-weighted index of the hundred largest non-financial companies listed on Nasdaq. All three publish, and periodically revise, their own eligibility and weighting-cap methodologies; treat any version of this paragraph more specific than "float-adjusted, cap-weighted, committee-reviewed" as needing a current-methodology check.

**Index effects.** Because index-tracking funds must hold (or, for a synthetic tracker, reference) a benchmark's constituents in its published weights, a change to the index — a constituent addition or deletion — forces real trades that have nothing to do with any view on the company's fundamentals: it is a change in who is a mechanical buyer or seller, not a change in cash flows. Shleifer (1986) frames the resulting question sharply: if the demand curve for a stock is downward-sloping rather than perfectly elastic (the classical view, in which a stock has innumerable near-perfect substitutes and any one investor's demand shift moves the equilibrium quantity, not the price), then a purely mechanical, information-free shift in demand from newly mandated index-tracking buyers should move the price permanently, not just transiently. Harris and Gurel (1986) documented exactly this empirically around S&P 500 addition announcements: measurable price and volume effects clustered around the index-effective date, consistent with demand curves that do not, in practice, slope perfectly flat. For a systematic desk this is a genuine, published anomaly family, not folklore — and, exactly like the announcement-drift material in Section 8.5, it is a hypothesis to test out-of-sample against current data and current crowding, not a standing instruction to buy every announced addition.

**Rebalance flows.** The mechanical consequence of the above is concentrated, dateable trading volume: on an index's reconstitution effective date, every fund benchmarked to it must move its holdings to the new weights, and because tracking error is measured against the benchmark's own official closing print, that trading is disproportionately routed into the closing auction described in Section 8.2 — one more reason, alongside the auction's general price-discovery role, that index-reconstitution days produce some of the heaviest closing-auction volumes of the year on both the ASX and the major US venues.

## 8.4 Corporate Actions

Section 8.3 treated the divisor as the mechanism that keeps an *index* continuous through a corporate action. Exactly the same hazard exists one level down, at the level of a single stock's own price series, and there is no divisor sitting underneath an individual share to fix it automatically: a raw price history that does not correct for dividends, splits, rights issues and buybacks contains mechanical, predictable jumps that look, to any model or backtest that reads price alone, exactly like real returns.

**Dividends and the ex-date.** A company that declares a cash dividend fixes an ex-dividend date: a buyer of the share on or after that date does not receive the upcoming payment, so from that date the share is worth strictly less by (in a frictionless, no-other-news benchmark) the dividend itself. This is not a market view; it is an accounting identity about who is entitled to the cash.

**Worked Example 8.3 — Ex-dividend price mechanics and franking, Southern Bank Ltd**

Southern Bank Ltd (illustrative scenario) closes at \$28.40 cum-dividend and declares a fully franked dividend of \$0.85 per share. The theoretical ex-dividend price is

$$P_{\text{ex}} = P_{\text{cum}} - D = 28.40 - 0.85 = \$27.55.$$

Read naively as a price return, the ex-date print is $(27.55 - 28.40)/28.40 = -2.9930\%$ — a move larger than a typical day's volatility for a large-capitalisation bank stock, occurring on a day when, by construction, nothing informative need have happened at all. The *total* return, crediting the dividend against the price move, is $(27.55 + 0.85 - 28.40)/28.40 = 0.0000\%$: exactly flat, as it must be in this frictionless benchmark, because the ex-date mechanically reallocates value between the share price and a declared cash entitlement without destroying or creating any of it.

Fully franked is an Australian-specific attribute of the dividend, not of the price mechanics above: it certifies that the company has already paid corporate tax on the profit being distributed, and attaches a franking credit that a (typically resident) shareholder can offset against their own tax liability. For a dividend fully franked at the company tax rate $\tau$ (most large companies paid $\tau = 30\%$ through the mid-2020s, with a lower rate for smaller "base rate entities"; verify the rate current for the entity and year in question), the attached credit and the grossed-up amount are

$$FC = D \times \frac{\tau}{1-\tau}, \qquad \text{grossed-up} = D + FC = \frac{D}{1-\tau}.$$

At $\tau = 30\%$: $FC = 0.85 \times (0.30/0.70) = \$0.3643$, and the grossed-up dividend is $0.85 + 0.3643 = \$1.2143$ (check: $0.85/0.70 = \$1.2143$, matching).

**Definition — Franking credit**

A franking credit is a tax credit attached to an Australian dividend, certifying corporate tax already paid on the underlying profit, and is a structural feature of Australia's dividend imputation system, not a cash flow paid by the company on top of the dividend itself. Only the cash dividend $D$ moves through the price mechanics of Worked Example 8.3; the franking credit is a downstream tax attribute realised on the shareholder's own return, worth different amounts (up to a full refund, for some eligible resident holders) depending on the holder's own tax position, and worth nothing at all to a holder who cannot use Australian franking credits, such as most short sellers and most non-resident holders. Empirical "dividend drop-off" studies ask exactly how much of that clientele-dependent value shows up in the observed cum-to-ex price change; this book's mechanical baseline above is the frictionless benchmark that literature measures deviations from, not a claim that every ex-date print matches it to the cent.

**Splits and consolidations.** A $k$-for-1 split multiplies shares outstanding by $k$ and divides the price by $k$, leaving market value, and every shareholder's wealth, unchanged — the same $P \times Q$ invariance already used for the index in Worked Example 8.2(a), now applied to one stock's raw price series rather than to an index's aggregate. A raw, unadjusted price history nonetheless shows a $(k-1)/k$ overnight "drop" on the split date that a naive return calculation will read as a real, and typically enormous, one-day loss; a consolidation (reverse split, $k < 1$ in the same formula) produces the mirror-image fictitious jump.

**Result — Theoretical ex-rights price (TERP)**

A rights issue offers $M$ new shares at subscription price $P_{\text{sub}}$ for every $N$ shares already held. If every right is exercised, the post-issue price that leaves total shareholder value unchanged (existing value plus new cash raised, spread over the enlarged share count) is

$$\text{TERP} = \frac{N\,P_{\text{cum}} + M\,P_{\text{sub}}}{N+M},$$

and the theoretical value of one right is $P_{\text{cum}} - \text{TERP}$: the discount the rights issue mechanically builds into the stock, exactly analogous to the dividend drop above, and no more informative about the company's prospects on its own.

Buybacks run the dividend mechanic in reverse and add a share-count effect on top: an on-market buyback retires shares at the prevailing market price, shrinking shares outstanding with (in the frictionless benchmark) no immediate price effect of its own, the same mechanical share-count change already worked for the index divisor in Worked Example 8.2(b). An off-market buyback — an Australian-specific structure with no close US analogue — typically splits the buyback price into a smaller capital component and a much larger fully franked dividend component for tax reasons, a structural feature worth naming precisely because it means an off-market buyback is, economically, mostly a dividend event wearing a buyback's name; the exact capital/dividend split each issuer sets is company- and offer-specific and must be read from the buyback documentation, not assumed.

**Trap — Unadjusted series are not a rounding error**

Every mechanic in this section produces a price-series discontinuity that is real, mechanical, and utterly uninformative about the company: a $-2.9930\%$ ex-dividend print, a $-50\%$ two-for-one split print, a rights-issue discount, a buyback-driven share-count shift. Feed

any of these into a model or a backtest as if they were ordinary returns and the damage is not random noise — it is systematic, dated exactly to the corporate-action calendar, and large enough on its own to dominate a volatility estimate, trip a stop-loss, or manufacture an entirely fictitious "signal" that fires every reporting season. This is a specific instance of the Course Book's leakage taxonomy under survivorship and instrument selection [Course Book §18.1.6]: a universe or price history assembled without point-in-time corporate-action correction is not simply noisy, it is wrong in a way that correlates with exactly the calendar dates a systematic strategy is most likely to trade around. It is also, concretely, why the platform's v3 equity-pairs work is specified against corporate-action-adjusted series as a hard precondition, not a nice-to-have: a cointegration test or a spread-reversion signal run on raw prices is, in part, a test of whose dividend calendar looked busier last quarter.

## 8.5 The Information Cycle

Listed companies release scheduled, market-moving information on a predictable calendar — a structural fact worth naming even though its exact windows drift by market and by year. US-listed companies report quarterly, with results clustering in dense weeks beginning three to five weeks after each quarter-end; ASX-listed companies report half-yearly and full-yearly, clustering into the February and August "reporting seasons" that any Australia-based systematic desk learns to route around. Outside the scheduled print, continuous-disclosure obligations in both jurisdictions require a listed company to release material price-sensitive information as soon as it is aware of it, which is exactly why the company-requested trading halt introduced in Section 8.2 exists: a company anticipating a material announcement can and typically does request a halt to stop trading on an information asymmetry it is in the process of correcting, rather than let the market trade on rumour until the release is ready.

**Announcement drift.** The efficient-markets programme that gives concepts like "the price already reflects available information" a testable form (Fama, 1970) also produced its own most durable catalogue of counterexamples, documented departures from strict post-announcement efficiency that are named, studied and re-tested rather than assumed away. Post-earnings-announcement drift — a tendency for a stock's return to continue, for some weeks after an earnings surprise, in the same direction as the initial reaction, rather than jump once and stop — is the best-known member of that family; the weekend effect documented by French (1980), a calendar-day pattern in average returns with no obvious information-arrival story behind it, is another, unrelated example from the same broad literature of tested and re-tested anomalies. The right posture toward both, and toward every other named anomaly this book will meet, is the one this series applies throughout: these are hypotheses with a citation and a documented history, not standing trading rules, and whatever edge any of them once carried is a live empirical question against current data, current costs and current crowding — test, do not trust.

**Short selling and borrow.** A short sale — selling a share the seller does not own, to be bought back later — requires, mechanically, a real security borrowed from somewhere, and that borrow has genuine economics a long position never faces.

**Definition — Locate, borrow fee, and days-to-cover**

Before executing a short sale, a broker must have a *locate*: a reasonable basis for believing the shares can be borrowed and delivered by settlement, a regulatory requirement in both the US (under Reg SHO) and Australia (under ASIC's securities-lending and short-sale reporting regime), with the exact rule text and disclosure thresholds in each jurisdiction

subject to periodic revision and best verified current. Once located, shares are borrowed — typically via a securities-lending market intermediated by custodians and prime brokers, against posted collateral — for an annualised *borrow fee*: a few tens of basis points a year for an easily borrowed, high-float name ("general collateral"), rising to high single digits or more for a "hard-to-borrow" or "special" name where demand to short outstrips available lendable supply. *Days-to-cover* (aggregate short interest divided by average daily volume) is the standard published proxy for how many trading days it would take the market's short base to close out simultaneously, and hence for how disorderly a rush to cover could become.

The lender can recall borrowed shares at any time, forcing the borrower to return them by buying in the market, closing the short outright, or arranging a substitute loan — and, separately from the per-day fee, a short seller who is short over a dividend's ex-date does not receive the dividend paid by the company; instead the short seller must pay a *manufactured dividend* of the same cash amount to the lender of the shares, compensating the lender for the payment the borrowed shares would otherwise have earned. For a CFD-referenced short of the kind the platform actually holds, the provider typically collapses locate, borrow fee and any short rebate into a single daily financing rate charged on the short leg — the same "notional $\times$ rate, per day" structure Chapter 3 used for CFD financing generally, just usually running in a less favourable direction and at a materially higher rate for a hard-to-borrow name than for an easily financed long.

**Worked Example 8.4 — Short position financing cost, Ridgeline Mining Ltd**

A trader sells short 10,000 shares of Ridgeline Mining Ltd (illustrative scenario) at \$18.50, a notional of $10{,}000 \times 18.50 = \$185{,}000.00$. The CFD provider charges an annualised financing/borrow rate of 3.25% on the short leg, and the position is held 45 days on an Actual/365 basis. The financing cost is

$$\text{cost} = \text{notional} \times \text{rate} \times \frac{\text{days}}{365} = 185{,}000 \times 0.0325 \times \frac{45}{365} = \$741.27,$$

which is $741.27/185{,}000 = 0.4007\%$ of notional over the 45-day holding period — recovering the annual rate exactly once re-annualised, $0.4007\% \times (365/45) = 3.25\%$, confirming the arithmetic. This cost accrues whether or not the short thesis is ever right; it is a certain cash outflow set against an uncertain price outcome.

**Trap — Recalls arrive exactly when you need the borrow most**

A name becomes hard to borrow, and its fee spikes, precisely when short demand is heaviest relative to lendable supply — which is disproportionately likely to be exactly when the stock is rallying against the short (squeezing the shorts is what draws more shorting demand in the first place) and exactly when an existing lender is most tempted to recall shares to sell them outright rather than keep earning a lending fee. The moment a short position most needs patient, cheap financing to ride out an adverse move is systematically correlated with the moment financing becomes expensive or disappears altogether. A backtest that prices the borrow leg at today's easy-to-borrow, general- collateral rate and holds that rate fixed for the life of a simulated short is not being conservative by a little; it is assuming away the specific tail risk — fee spikes and forced buy-ins clustering with adverse price moves — that most differentiates shorting from a mirrored long, exactly the asymmetry the platform's pairs work [Course Book §21.4] must budget for on its short leg.

**Your project — Your project — v3's short leg is not a costless mirror of its long leg**

The Course Book's pairs-trading treatment [Course Book §21.4] derives entry and exit thresholds from a spread's estimated mean-reversion dynamics, implicitly pricing both legs of a spread trade symmetrically. This chapter's arithmetic is the reminder that the short leg of any v3 equity pair is a real borrowed security (or a CFD referencing one) with its own locate requirement, its own borrow fee that can move independently of the statistical spread the model is trading, and its own recall risk that can force an exit unrelated to the spread's level. A pair-mining process that screens candidates on cointegration statistics alone, without a borrowability and cost screen on the short leg, will pass pairs that are statistically attractive and operationally untradeable — the 4.4384% over 90 days worked in Exercise 8.3 below is exactly the size of number that needs to sit inside the leg-screening checklist, not be discovered after the position is on.

## 8.6 Sectors and Factors

**Definition — GICS**

The Global Industry Classification Standard (GICS), maintained by S&P and MSCI and used across most equity index providers and screening platforms, assigns every listed company to exactly one sub-industry within a four-level hierarchy: eleven sectors, each split into industry groups, each split into industries, each split into sub-industries, classified by principal business activity. The hierarchy itself is periodically revised (a 2018 revision, for instance, created a Communication Services sector out of parts of what had been Telecommunication Services and Consumer Discretionary) — stable enough to state as structural fact, but with its exact current sector count and boundaries worth a check against the current standard rather than assumed frozen.

Beneath sector classification sits a second, cross-cutting vocabulary this book's reader will meet constantly in factor and quant literature: a small set of firm characteristics used to sort the entire cross-section of stocks into groups whose average returns are then compared. *Size* sorts by market capitalisation; *value* sorts by cheapness relative to a fundamental anchor (book value, earnings, cash flow) against "growth" at the expensive end; *momentum* sorts by a stock's own relative return over a recent formation window (commonly six to twelve months) as a predictor of near-term continuation; *quality* sorts by a profitability and balance-sheet- strength composite (return on equity, leverage, earnings stability); and *low-volatility* sorts by trailing realised volatility or market beta. Each is, at bottom, nothing more exotic than a rule for grouping stocks along one cross-sectional dimension and measuring whether the groups' average returns differ — a description of cross-sectional *structure*, not a proof that any one grouping currently carries exploitable, cost-net edge. This book applies the same discipline to every one of them that Section 8.5 applied to announcement drift: each is a named, testable regularity with a documented history of both confirmation and, in several cases (momentum crashes being the sharpest), spectacular reversal, to be measured out-of-sample against current data rather than assumed permanent.

**Your project — Your project — sector and factor structure as v3's economic pre-filter**

The Course Book's pair-mining factory [Course Book §21.6] runs cointegration and stationarity tests across a large candidate universe, and its own discipline warns that testing enough pairs will find statistically significant relationships by chance alone. Sector and factor structure is the *economic* pre-filter that discipline needs before the statistical filter runs: two stocks in the same GICS sub-industry, or matched on similar size, value and quality characteristics, share

common exposures (input costs, regulation, the same macro and sector shocks) for a reason a human can state in one sentence, which is exactly the kind of prior a purely statistical cointegration search cannot supply on its own. A pair that clears cointegration testing *and* has an economic story — same sub-industry, similar factor loadings — is a materially safer candidate for v3's pair universe than one that clears the statistics with no economic rationale at all, which is far more likely to be exactly the kind of mined coincidence the leak-injection and multiple-testing discipline built earlier in this library exists to catch.

## 8.7 Exercises

**Exercise 8.1 — A four-stock index drill: weights, a move, a split, and an issuance**

Four hypothetical companies, W, X, Y, Z, have the following prices and shares outstanding.

| Stock | Price ($) | Shares (m) |
|---|---|---|
| W | 50 | 300 |
| X | 20 | 900 |
| Y | 15 | 1,000 |
| Z | 80 | 150 |

(a) Using a base index value of 1,000, compute the divisor, the index level, and each stock's cap-weight.

(b) Starting fresh from this base index, X rises 6% with the other three prices unchanged. Compute the new total market cap, the new index level, and the index return; then check your answer using Result 8.3.

(c) Starting fresh from the base index again, Z executes a 4-for-1 split. Does the divisor need to change? Compute Z's new price, new shares outstanding, new market cap, and the resulting index level.

(d) Starting fresh from the base index once more, Y issues 200m new shares in a placement at the prevailing $15.00 price. Compute the naive index level if the divisor is left unchanged, the size of the resulting fictitious jump, and the corrected divisor and index level.

▷ working space

**Solution — Exercise 8.1**

**(a)** Market caps: W $= 50 \times 300 =$ \$15,000m, X $= 20 \times 900 =$ \$18,000m, Y $= 15 \times 1{,}000 =$ \$15,000m, Z $= 80 \times 150 =$ \$12,000m; total = \$60,000m. Divisor $D_0 = 60{,}000/1{,}000 = 60$, so $I_0 = 60{,}000/60 = 1{,}000.00$. Cap-weights: W $= 15{,}000/60{,}000 = 25.00\%$, X $= 18{,}000/60{,}000 = 30.00\%$, Y $= 15{,}000/60{,}000 = 25.00\%$, Z $= 12{,}000/60{,}000 = 20.00\%$ (summing, as required, to 100.00%).

**(b)** X's new price is $20 \times 1.06 =$ \$21.20, so its new cap is $21.20 \times 900 =$ \$19,080m. The new total is $15{,}000 + 19{,}080 + 15{,}000 + 12{,}000 =$ \$61,080m, and the new index is $61{,}080/60 = 1{,}018.00$, an index return of 1.80%. Check via Result 8.3: only X moved, so $R^I = w_X \times r_X = 30.00\% \times 6\% = 1.80\%$ — matches exactly.

**(c)** No: a split leaves $P_i Q_i$ unchanged for the stock that splits, so by Result 8.3 the divisor does not move. Z's new price is $80/4 =$ \$20.00, new shares $150 \times 4 = 600$m, new cap $20.00 \times 600 =$ \$12,000m — identical to before the split. The index total is still \$60,000m and, at the unchanged divisor of 60, the index is still $60{,}000/60 = 1{,}000.00$.

**(d)** Y's new shares are $1{,}000 + 200 = 1{,}200$m, so its new cap is $15.00 \times 1{,}200 =$ \$18,000m (a \$3,000m increase driven entirely by the extra shares, at an unchanged price). The new total is $15{,}000 + 18{,}000 + 18{,}000 + 12{,}000 =$ \$63,000m. Left at the old divisor of 60, the naive index reads $63{,}000/60 = 1{,}050.00$, a fictitious +5.00% jump on a placement that changed nothing about any stock's price. The corrected divisor is $D_1 = 60 \times (63{,}000/60{,}000) = 63.00$, restoring $I = 63{,}000/63 = 1{,}000.00$: unchanged across the placement, exactly as Result 8.3 requires.

**Exercise 8.2 — Ex-date P&L on both sides of a pair: long, short, and the raw-series illusion**

Northern Retail Ltd (illustrative scenario) closes cum-dividend at \$15.20 and declares a \$0.42 dividend, franked at 60% (company tax rate $\tau = 30\%$).

(a) Compute the theoretical ex-dividend price, and a long position of 10,000 shares' price P&L, dividend received, and net P&L across the ex-date.

(b) Compute a short position of 5,000 shares' price P&L, manufactured dividend paid, and net P&L across the same ex-date.

(c) A raw backtest system reads only price changes, with no dividend adjustment. What P&L would it report for each side, and what systematic bias does this introduce for a strategy that runs both long and short equity legs (such as a pairs trade)?

(d) Compute the per-share franking credit and grossed-up dividend at the stated 60% franking level.

▷ working space

**Solution — Exercise 8.2**

**(a)** Ex price $= 15.20 - 0.42 = \$14.78$. Long 10,000 shares: price P&L $= (14.78 - 15.20) \times 10{,}000 = -\$4{,}200.00$; dividend received $= 0.42 \times 10{,}000 = \$4{,}200.00$; net $= \$0.00$. The mechanical price drop and the cash dividend offset exactly, as Worked Example 8.3 showed for the index-free single-stock case.

**(b)** Short 5,000 shares: price P&L $= (15.20 - 14.78) \times 5{,}000 = \$2{,}100.00$ (a short gains when price falls); manufactured dividend paid to the share lender $= 0.42 \times 5{,}000 = \$2{,}100.00$. Net $= 2{,}100.00 - 2{,}100.00 = \$0.00$ — the short's mechanical price gain and its dividend obligation also offset exactly, independent of franking (a manufactured dividend payment carries no franking credit to the payer).

**(c)** Reading price alone, the raw system reports $-\$4{,}200.00$ for the long side and $+\$2{,}100.00$ for the short side. This is not symmetric noise: it makes every long position look worse, and every short position look better, than reality by exactly the dividend on every name that goes ex-dividend during the backtest window. For a strategy running both legs of a pair, this is a directional bias baked into the sign of the position, not a random error that cancels across many trades — it will systematically overstate the profitability of short legs and understate the profitability of long legs, precisely the kind of leakage Section 8.4's trapbox [Course Book §18.1.6] warns against.

**(d)** At full franking, the credit per share would be $0.42 \times (0.30/0.70) = \$0.18$ (using $0.42 \times 3/7 = \$0.18$ exactly). At 60% franking, the credit is $0.60 \times 0.18 = \$0.108$ and the grossed-up dividend is $0.42 + 0.108 = \$0.528$ per share.

**Exercise 8.3 — Borrow economics of a 90-day hard-to-borrow short**

Redline Retail Ltd (illustrative scenario), a heavily shorted, low-float stock, has gone "special." A trader sells short 20,000 shares at \$6.40, planning to hold for 90 days on an Actual/365 basis, at an annualised hard-to-borrow fee of 18.00%. Independently, the trader's pairs model estimates a gross (pre-cost) reversion profit over a comparable 90-day window of 4.00% of notional.

(a) Compute the notional, the dollar borrow cost, and the borrow cost as a percentage of notional.

(b) Compute the assumed gross profit in dollars, and the net result after subtracting the borrow cost alone (before any spread, commission, or slippage cost).

(c) Does the borrow fee alone exceed the strategy's entire assumed gross edge? What does this imply for whether the trade should be taken, independent of whether the statistical thesis is correct?

▷ working space

**Solution — Exercise 8.3**

**(a)** Notional $= 20{,}000 \times 6.40 = \$128{,}000.00$. Borrow cost $= 128{,}000 \times 0.18 \times (90/365) = \$5{,}681.10$, which is $5{,}681.10/128{,}000 = 4.4384\%$ of notional (check: $18.00\% \times 90/365 = 4.4384\%$, matching).
**(b)** Assumed gross profit $= 0.04 \times 128{,}000 = \$5{,}120.00$. Net after the borrow fee alone $= 5{,}120.00 - 5{,}681.10 = -\$561.10$: a loss, from financing alone, before a single dollar of spread, commission or slippage is deducted.
**(c)** Yes: the borrow cost, 4.4384% of notional, exceeds the assumed 4.00% gross edge by 0.4384 percentage points. The trade is unviable purely on financing grounds — this is exactly "the fee that kills the trade": no improvement in execution quality or entry timing can rescue a position whose financing cost alone already exceeds its entire expected gross profit, which is why a borrow-cost screen belongs *before* capital is committed, not as a post-mortem line item, as the projbox in Section 8.5 argued for v3's leg-screening process.

**Exercise 8.4 — Classify six corporate actions by backtest hazard**

For each corporate action below, state (i) whether a raw, unadjusted price series shows a spurious jump on the event date, (ii) the mechanical driver if so, and (iii) the fix: which adjustment factor must be applied, and to which historical prices, to remove it.

(1) A 2-for-1 stock split.
(2) An ordinary, fully franked cash dividend.
(3) A 1-for-5 rights issue at a subscription price of \$8.00 against a \$10.00 cum-rights price.
(4) An on-market buyback.
(5) An off-market buyback structured with a large fully franked dividend component.
(6) Addition to the S&P/ASX 200 at the next quarterly reconstitution.

▷ working space

**Solution — Exercise 8.4**

| Action | Spurious jump? | Mechanical driver | Fix |
|---|---|---|---|
| (1) Split | Yes | Price divides, shares multiply by $k$; raw series shows a $-(k-1)/k$ "return" | Multiply all pre-split historical prices by $1/k$ (a split-adjustment factor), so the series is continuous in total value |
| (2) Ordinary dividend | Yes | Price falls by $\approx D$ on the ex-date | Add back the dividend to all prior prices (a total-return adjustment factor), or track total return directly |
| (3) Rights issue | Yes | Price falls to TERP $= (5 \times 10.00 + 1 \times 8.00)/6 =$ \$9.6667, a \$0.3333 mechanical discount | Apply a rights-adjustment factor (analogous to the dividend factor, scaled by the TERP-to-cum-price ratio) to prior prices |
| (4) On-market buyback | No price jump; share-count changes | Shares retired at the prevailing price, no price effect in the frictionless benchmark | No price adjustment; only shares-outstanding (market-cap) series needs updating |
| (5) Off-market buyback | Yes, on the dividend portion | Economically mostly a dividend event (large franked component) wearing a buyback's name | Treat the dividend-equivalent portion with the same total-return adjustment as (2); treat the capital portion as a partial redemption |
| (6) Index addition | No adjustment to the *stock's* own series; the *index's* constituent list and divisor change | Mechanical, demand-driven price/volume effects clustering at the effective date (Shleifer, 1986; Harris-Gurel, 1986), not a price-series discontinuity to correct | No stock-price fix; the index divisor is reset per Result 8.3, and the event is flagged as a demand-driven anomaly window, not detrended away |

The pattern across (1)–(3) and the dividend leg of (5) is the same one this chapter has used throughout: each is a computable, mechanical price change with a known formula, correctable by a multiplicative adjustment factor applied to history. (4) is the reminder that not every corporate action moves price at all — some only move the share count the index divisor (or a market-cap series) must track. (6) is different in kind again: nothing about the stock's own price mechanics requires correction, but the event is a genuine, documented source of demand-driven return that a systematic desk should recognise and test for, not confuse with either of the other two categories.

**Exercise 8.5 — Essay: why an index-CFD trader still needs this chapter**

The platform trades index CFDs, not individual equities. Explain, with reference to at least dividends, rebalancing, and trading halts, why a trader who only ever holds index-level exposure still needs to understand single-stock corporate-action and market-structure mechanics. Use the $-0.0658\%$ figure below as part of your answer.

▷ working space

**Solution — Exercise 8.5**

An index CFD's price is derivative, ultimately, of the cash index level (Chapter 9 derives the fair-value chain in full), and the cash index is nothing but the weighted aggregate of its constituents' own prices (Section 8.3). Every mechanic this chapter worked at the single-stock level therefore propagates upward, scaled by weight, into the instrument the platform actually holds. Take Southern Bank Ltd from Worked Example 8.3 again, now as one constituent of a broad index at an assumed weight of 2.20%: its ex-dividend drop contributes an index-level price return of $-w \times (d/P) = -0.0220 \times (0.85/28.40) = -0.0658\%$ on that one name alone, on a day when, mechanically, nothing informative need have happened to the bank at all. Multiply that by dozens of constituents going ex-dividend through a reporting season and the cumulative mechanical drag on a *price* (not total-return) index over a quarter is systematic and measurable — exactly the raw-index-versus-total-return confusion Chapter 9's trapbox singles out, because an index CFD priced off a price index inherits that drag directly, and a naive backtest that reads it as alpha or as a risk signal is repeating this chapter's central error one level up the aggregation chain. Rebalancing flows (Section 8.3) matter for the same reason: a reconstitution changes what the index *is*, discontinuously, on the effective date, and an index CFD tracks whatever the index becomes, not a fixed basket frozen at the position's entry date. And a trading halt in even one heavily weighted constituent (Section 8.2) can freeze part of the index's own price discovery while the index CFD, quoting far closer to continuously against the futures chain, must decide how to behave regardless. None of this requires trading a single share directly; all of it requires understanding, as this chapter has, exactly how a share's price, a dividend, a split and a halt turn into the number printed on the one instrument the platform is actually long or short.

**Key idea — What to remember**

A share is a residual, limited-liability claim traded across two structurally distinct layers — primary issuance and secondary exchange — and every venue this chapter examined, from the deep, fragmented US market to the reader's own ASX, is the continuous-auction and call-auction machinery of Chapter 2 wearing a listing regime, a settlement cycle and a halt policy that drift and must be re-verified, not memorised, exactly as the US's 2024 move to T+1 settlement demonstrated live. An index is a formula, not an observation: its divisor exists to keep the published level continuous through corporate actions, and this chapter's arithmetic showed precisely which of those actions — a split, for a cap-weighted benchmark — leave it untouched, and which ones — a buyback, an issuance, a reconstitution — force it to move, a distinction with no partial credit for guessing wrong. The same mechanics, one level down, turn an unadjusted single-stock price series into a source of large, dated, entirely uninformative jumps at every dividend, split and rights issue, precisely the leakage the Course Book's survivorship discipline warns against and precisely what v3's adjusted-series rule exists to prevent. Short selling adds a second, genuinely asymmetric layer of cost on top of price risk — a locate, a borrow fee that tends to spike exactly when a position is under the most pressure, and a recall risk with no long-side analogue — all of which must clear a pair's expected edge before capital is committed, not after a fee notice arrives; sector and factor structure is this chapter's answer to what turns a statistically cointegrated pair into an economically credible one. Chapter 9 carries this machinery one layer up, to the instrument the platform actually holds: the index CFD, its fair-value basis against the futures chain it references, and the dividend and session-gap mechanics that make raw-versus-total-return index confusion a live hazard at exactly the layer this chapter has just made precise.

# 9

# Index CFDs and Trading Single Names

*Chapter 8 built the share and the index from the venue upward: how a price is discovered, how a cap-weighted level and its divisor are computed, and why a dividend or a split rewrites the constituent series in a mechanically predictable way. None of that arithmetic is itself a tradeable instrument. This chapter follows the index one further step, from the cash level Chapter 8 computed, through the futures contract where the index's forward value is actually discovered, to the contract-for-difference a retail account can hold — the platform's own instrument, and the wrapper through which every mechanic derived below reaches an account statement. The chain has three links, and each one changes what a backtest must model: the cash index does not trade, so nothing in Chapter 8's construction arithmetic is directly executable; the future prices a cost of carry that a naive spot-tracking model omits entirely; and the CFD inherits that carry while adding a session structure — near-continuous quoting against a cash market that opens and closes — which manufactures a genuine, attributable gap at every local open. The chapter closes by returning to single names rather than baskets: what a share must clear, individually, before it is allowed to be one leg of a version-3 pair, and the full cost stack — spread, commission, financing, borrow — a pairs round trip actually pays once every line is added, not only the spread the Course Book's simplified toll already counted.*

## 9.1 The Index Exposure Chain

A cash index is a number: Chapter 8 derived it as a divisor-adjusted, weighted sum of constituent prices, recomputed every time any constituent trades. It is not a security. Nobody holds it, nobody delivers it, and no order book matches buyers and sellers of "the index" directly — every route to an index-shaped exposure is synthetic, built from something else that does trade. Three such routes matter to this book, layered in a fixed order.

**Definition — The index exposure chain**

**Cash index ($S$).** The constituent-weighted level itself, computed continuously while the underlying constituent market is open and static (producing no new print) while it is closed. Not a tradeable instrument.

**Index future.** An exchange-traded, centrally cleared, margined contract on the index's value at a fixed expiry $T$, quoted nearly continuously — commonly close to 24 hours across the trading week on the major contracts — independent of the constituent market's own session. This is where the index's *forward* value is actually discovered.

**Index CFD.** An over-the-counter contract for difference between a provider and a client, struck against the future's price (directly, or via a constructed feed) rather than against the cash index directly, requiring no delivery and no clearing house, and margined bilaterally against the provider. This is the retail-accessible wrapper, and the platform's own instrument.

The middle link exists because the first cannot be traded and a genuine forward market in the index is economically useful — hedgers and speculators need forward exposure, not only

spot — and the third link exists because a futures contract's minimum size, margining regime and exchange membership requirements put it out of reach of most retail accounts. Each step repackages the same underlying economic claim, and each repackaging is priced, not assumed identical to the step before it: a future is not simply "the index, dated," and a CFD is not simply "the future, retail-sized." The relationship between the first two links is exact and derivable; the relationship between the second two is structural and provider-specific.

### Fair value: the cost-of-carry relationship

The future's price is pinned to the cash index today by a standard no-arbitrage cost-of-carry relationship (Hull, 2021), and the derivation is short enough to give in full.

**Result — Index fair value: the cost-of-carry relationship**

Let $S$ be the cash index today, $F$ the price of a future expiring in $T$ years, $r$ the simple annual rate at which cash can be borrowed or lent over $[0, T]$, and $q$ the constituent basket's dividend yield, paid over $[0, T]$ and approximated, for a tenor of a few months or less, as a simple-interest flow worth $qST$ in aggregate. Two ways of holding a claim on $S_T$ must cost the same today, or an arbitrageur trades one against the other for a riskless profit. Buying one future costs nothing to enter and pays $S_T - F$ at $T$. Buying the basket for $S$ today, financed by borrowing $S$ at $r$ and collecting its dividends along the way, pays

$$S_T + qST - S(1 + rT) \;=\; S_T - S\big(1 + (r - q)T\big)$$

at $T$: the basket's terminal value, plus dividends collected, less the loan repaid. Both routes deliver $S_T$ minus a fixed amount fixed today; if the two fixed amounts differed, selling the dearer claim and buying the cheaper one is a riskless profit locked in today for zero net cost, so no-arbitrage forces them equal:

$$F \;=\; S\big(1 + (r - q)T\big), \qquad \text{basis} \;:=\; F - S \;=\; S(r - q)T. \tag{9.1}$$

Compounding continuously instead of simply gives the equivalent $F = Se^{(r-q)T}$, used throughout derivatives pricing wherever a continuously paid yield stands in for a foreign rate or a storage cost; the two forms agree to a fraction of an index point at the tenors and rate gaps typical of listed index futures (Worked Example 9.1), and nothing below turns on the choice.

**Worked Example 9.1 — Fair basis and the premium/discount reading**

Two hypothetical cash indices illustrate both directions of (9.1), on an ACT/365 day-count basis ($T = \text{days}/365$) used throughout this chapter.

**Part A — premium (contango).** A broad cap-weighted cash index stands at $S = 7{,}850.00$ points. Money-market cash yields $r = 4.35\%$ p.a., the basket's dividend yield is $q = 1.75\%$ p.a., and the nearest quarterly future expires in 47 days, so $T = 47/365 = 0.128767$ years. Then

$$(r - q)T = (0.0435 - 0.0175) \times 0.128767 = 0.00334795,$$

$$F = 7{,}850.00 \times (1 + 0.00334795) = \mathbf{7{,}876.28} \text{ points},$$

$$\text{basis} = F - S = +\mathbf{26.28} \text{ points}.$$

The future trades *above* cash: a positive basis, or contango, is the normal state whenever the financing rate exceeds the dividend yield, the common case for a low-yield, high-nominal-rate benchmark. (The continuous form gives $F = 7{,}876.33$, a difference of 0.04 points.)

**Part B — discount (backwardation).** A second, higher-yielding cash index stands at $S = 8{,}120.00$ points, with $r = 4.10\%$, a materially higher dividend yield $q = 5.60\%$, and $T = 62/365 = 0.169863$ years to expiry. Now $r < q$:

$$(r - q)T = (0.0410 - 0.0560) \times 0.169863 = -0.00254795,$$

$$F = 8{,}120.00 \times (1 - 0.00254795) = \mathbf{8{,}099.31} \text{ points},$$

$$\text{basis} = -\mathbf{20.69} \text{ points}.$$

The future trades *below* cash. This is not a malfunction: a high-yield index financed at a comparatively low cash rate makes carrying the physical basket forward more expensive, in dividends foregone, than the interest saved by not paying for it today, and the future prices that in. Reading the sign of $r - q$ is the whole exercise; the direction of the basis is never itself a forecast of where the cash index is going.

**Part C — rich or cheap?** Suppose Part A's future is actually quoted in the market at $F_{\text{mkt}} = 7{,}878.00$, against the fair value $F = 7{,}876.28$ just computed. The deviation is

$$F_{\text{mkt}} - F = 7{,}878.00 - 7{,}876.28 = +\mathbf{1.72} \text{ points } \ (0.02\% \text{ of } S),$$

so the future is trading *rich* to fair value: cash-and-carry (sell the future, buy the basket, finance it to expiry) is the theoretical arbitrage direction. At 0.02% of the index this is unexploitable by anyone who has to buy the actual constituent basket to do it — transaction costs and financing terms on a many-name basket dwarf a deviation this size — and it should be read as noise around fair value, not a signal. A deviation an order of magnitude larger, at quarter-end or around an unusual dividend-timing event, is the signature of genuine funding stress rather than a trading opportunity a retail CFD account can act on directly.

### From future to CFD: the provider's construction

The CFD link in the chain is not pinned down by no-arbitrage the way (9.1) pins down the future, because a bilateral OTC contract has no delivery mechanism forcing convergence and no exchange enforcing a single price. Providers describe, in broad terms, essentially two constructions. The first prices the CFD directly off the front future's continuous feed and subtracts a *fair-value adjustment* equal to the remaining basis, $S(r - q)(T - t)$ at time $t$, so the quoted CFD price tracks the cash index rather than the future's premium or discount, and the adjustment shrinks linearly to zero as $t \to T$ — consistent with $F_T = S_T$ at expiry, which (9.1) already implies by setting the remaining tenor to zero. The second prices the CFD at an estimate of the cash index level directly and charges or credits financing separately as an explicit daily line: functionally equivalent to the first at the end of any one day, but operationally different intraday, since the first construction can show a small adjustment-driven jump at each futures roll while the second shows a smoother tick-by-tick path referenced to the cash proxy. Which construction a given provider uses for a given instrument, and the exact benchmark rate and administrative spread layered on top of it (Section 9.5), are provider-specific facts to verify against current documentation, not structural facts this book can state once and rely on.

## 9.2 Dividends in Index CFDs

Chapter 8 derived the ex-dividend mechanics of a single stock: on the ex-date, the price drops by approximately the dividend, because a share bought from that date no longer carries the right to the payment a share bought the day before did. A cap-weighted index is a sum of constituent prices divided by a divisor, so every constituent's ex-date drop passes straight through, weighted

by that constituent's index weight, into the index level itself. Across a market's worth of staggered ex-dates this is not an occasional event but a near-continuous drag on the printed level — small on most days, and additive to an economically meaningful amount over a dividend-heavy week or quarter.

**Definition — Price index vs. total-return index**

A *price* (or "clean") index reinvests nothing: every ex-date drop is a permanent reduction in the printed level relative to where it would otherwise have stood, so a price index's long-run return understates the return actually available to a holder of the underlying basket by approximately the accumulated dividend yield. A *total-return* index reinvests each dividend back into the index on the ex-date, so the identical capital-only price moves are present but no permanent ex-date drop appears: the level itself already embeds every historical dividend. The two series diverge monotonically and, over years, substantially. An index CFD is priced off the price index, via the future of Section 9.1 — the object that actually trades and actually gaps on an ex-date — never off a total-return series, which is a research and benchmarking construct with no directly tradeable analogue.

A long index CFD is meant to replicate holding the constituent basket, and a basket holder receives the dividend in cash on the ex-date, so replication requires the long CFD to be credited the equivalent amount — otherwise a long CFD systematically underperforms the basket it claims to track by exactly the accumulated ex-date drag. The mirror argument prices the short side: a trader short the physical basket, via a securities loan, owes the dividend to the lender who supplied the stock, so a short CFD, synthetically equivalent to a short basket position, is debited the same amount it would owe if the position were physical.

**Result — Dividend adjustment on an index CFD**

On each ex-date, a long index CFD position is credited, and a short position debited, an amount equal to the index points the constituent's dividend contributes to the (price) index level, converted at the position's exposure per point. The credit and debit are equal in magnitude and opposite in sign; whether a provider pays the long side the full gross amount or applies a withholding-style haircut on indices with cross-border constituents is provider-specific and must be verified against current terms, but the structure — long credited, short debited, on the ex-date, sized to the index-point contribution — is the mechanism common to every retail index CFD provider, because it is simply what replicating the underlying basket requires.

**Worked Example 9.2 — A dividend-heavy week, long and short**

A broad-market index CFD is quoted at \$10 per index point. Over one week, three constituents go ex-dividend on the Monday (combined index-point contribution 1.35), one on the Tuesday (0.42), and one on the Thursday (0.68), for a total weekly dividend drag of

$$1.35 + 0.42 + 0.68 = 2.45 \text{ points.}$$

Suppose the index's raw print is unchanged over the week: it opens Monday and closes Friday both at 7,850.00. A backtest reading only the price index series over this week records a price move of zero and a P&L of zero for a position of either sign, at \$10/point:

$$\text{price P\&L} = (7{,}850.00 - 7{,}850.00) \times \$10 = \$0.00.$$

The live CFD ledger, however, applies the credit/debit rule above on each of the three

ex-dates regardless of what the index printed that week. The long position is credited

$$2.45 \times \$10 = \$24.50,$$

and the short position is debited the same \$24.50. Netting both effects,

$$\text{LONG net} = \$0.00 + \$24.50 = +\$\mathbf{24.50},$$

$$\text{SHORT net} = \$0.00 - \$24.50 = -\$\mathbf{24.50}.$$

The raw index said nothing happened. The account statement says otherwise on both sides of the trade, by an identical \$24.50 that a raw-price backtest never sees.

**Trap — Raw index levels are not a tradeable P&L series**

Two opposite backtesting failures share one root cause: treating an index series as if its levels alone determine a CFD position's return. First, simulating a strategy against a *price* index series while ignoring dividend adjustments entirely reproduces neither the CFD's real long P&L (understated, by the accumulated credits, exactly as in Worked Example 9.2) nor its real short P&L (overstated — a systematic short-index strategy looks better in the backtest than any live account of it would ever be, because the recurring debit never appears). Second, and just as common, simulating against a *total-return* series while *also* modelling explicit dividend adjustments on top of it double-counts: total-return levels already have every historical dividend reinvested into the price, so adding a credit on an ex-date the series has already absorbed pays for the same dividend twice. The only correct simulation uses price-index level changes for the mark-to-market leg and adds the credit/debit ledger entries explicitly, on the actual ex-dates — exactly mirroring what the live account statement does, because that is, quite literally, the only thing a live index CFD account statement can do.

## 9.3 Sessions and Gaps

Chapter 8's cash session is short and bounded: a few hours of continuous trading bracketed by opening and closing auctions, closed evenings, weekends and local holidays. The future (Section 9.1) that the index CFD is priced off keeps trading through almost all of that closed time, because it is a single global instrument referencing an index whose economic drivers — overnight news, other markets' sessions, currency moves — do not stop simply because the local cash exchange has. An index CFD therefore quotes a genuinely continuous price across a period when the thing it is meant to track is, in the most literal sense, not being priced at all: the cash index has no fresh print between the previous close and the next open, because no constituent has traded.

**Result — Decomposing the overnight gap**

Let $S_{\text{close}}$ be the previous cash close, $Q_{\text{pre}}$ the index CFD/future's quote in the moments just before today's cash open, and $S_{\text{open}}$ today's actual cash open print, from the local opening auction. The total open gap decomposes exactly as

$$\underbrace{S_{\text{open}} - S_{\text{close}}}_{\text{total}} = \underbrace{Q_{\text{pre}} - S_{\text{close}}}_{\text{overnight}} + \underbrace{S_{\text{open}} - Q_{\text{pre}}}_{\text{residual}}. \tag{9.2}$$

The overnight component is exactly what an account holding the CFD/future through the closed period already experienced, marked to market continuously as it accrued. The residual is information that crystallises only when the local market itself reopens — domestic order

imbalance, local news between the last CFD tick and the auction, or the auction mechanism's own price discovery — and it is unavailable to anyone, at any price, before the opening auction uncrosses.

**Worked Example 9.3 — Gap attribution at the cash open**

The previous cash close was $S_{\text{close}} = 7{,}905.20$. Overnight, the index CFD/future drifted on offshore developments, quoting $Q_{\text{pre}} = 7{,}922.65$ in the moments before today's local open. The actual opening auction then prints $S_{\text{open}} = 7{,}930.10$. By (9.2),

$$\text{total gap} = 7{,}930.10 - 7{,}905.20 = +\mathbf{24.90} \text{ points},$$

$$\text{overnight component} = 7{,}922.65 - 7{,}905.20 = +\mathbf{17.45} \text{ points} \ \ (70.08\% \text{ of the total gap}),$$

$$\text{residual component} = 24.90 - 17.45 = +\mathbf{7.45} \text{ points} \ \ (29.92\% \text{ of the total gap}).$$

An account that held the CFD long overnight had already marked 17.45 of the eventual 24.90-point move to market before the local exchange opened; only the remaining 7.45 points were genuinely new information at the open. A pure cash-market participant — unable to trade the overnight move at all — experiences the entire 24.90 points as a single discontinuity at the open. The CFD trader experiences most of the same move smoothly and a smaller jump at the open; the jump is real, but it is not the whole gap, and treating it as though it were overstates the risk actually concentrated in that one bar.

**Your project — Session buckets and gap handling for the platform's two index instruments**

The platform's two index CFDs need the FX cost surface's session-bucketing logic (Chapter 4's Wellington-to-NY liquidity clock, restated for equities) applied to a different calendar: bucket every bar by whether it falls inside the underlying cash session, in an overlapping-but-outside- cash-hours window (typically thinner, wider-spread, futures-led), or in the remaining low-liquidity hours, and fit the spread distribution separately per bucket rather than pooling a cash-session tick with a small-hours tick as though they were exchangeable observations. The cash-open bar itself needs a second, orthogonal flag, independent of which spread bucket it falls in: it is a structurally gap-prone bar by (9.2), in exactly the sense that a scheduled-news bar or a rollover bar is already flagged elsewhere on the platform, and the evaluation harness's calibration checks (coverage, PIT) should be run on it separately from ordinary intraday bars rather than pooled with them — a quantile band calibrated on the typical bar will not, and is not expected to, cover the open bar's fatter, gap-driven realised distribution at the same nominal rate.

## 9.4 Single Names for Pairs

Everything so far in this chapter has concerned an instrument referencing a basket. Version 3 trades single names, two at a time, and the Course Book built the statistics of that trade in full: cointegration, the Ornstein–Uhlenbeck spread, the Kalman-filtered hedge ratio, and the pair-mining factory's multiple-testing arithmetic. None of that is repeated here. What the Course Book's treatment does not cover, because it is not a statistical question, is whether a candidate name is fit to be a leg at all before a single cointegration test is run against it — and that is a prior, operational screen this book supplies.

**Liquidity.** A leg must be large enough, in average daily (dollar) volume (ADV), that entering and exiting the position the pair's own trade cadence requires does not itself move the price against it — the implementation-shortfall logic of Almgren and Chriss (2001) applied at the

individual-name level, before any pair-level economics are considered. A floor stated as a multiple of the largest single order the strategy would ever place in that name (so the position can be built or unwound inside a small fraction of one day's volume) is more useful than an absolute ADV threshold, because it scales with the strategy's own size rather than with the name's popularity.

**Spread.** [Course Book §21.4] priced a pair's round trip at four times a single leg's per-trade toll, $c = 4c_{\text{leg}}$, profitable only once the dispersion harvested between entry and exit exceeds it. For a liquid name $c_{\text{leg}}$ is dominated by half the quoted spread, so a name whose own half-spread alone already approaches, say, one-eighth of the pair's stationary dispersion $\hat{\sigma}_z$ is a poor candidate regardless of how convincingly it cointegrates with anything: the *other* leg's cost has not even been added yet, and half the margin for error implied by that relationship is already spent. Screening on spread before screening on cointegration discards the arithmetic failures cheaply, before the Course Book's more expensive statistical machinery ever runs.

**Borrowability.** Chapter 8 covered the mechanics of a locate, a borrow fee and recall risk for a single short position. A pairs book adds one twist: either leg can be the short one depending on the sign of the day's signal, so both names, not only the one currently short, need a history of reliable, persistently available borrow rather than a locate confirmed just for today. Recall risk is worse here than for an outright short, because a forced buy-in on one leg does not close the position — it strips the hedge and leaves a naked directional exposure in the other leg, at precisely the moment (the pair already diverged, margin already stressed) that a forced, unplanned trade is most costly.

**Corporate-action-adjusted series.** [Course Book §21.6]'s leak inventory already named the fix — continuous, point-in-time adjustment, with factors stored at their announcement timestamp rather than applied retroactively — for the pair-level backtest. At the name-screening stage the requirement is upstream and simpler to state: no candidate leg should carry a *pending*, announced-but-not-yet-effective corporate action inside the intended holding window. A pending demerger, tender or special dividend changes exactly the return-generating process the cointegration test assumed stable, on a known date, which makes it a screen-out — or at minimum a flagged, size-capped exception — rather than a risk discovered mid-trade.

**Halts policy.** If one leg halts, pending a material announcement or on an automated single-stock circuit breaker (Chapter 8), while the other keeps trading, the position is instantly a naked directional bet in whichever leg is still moving, with no ability to hedge or exit the halted side. This must be a pre-registered decision, not an improvisation made while the halt is live: hedge the halted leg synthetically with an index or sector proxy for the halt's duration, hold within a hard, pre-set size limit and accept the temporary directional exposure, or exit the tradeable leg outright and accept a temporarily unhedged residual. Any of the three is defensible; discovering which one your own book will do only after the halt has already started is not.

Table 9.1 collects the five screens into the operational gate a candidate name must clear before it is even offered to the cointegration test.

**Sizing the two legs.** [Course Book §21.4] settled this already and it is not re-derived here: size beta-neutral, one unit of the dependent leg against $\hat{\beta}$ units of the other, because beta-neutrality is the only sizing consistent with the cointegrating relation itself — any other ratio leaves a live, unbounded stochastic trend inside a position meant to be stationary. Dollar-neutral sizing is a separate risk convention, useful for controlling net market exposure, and it coincides with beta-neutral sizing only by coincidence of price levels. A book run beta-neutral inherits whatever residual factor exposure $\hat{\beta}$ does not itself cancel, and the platform's own index CFD (Sections 9.1–9.3) is a natural instrument for stripping that residual out separately, rather than distorting the pair's own sizing to chase a zero-net-exposure target the cointegrating vector was never trying to deliver.

**Where this checklist sits in the pipeline.** [Course Book §21.6]'s pipeline begins with an economic pre-filter over price data — same GICS sub-industry, dual-class pairs, ADR/ordinary pairs — and only then runs the cointegration test, because that ordering shrinks the trial count before

| Screen | What it tests | Illustrative threshold | Consequence if failed |
|---|---|---|---|
| ADV floor | Can the position be built/unwound without moving the price | $\geq$ a small multiple of the largest planned single order (indicative) | Excluded regardless of statistical fit |
| Spread cap | Whether $c_{\text{leg}}$ alone already threatens the breakeven margin | Half-spread a small fraction of typical $\hat{\sigma}_z$ (indicative) | Excluded, or re-tested only against slower, wider-dispersion pairs |
| Borrow | Locatable now *and* persistently, at a bounded fee, for either leg direction | Pre-registered maximum borrow fee (% p.a.), general collateral preferred | Excluded if fee cap breached even when cointegrated |
| Corporate actions | No pending, announced action inside the holding window | Point-in-time adjusted series mandatory; no live pending action | Excluded, or flagged and size-capped |
| Halts policy | Pre-registered response if one leg halts | One of: synthetic hedge, capped hold, or exit-tradeable-leg | Position management fixed before entry, not during a halt |

**Table 9.1.** The v3 leg-screening checklist: operational conditions a single name must clear before it is offered to the cointegration test of [Course Book §21.6]. Thresholds are illustrative and must be calibrated, and periodically re-verified, against the platform's own cost structure and borrow access.

any statistical power is spent. This chapter's checklist is a second, complementary filter at a different level: it screens *names*, not pairs, and it must run first, because a candidate pair inherits the worse of its two legs' liquidity, borrow and corporate-action standing regardless of how well the two prices move together. Running the economic and statistical filters of [Course Book §21.6] over a universe that has not first cleared this chapter's name-level screen wastes exactly the statistical power [Course Book §21.6] was built to conserve, on candidates that were never operationally tradeable in the first place.

## 9.5 Costs at the Index/Equity Layer

Three cost lines apply at this layer that the FX chapters priced differently, or not at all: the index CFD's own spread, now sessioned by the cash-hours structure of Section 9.3 rather than by the Wellington–NY clock; financing, still a carry rate exactly as in (9.1) but now the provider's own rate rather than a rate differential between two currencies; and, for the single-name legs of Section 9.4, an explicit commission alongside the spread, and a borrow fee whenever a leg is short.

**Index CFD spreads.** Spreads on the platform's two index instruments widen and narrow with the session structure Section 9.3 described: tightest while the underlying cash market is open and liquid, wider outside those hours when the quote is carried mostly by the future's own thinner order book, and prone to brief spikes around scheduled macro releases and at the daily rollover moment, mirroring the FX session-cost pattern already derived in Chapter 4. Exact current spread levels by session are provider- and instrument-specific and drift with market conditions; treat any specific figure as indicative, as of the mid-2020s, and verify current specifications before sizing a live trade against it.

**Financing.** (9.1)'s $r$ becomes, for a position actually held overnight rather than replicated theoretically, the provider's own financing rate: typically a short-term reference rate in the CFD's currency, plus an administrative spread on the long side and minus a usually smaller spread on the short side, applied per day held on the position's full notional — structurally the same construction

as the FX rollover swap and this chapter's Section 9.1, expressed in the index's own points or as a percentage of notional rather than in pips. The benchmark rate and the administrative spread on each side are provider-specific and must be verified current, not assumed.

**Single-name commissions and borrow.** Chapter 8 covered the borrow-fee mechanics; the structural addition at this layer is that single-name CFDs commonly carry an explicit, visible commission per trade alongside a comparatively tight spread, in contrast to the spread-only pricing typical of FX and index CFDs — one more line to itemise, not fold silently into an assumed all-in spread. Both the commission schedule and the borrow-fee curve by name are provider- and market-specific and, again, drift.

**Worked Example 9.4 — A 10-day pairs round trip, the full cost stack**

A beta-neutral pair: long 600 shares of $Y$ at \$84.50 (notional \$50,700.00), short 1,032 shares of $X$ at \$46.20 (notional \$47,678.40, $\hat{\beta} = 1{,}032/600 = 1.72$), held for 10 calendar days ($T = 10/365$).

**Commission.** At \$0.005 per share per side, a round trip (entry plus exit) costs \$0.01 per share:

$$\text{commission}_Y = 600 \times \$0.01 = \$6.00, \qquad \text{commission}_X = 1{,}032 \times \$0.01 = \$10.32.$$

**Spread.** One full spread paid per round trip: $Y$ quoted 6c wide, $X$ quoted 4c wide,

$$\text{spread}_Y = 600 \times \$0.06 = \$36.00, \qquad \text{spread}_X = 1{,}032 \times \$0.04 = \$41.28.$$

**Financing and borrow.** A reference rate of 4.35% p.a. plus a 2.50% administrative spread prices the long leg's financing at 6.85% p.a.; the short leg earns a financing credit of $4.35\% - 2.50\% = 1.85\%$ p.a. on its notional but pays a stock-borrow fee of 2.75% p.a. to the lender of $X$ shares — a mildly special name, not a hard-to-borrow one:

$$\text{financing}_Y = \$50{,}700.00 \times 0.0685 \times \tfrac{10}{365} = \$95.15,$$

$$\text{credit}_X = \$47{,}678.40 \times 0.0185 \times \tfrac{10}{365} = \$24.17,$$

$$\text{borrow}_X = \$47{,}678.40 \times 0.0275 \times \tfrac{10}{365} = \$35.92,$$

for a net financing-and-borrow cost on the short leg of \$11.76, computed from the unrounded components: the credit does not clear the borrow fee once the name is even mildly special, because the two accrue from different desks against different rates and only happen to net against each other on the account statement.

**The full stack.**

| | |
|---|---|
| Commission (both legs) | \$16.32 |
| Spread (both legs) | \$77.28 |
| Financing, leg $Y$ | \$95.15 |
| Financing credit, leg $X$ | −\$24.17 |
| Borrow fee, leg $X$ | \$35.92 |
| **Total cost** | **\$200.51** |

Rows are rounded to the cent; the total is computed from unrounded components and can differ from the sum of the rounded rows by a cent. By cost type this is 8.14% commission, 38.54% spread and 53.32% financing-and-borrow: over a ten-day hold, financing — a cost [Course Book §21.4]'s per-trade toll $c_{\text{leg}}$ does not itemise at all, being a per-trade rather than a per-day-held cost — is the single largest line, ahead of spread.

**Against a 2-sigma spread-reversion gross.** Suppose this pair's stationary spread dispersion is $\hat{\sigma}_z = \$1.35$ per share of $Y$ held, entered at $z_{\text{in}} = 2$ and exited at $z_{\text{out}} = 0$ — exactly [Course Book §21.4]'s canonical cycle. The gross, scaled to the 600-share position, is

$$\text{gross} = 600 \times (2 - 0) \times \$1.35 = \$1{,}620.00,$$

$$\text{net} = \$1{,}620.00 - \$200.51 = \$\mathbf{1{,}419.49},$$

$$\frac{\text{cost}}{\text{gross}} = 12.38\%, \qquad \frac{\text{net}}{\text{gross}} = 87.62\%.$$

This pair clears its costs far more comfortably than [Course Book §21.4]'s own illustration, precisely because its $\hat{\sigma}_z$ is large relative to the toll — exactly what [Course Book §21.6]'s speed and cost screens are built to select for, and exactly what Section 9.4's name-level checklist exists to protect before the statistics are even run: a wide, liquid, cheaply-borrowed pair like this one has no reason to fail on cost; a pair that limps through the checklist with a borrow fee near the cap and a spread near the ceiling will not.

[Course Book §22.4]'s master table already lists this chapter's two cost structures side by side with FX and commodities: "double-leg spread per trade; borrow/short availability" for the equity-pairs column, against a single spread-plus-rollover for FX and a financing charge that embeds roll yield for commodities. Nothing in that row is generic — Worked Example 9.4 is what it costs out to in dollars, for one illustrative pair, once financing is added to the two items the table cell names explicitly.

## 9.6 Exercises

**Exercise 9.1 — Fair-basis drills and the premium/discount reading**

(a) A cash index stands at $S = 6{,}420.50$. Money-market cash yields $r = 4.60\%$ p.a., the basket's dividend yield is $q = 2.05\%$ p.a., and the nearest future expires in 91 days. Compute the fair value $F$ and the basis, and state whether the future trades at a premium or a discount to cash.

(b) A second cash index stands at $S = 5{,}150.00$, with $r = 3.75\%$ p.a., $q = 4.90\%$ p.a., and 35 days to expiry. Compute $F$ and the basis, and state the reading.

(c) The future in part (a) is actually quoted in the market at $F_{\text{mkt}} = 6{,}431.00$. Is it trading rich or cheap to your part (a) fair value, by how many points, and as what percentage of $S$?

▷ working space

**Solution — Exercise 9.1**

**(a)** $T = 91/365 = 0.249315$ years, $(r-q)T = (0.0460 - 0.0205) \times 0.249315 = 0.00635753$,

$$F = 6{,}420.50 \times (1 + 0.00635753) = \mathbf{6{,}461.32}, \qquad \text{basis} = +\mathbf{40.82} \text{ points.}$$

$r > q$, so the future trades at a **premium** (contango).
**(b)** $T = 35/365 = 0.095890$ years, $(r-q)T = (0.0375 - 0.0490) \times 0.095890 = -0.00110274$,

$$F = 5{,}150.00 \times (1 - 0.00110274) = \mathbf{5{,}144.32}, \qquad \text{basis} = -\mathbf{5.68} \text{ points.}$$

$q > r$, so the future trades at a **discount** (backwardation).
**(c)** Deviation $= 6{,}431.00 - 6{,}461.32 = -\mathbf{30.32}$ points, or $-0.47\%$ of $S = 6{,}420.50$. The market price sits *below* fair value, so the future is trading **cheap**: the theoretical arbitrage direction is reverse cash-and-carry (buy the future, sell the basket), and a deviation this size is still well inside the range that basket-execution costs alone can explain without any genuine mispricing.

**Exercise 9.2 — Dividend-adjustment arithmetic, long and short, with a haircut**

An index CFD is quoted at \$25 per index point. Over one week its constituents contribute a total dividend drag of 3.10 index points, and the provider credits long positions 90% of the gross point value — a withholding-style haircut on this particular index's foreign constituents — while debiting short positions the full 100%. Independently of the dividend drag, the index's net observed change over the week, its actual printed capital-return move, is $+5.30$ points.

(a) Compute the week's price P&L for a long position and for a short position.
(b) Compute the dividend credit (long) and debit (short).
(c) Compute each side's true net P&L, and compare it with what a raw-price-only backtest of the same week would have shown.

▷ working space

**Solution — Exercise 9.2**

**(a)** Price P&L $= 5.30 \times \$25 = \$132.50$ for the long; $-\$132.50$ for the short.
**(b)** Long credit $= 0.90 \times 3.10 \times \$25 = \$69.75$. Short debit $= -1.00 \times 3.10 \times \$25 = -\$77.50$.
**(c)**

$$\text{LONG net} = \$132.50 + \$69.75 = \$\mathbf{202.25},$$

$$\text{SHORT net} = -\$132.50 - \$77.50 = -\$\mathbf{210.00}.$$

A raw-price-only backtest reads $\pm\$132.50$ for the two sides. It **understates** the long's true result by exactly the \$69.75 credit, and it **overstates** the short's true result — makes it look \$77.50 better than reality — by hiding the debit entirely. The two errors are not mirror images of each other in size, because the credit and debit fractions here are themselves asymmetric (90% vs. 100%), a detail that matters only once you already know to look for it, which is the entire point of Section 9.2's trapbox.

**Exercise 9.3 — A gap-attribution variant**

The previous cash close was $S_{\text{close}} = 5{,}212.40$. Overnight, the index CFD/future drifted down, quoting $Q_{\text{pre}} = 5{,}198.75$ just before today's local open. The actual opening auction prints $S_{\text{open}} = 5{,}190.00$.

(a) Compute the total gap, the overnight component, and the residual (open-auction) component.
(b) Express each component as a percentage of the total gap.
(c) A trader was short the CFD overnight. How much of the eventual move had they already captured, mark-to-market, before the local open?

▷ working space

**Solution — Exercise 9.3**

**(a)**

$$\text{total gap} = 5{,}190.00 - 5{,}212.40 = -\mathbf{22.40} \text{ points},$$

$$\text{overnight component} = 5{,}198.75 - 5{,}212.40 = -\mathbf{13.65} \text{ points},$$

$$\text{residual component} = -22.40 - (-13.65) = -\mathbf{8.75} \text{ points}.$$

**(b)** Overnight: $-13.65/-22.40 = \mathbf{60.94}\%$. Residual: $-8.75/-22.40 = \mathbf{39.06}\%$ (sum 100%).
**(c)** The short trader's overnight mark-to-market gain tracks the overnight component exactly, since the position gains as the quote falls: 13.65 of the eventual 22.40-point decline was already captured before the open, and the remaining 8.75 points arrived as a fresh move at the opening auction — new information the short position benefited from but could not have traded into any earlier than everyone else, since it did not exist until the auction uncrossed.

**Exercise 9.4 — Building the v3 leg-screening checklist**

Six candidate names have been proposed for a pairs universe. Using the illustrative thresholds of Section 9.4 (ADV $\geq$ \$10m, spread $\leq$ 10bps, borrow fee $\leq$ 3.00% p.a., no pending corporate action, no recent halt), screen each name and report which of the $\binom{6}{2}$ raw candidate pairs remain operationally viable.

| Name | ADV (\$m) | Spread (bps) | Borrow (% p.a.) | Pending CA | Recent halt |
|---|---|---|---|---|---|
| A | 42.0 | 3 | 0.35 | N | N |
| B | 8.5 | 11 | 1.20 | N | N |
| C | 65.0 | 2 | 4.75 | N | N |
| D | 3.2 | 18 | 0.60 | N | N |
| E | 51.0 | 4 | 0.55 | Y | N |
| F | 37.5 | 5 | 0.90 | N | N |

▷ working space

**Solution — Exercise 9.4**

Apply all five screens to each name:

| Name | Verdict |
|---|---|
| A | PASS (clears all five) |
| B | FAIL — ADV 8.5 $<$ 10 *and* spread 11 $>$ 10bps |
| C | FAIL — borrow 4.75% $>$ 3.00% cap |
| D | FAIL — ADV 3.2 $<$ 10 *and* spread 18 $>$ 10bps |
| E | FAIL — pending corporate action |
| F | PASS (clears all five) |

Only A and F pass every screen. The raw candidate count is $\binom{6}{2} = \frac{6\times5}{2} = 15$ pairs, of which only $\binom{2}{2} = 1$ — the pair $(A, F)$ — has both legs operationally viable; the other 14 are excluded before a single cointegration test runs, regardless of what the price history of any of those pairs would have shown. This is [Course Book §21.6]'s combinatorial arithmetic playing out at the smallest possible scale: just as a 500-name universe's 124,750 raw pairs survive an economic pre-filter down to a low single-digit percentage before cointegration is even tested, this six-name universe survives its own prior, operational filter down to one pair in fifteen (6.67%). Passing this checklist is necessary, not sufficient: $(A, F)$ is now eligible to be tested for cointegration, not certified to be cointegrated.

**Key idea — What to remember**

An index CFD is two removes from the number Chapter 8 taught you to compute: the cash index does not trade, the future prices it forward by a derivable cost of carry, and the CFD inherits that carry while adding a near-continuous session against a cash market that is not. Every dividend that Chapter 8's ex-date arithmetic drops from the cash print reappears as an explicit credit or debit on the CFD ledger, in opposite directions for long and short, and a backtest that reads only the price index — or, just as wrongly, a total-return series with adjustments bolted on top — will misprice both sides of that trade in opposite directions. The same near-continuous quoting that lets a CFD track its future through the night manufactures a genuine, decomposable gap at every local cash open, concentrated in one bar the evaluation harness must treat as structurally different from its neighbours. Single names inherit none of this basket machinery, but add their own prior discipline — liquidity, borrow, corporate-action timing, halts policy — that must clear before [Course Book §21.4] and [Course Book §21.6]'s statistics are ever run, plus a full cost stack, financing and borrow included, that a spread-only mental model built for FX will systematically underprice. The chapters ahead turn to derivatives and volatility instruments, where this carry arithmetic generalises again, and to commodities, where it generalises once more.

PART IV

# Derivatives and Commodities



LUNA LABS

# 10

# Futures Markets

*Chapter 9 built one instrument's entire route from a cash index to a tradeable CFD, and the middle link of that chain — the exchange-traded, centrally cleared future — was treated as a black box, priced by a no-arbitrage formula and otherwise left alone. This chapter opens that box. It is not about any one underlying: the machinery here is what every listed future shares, whether the underlying is a stock index, a currency, a bond, or a barrel of crude, and it is the machinery the reader's own WTI CFD sits on top of even though that account never itself submits a futures order. Four objects recur through every section. The* contract *standardises what Chapter 1 left generic — underlying, size, tick, expiry, settlement — into a single fungible instrument two strangers can trade without negotiating terms. The* clearing house *interposes itself between every buyer and seller, converting Chapter 1's bilateral-credit-risk OTC world into something structurally different, and* margin *is what makes that guarantee credible, marked to market in cash every single day rather than settled once at expiry. The* curve *— the same contract at successive expiries, quoted simultaneously — carries information the spot price alone cannot, generalises Chapter 9's cost-of-carry relationship to commodities, and forces every position held past one contract's expiry through a* roll *whose cash cost or credit is neither noise nor a broker's invention. Options markets, taken up next, are built on exactly this exchange-and-clearing chassis; commodities, two chapters further on, are where the curve's storage-and-convenience-yield term does most of its work. Both borrow this chapter's vocabulary without re-deriving it.*

## 10.1 The Contract

A future is what a forward (Chapter 4) becomes once an exchange, rather than two counterparties, writes every term of the agreement except the traded price. Chapter 4 priced the FX forward as a bilateral, customisable contract — any notional, any value date, negotiated once with a single dealer — and noted only in passing that exchange-traded FX futures exist, standardising size, value date and margining in exchange for a liquidity and counterparty structure the OTC alternative cannot offer. This section makes that trade-off precise, for any underlying, not only currencies.

**Definition — The futures contract**

A **futures contract** is a standardised, exchange-listed agreement to buy (the long) or sell (the short) a fixed quantity of an underlying at a price agreed today, for settlement at a fixed future date, with every term besides that price fixed by the exchange rather than negotiated bilaterally.

**Underlying.** The asset, index, or rate the contract references — a currency pair, an equity index level, a government bond, a barrel of crude, an interest rate.

**Contract size (multiplier), $m$.** The fixed quantity of the underlying one contract represents, or, for a cash-settled financial future, the dollar value of one full point of the quoted price. Every price quoted on the contract is implicitly multiplied by $m$ to reach a cash value; $m$ is

fixed by the exchange and never varies contract to contract within one listing.
**Tick size, $\tau$, and tick value.** The minimum price increment the exchange will accept on an order or print on the tape; no valid price on the contract can differ from another by less than $\tau$. The *tick value* $v_\tau = \tau m$ converts one minimum price increment directly into a cash amount per contract, derived in full below.
**Expiry (delivery) cycle.** The fixed calendar of dates on which listed contracts expire, conventionally named by month and, for many financial futures, clustered on a quarterly cycle (March, June, September, December, often coded H/M/U/Z after the IMM's original lettering) while many physical-commodity futures list monthly or near-monthly to track a harvest, delivery, or storage-drawdown calendar; the exact set of listed months is contract-specific and must be read off the exchange's own listing, not assumed from the underlying alone.
**Settlement method.** *Physical delivery*, in which the short who has not closed the position by expiry must deliver the actual underlying (or, for some contracts, one of a defined basket of deliverable grades or bonds) to the long against payment of the final settlement price; or *cash settlement*, in which no delivery occurs and the contract simply pays, at expiry, the difference between the entry price and a defined final settlement value computed from the underlying market (an index close, a benchmark fixing). Financial index futures are typically cash-settled; physical commodities split between the two, and which applies to a given contract is, again, a fact to read off the specification, never to assume.

**Result — Tick value and contract P&L**

If a contract's tick size is $\tau$ and its contract size (multiplier) is $m$, the **tick value** is

$$v_\tau \;=\; \tau m,$$

the cash consequence of a one-tick price move on a single contract. A price change of $\Delta F$ — necessarily an exact multiple of $\tau$, for any price the exchange will actually print — on a position of $N$ contracts produces

$$\text{P\&L} \;=\; \Delta F \cdot m \cdot N \;=\; \left(\frac{\Delta F}{\tau}\right) v_\tau N, \tag{10.1}$$

two equivalent routes to the same cash figure: the first works directly from the raw price change, the second from a tick count a trading screen typically displays already, and neither requires knowing anything about the contract beyond $m$ and $\tau$.

**Worked Example 10.1 — Tick value on two hypothetical contracts**

Two contracts invented for this example only — no real exchange's current specification is implied by either, and the multiplier and tick size of any real contract must be read off that exchange's own listing before it is ever traded or sized.
**Contract Alpha**, a hypothetical stock-index future, has multiplier $m = \$20$ per index point and tick size $\tau = 0.10$ points, so by (10.1)'s definition,

$$v_\tau = 0.10 \times \$20 = \$\mathbf{2.00}.$$

Its settlement price moves from 4,215.30 to 4,219.80, a rise of 4.50 points, or 45 ticks. A trader long 6 contracts realises

$$\text{P\&L} = 4.50 \times \$20 \times 6 = \$\mathbf{540.00},$$

which matches the tick count route exactly: $45 \times \$2.00 \times 6 = \$540.00$.

**Contract Beta**, a hypothetical soft-commodity future, has contract size $m = 2{,}500$ units and tick size $\tau = \$0.0025$ per unit, so $v_\tau = 0.0025 \times 2{,}500 = \$\mathbf{6.25}$. Its price falls from 118.400 to 118.375, a move of $-0.025$, or 10 ticks. A trader *short* 3 contracts gains on the fall:

$$\text{P\&L} = -(-0.025) \times 2{,}500 \times 3 = \$\mathbf{187.50}.$$

Two contracts, two unrelated multipliers, two unrelated tick sizes: nothing about "one tick" is comparable across contracts, and no position-sizing arithmetic in this book skips the step of reading $m$ and $\tau$ off the specific contract actually being traded.

### Central clearing and novation

Chapter 1 named the central counterparty in passing, as the feature separating an exchange from an OTC market; this section derives exactly what that interposition does, and Section 10.2 derives what it costs the exchange to make the resulting guarantee credible rather than assumed.

Consider two strangers, a buyer and a seller, matched on an exchange's order book — the same continuous-double-auction matching Chapter 2 built from first principles, applied here to a standardised, expiring contract rather than a listed share. Absent anything more, each now bears the other's default risk for the life of the contract: exactly the bilateral exposure an FX forward or swap (Chapter 4) carries for its own tenor, and exactly what an OTC market structurally is (Duffie, 2012). A **central counterparty (CCP)** removes this by **novation**: immediately after the trade is matched, the original bilateral contract is legally replaced by two new contracts on identical economic terms — one between the buyer and the CCP, one between the CCP and the seller. Neither original party faces the other again; both face the CCP, and only the CCP. The buyer's counterparty risk is now the CCP's solvency, not the seller's, and symmetrically for the seller — a substitution worth making only because the CCP is built, deliberately, to be a better credit risk than either original party could individually offer the other (Hull, 2021).

Novation has a second consequence, independent of credit substitution, and it is the one that makes an exchange-traded future liquid in a way a purely bilateral contract cannot be. Because every novated contract is, from either side, a claim on the CCP rather than on a named counterparty, closing a position never requires finding the specific party who was on the other side of the original trade: it requires only trading an equal and opposite quantity of the same contract with anyone, after which the CCP nets the two positions against each other and the trader's obligation to it disappears. This is the structural reason a futures position can be closed out, in full or in part, at any moment before expiry with no counterparty search at all — where an OTC forward (Chapter 4), absent a separate unwind negotiated with the original dealer, is generally held open to its value date. Major derivatives exchanges each operate, or are served by, their own clearing house (CME Clearing and ICE Clear among the deepest global examples; ASX Clear (Futures) the reader's local venue for exchange-traded derivatives) — naming which institutions perform this role is structural and stable; the specific margin schedules and product lists any of them run are not, and are flagged as such throughout this chapter.

**Trap — A CCP does not delete risk, it concentrates and re-prices it**

Novation removes bilateral credit risk between the original two parties; it does not remove credit risk from the system, it re-routes every trade's risk onto one balance sheet built, specifically, to net thousands of offsetting positions and to hold enough collateral and mutualised default resources to survive the failure of its largest member in stressed conditions. That balance sheet's own solvency is not a law of nature — it is the product of exactly the margining discipline Section 10.2 derives, collected without exception, from every member, every single trading day. A CCP that ever let that discipline lapse would simply relocate the 1980s-vintage

bilateral counterparty risk this section just removed back onto every participant at once, concentrated rather than diversified.

## 10.2 Margin and Mark-to-Market

Section 10.1's CCP guarantee is not free, and margin is where its cost — and its discipline — lands on the trader every day the position is open. Every futures position is revalued once at the end of every trading session against an official **settlement price** the exchange (via its clearing house) fixes for that contract — typically a defined average of trade prices in a closing window rather than simply the last print, precisely so that a single thinly-traded final trade cannot move an entire market's margin calculations. Every mechanic below operates off successive settlement prices; §10.6 returns to why the same official-price discipline governs convergence at expiry.

**Definition — Initial, maintenance and variation margin**

**Initial margin (IM).** The cash, or eligible collateral, a trader must post before a position can be opened, sized by the exchange or clearing house to cover, with high confidence, the largest loss the position could plausibly suffer before it could be closed out over a short holding horizon (a risk-based methodology such as SPAN, or a comparable value-at-risk-style model, recalibrates this continuously across thousands of contracts). The resulting figure, whether quoted in currency or as a percentage of notional, is a drifting one — indicative only, as of the mid-2020s, and to be verified against the exchange's current margin schedule before it is used for anything beyond illustration.
**Maintenance margin (MM).** A lower threshold, set below IM and commonly some fixed fraction of it (again exchange- and contract-specific, and equally subject to the same verify-current caveat), that the account's equity may fall to before a margin call is triggered — existing so that ordinary daily noise does not force a top-up on every single losing day.
**Variation margin (VM).** The cash amount credited to, or debited from, the margin account at each settlement, equal to the change in the position's mark-to-market value since the previous settlement: realised, not notional, cash that actually moves.

**Result — Daily settlement mechanics**

For a position of $N$ contracts (long positive, short negative) in a future with multiplier $m$, marked at settlement price $F_{t-1}$ yesterday, today's settlement at $F_t$ produces

$$\mathrm{VM}_t \;=\; (F_t - F_{t-1})\, m\, N, \tag{10.2}$$

credited to the margin account if positive, debited if negative, every session without exception, whether or not the position is ever closed. If the balance after posting $\mathrm{VM}_t$ falls below MM, the clearing member issues a **margin call**: the trader must deposit enough cash to bring the balance back up to IM — not merely back up to MM — by a stated deadline, or the position is liquidated, in whole or in part, without further reference to the trader. A shortfall against MM is never left to persist; the call is a demand with a deadline, not a warning. A balance above IM after a favourable settlement triggers no action at all: the excess is available to withdraw, but nothing compels withdrawal, and holding it invites no call of its own.

**Worked Example 10.2 — A margin sequence: five days, one call**

A hypothetical ASX 24-style equity-index future — illustrative multiplier and margin levels only, not a real contract's specification — has multiplier $m = \text{A\$50}$ per index point, initial margin IM = A\$6,000 and maintenance margin MM = A\$4,500 per contract. A trader buys one contract at $F_0 = 4{,}500.00$ on Day 0, posting IM, and holds it through five subsequent settlements.

| Day | Settle | $\Delta F$ | VM (A\$) | Balance (A\$) | Action |
|---|---|---|---|---|---|
| 0 | 4,500.00 | — | — | 6,000.00 | post IM |
| 1 | 4,483.00 | −17.00 | −850.00 | 5,150.00 | none (>MM) |
| 2 | 4,466.50 | −16.50 | −825.00 | 4,325.00 | **call**: top up +1,675.00 |
| | | | | 6,000.00 | (balance restored to IM) |
| 3 | 4,478.25 | +11.75 | +587.50 | 6,587.50 | none (>IM, no withdrawal forced) |
| 4 | 4,492.00 | +13.75 | +687.50 | 7,275.00 | none |
| 5 | 4,505.50 | +13.50 | +675.00 | 7,950.00 | none |

Day 2's settlement takes the balance to A\$4,325.00, below MM = A\$4,500.00: a margin call for 6,000.00 − 4,325.00 = A\$**1,675.00**, restoring the balance to IM exactly, not merely to MM. From Day 3 the balance runs freely above IM as the position recovers; no further action is triggered, and a trader who left the excess in the account (rather than withdrawing it) simply carries a larger buffer against the next adverse move.

Two reconciliations confirm the arithmetic. First, summing the five daily VMs,

$$-850.00 - 825.00 + 587.50 + 687.50 + 675.00 = +\mathbf{275.00},$$

which must equal, and does equal, the position's total price P&L computed directly: $(F_5 - F_0)\, m = (4{,}505.50 - 4{,}500.00) \times 50 = 5.50 \times 50 = +\mathbf{275.00}$. Second, total cash deposited — initial margin plus the Day 2 top-up — was 6,000.00 + 1,675.00 = A\$7,675.00, and the final balance is A\$7,950.00, so

$$\text{final balance} - \text{total deposited} = 7{,}950.00 - 7{,}675.00 = +\mathbf{275.00}$$

again — the margin account, despite being interrupted by a call partway through, is exactly self-consistent: whatever is left after backing out every dollar the trader put in is precisely the position's cumulative mark-to-market gain, no more and no less.

**Leverage.** The position's notional at entry is $F_0\, m = 4{,}500.00 \times 50 = \text{A\$225,000.00}$ against an initial outlay of only A\$6,000.00, a leverage ratio of

$$\frac{\text{notional}}{\text{IM}} = \frac{225{,}000.00}{6{,}000.00} = \mathbf{37.5{:}1}.$$

This ratio is arithmetic on the numbers given, not a market fact; a real contract's IM, and hence its implied leverage, is set by the exchange's current margin schedule and moves with volatility, never assumed constant from one worked example to a live account.

**Trap — Variation margin is realised cash, not paper P&L**

Every VM in Worked Example 10.2 is real money that actually left or entered a bank account behind the margin account — not a mark on a spreadsheet a trader can look past until the position closes. This is the single most consequential difference between a futures (or futures-referencing CFD) position and an unlevered cash position: an unrealised loss on a share held outright costs nothing in cash until it is sold, but an unrealised loss on a levered futures

position costs cash *today*, in the exact amount (10.2) computes, whether or not the position is ever closed and whether or not the loss is later reversed. Chapter 15's case studies return to this repeatedly: a position that is, on any honest assessment, right eventually can still be forced out at the worst possible moment if the cash to meet a margin call is not there on the day the call is made, and it is the size of that daily funding demand — not the position's ultimate profitability — that determines whether a levered account survives long enough to find out it was right. Position sizing against a margin schedule is a funding-liquidity problem before it is ever a directional one.

## 10.3 Volume, Open Interest, and Who Is in the Market

**Definition — Open interest and volume**

**Open interest (OI)** is the number of futures contracts of a given expiry currently outstanding — neither closed by an offsetting trade nor settled by delivery or cash expiry — counted once per contract; every long has an exactly offsetting short, so OI is unambiguous whichever side it is counted from. **Volume** is the number of contracts that changed hands over a stated period (a session, typically), counting every executed transaction once regardless of whether it opened a new position, closed an existing one, or simply transferred an existing position from one trader to another. OI is a *stock*, measured at a point in time; volume is a *flow*, measured over an interval; the two answer different questions, and neither can be inferred from the other alone.

Every trade is one buyer and one seller, but either side of a trade can be either opening a new position or closing an existing one, and the four combinations have four different consequences for OI while all four raise volume identically (Hull, 2021):

| Buyer's side | Seller's side | ΔOI | ΔVolume |
|---|---|---|---|
| Opens a new long | Opens a new short | $+1$ | $+1$ |
| Opens a new long | Closes an existing long (sells to close) | $0$ | $+1$ |
| Closes an existing short (buys to close) | Opens a new short | $0$ | $+1$ |
| Closes an existing short (buys to close) | Closes an existing long (sells to close) | $-1$ | $+1$ |

Rising OI alongside rising volume signals net new position formation — fresh commitments on both sides, the market's total risk-bearing capacity genuinely expanding. Volume with unchanged OI is pure turnover: existing exposure changing hands, no net new commitment created anywhere. Falling OI alongside volume signals unwinding — positions closing faster than new ones open, commonly (though not exclusively) as a contract nears expiry and holders roll or exit (§10.5). Reading OI in isolation, without volume, routinely misleads: a contract can show large, static OI simply because nobody has needed to trade it recently, not because it is liquid.

Chapter 1's hedger/speculator/arbitrageur trichotomy classifies a trade's economic motive without reference to any one instrument; applied to the open interest just defined, it says something concrete. Every contract of OI is a matched pair of one long and one short, and the economically interesting question is never the raw OI figure alone but *who* is on each side of it. A commercial hedger — a producer selling output forward, a consumer buying an input forward, an asset manager hedging an existing portfolio's exposure — uses the contract to reduce a risk arising from its underlying business; a speculator carries no offsetting exposure and is compensated, if the classical theory is right, for absorbing risk the hedging side wants to shed (§10.4 makes the

compensation mechanism explicit). Whether that compensation is real, and how large it is net of costs, is an empirical question the commodity-futures-returns literature has tested directly rather than assumed (Gorton, Rouwenhorst, 2006; Erb, Harvey, 2006), and §10.4's treatment of normal backwardation returns to what that literature found.

**The COT connection.** Chapter 7 built detailed machinery around the CFTC's weekly Commitments of Traders report without yet connecting it to the open interest defined here; the connection is definitional, not incidental. Every figure the COT report publishes — the Legacy report's Commercial, Non-Commercial and Non-Reportable long and short totals, or the finer Disaggregated and Traders-in-Financial-Futures splits Chapter 7 tabulated in full — categorises exactly the open interest this section defines, by which type of participant holds each side of it, and is not a separate dataset measuring something new. A COT $z$-score is therefore a statement about the *composition* of a contract's open interest — is the non-commercial long share unusually large relative to its own history — never a statement about price on its own, and never available for a market this chapter's machinery does not apply to: COT exists only for CFTC-regulated, exchange-cleared futures, which is exactly why Chapter 7 needed a proxy construction for currency pairs that list no CFTC-regulated future of their own.

## 10.4 The Curve

A single underlying rarely has just one listed future. An exchange typically lists the same contract at several simultaneous expiries — quarterly, monthly, or on whatever cycle Section 10.1's defbox left contract-specific — each trading, clearing and carrying its own open interest entirely independently of the others, priced off the same underlying but not required to move in lockstep with it, or with each other. The resulting cross-section, price plotted against time to expiry at one instant, is the **futures curve**, or **term structure**, and it carries information about financing, storage and expectations that the single spot or near-month price alone cannot.

### Financial futures: the cost-of-carry recap

Chapter 9 derived, for a stock-index future, the no-arbitrage relationship pinning the future's price to the cash index today:

$$F = S\big(1 + (r - q)T\big), \tag{10.3}$$

where $S$ is the cash price today, $r$ the financing rate over the contract's tenor $T$, and $q$ the yield — there, a dividend yield — the underlying pays out over the same horizon. Nothing in that derivation was specific to equity indices: it applies unchanged to any *financial* underlying with a well-defined carry cost and a well-defined yield — a government bond future (carry the bond, receive its coupon in place of $q$), a currency future (Chapter 4's covered interest parity is the identical relationship, with the foreign interest rate standing in for $q$). What every financial underlying shares is that holding it physically costs, at most, foregone interest and pays, at most, a cash yield; nothing about holding a bond or a basket of shares requires a warehouse.

### Commodities: storage cost and convenience yield

A physical commodity breaks that assumption. Holding a barrel of crude, a tonne of wheat, or an ounce of gold between today and $T$ costs something beyond foregone interest — storage, insurance, spoilage risk — and, more subtly, can also *pay* something beyond any cash yield: a manufacturer holding physical inventory avoids the risk of a stock-out disrupting production, a refiner holding crude can respond to a sudden demand spike without waiting on a shipment, and this optionality is worth something to whoever holds the physical good even though it generates no cash flow and appears on no dividend calendar. Working (1949) formalised this benefit as the **convenience yield**: an implicit return to holding the physical commodity itself, distinct from any return on a financial claim to it, that rises when inventories are scarce — a physical stock-out is a live risk — and falls toward zero when inventories are abundant.

**Result — Commodity futures fair value: storage cost and convenience yield**

Let $S$ be the spot commodity price, $r$ the financing rate, $u$ the proportional cost of storing and insuring the physical commodity (continuously compounded, per annum), and $y$ the convenience yield — the implicit, non-cash benefit of holding the physical good, likewise a continuously compounded annual rate. Extending Chapter 9's cost-of-carry argument by treating $u$ as a carry cost paid (a negative yield) and $y$ as a yield received (exactly like $q$ there, but non-cash and never directly observable) gives

$$F = S\,e^{(r+u-y)T}. \tag{10.4}$$

$r + u$ is the full cost of carrying the physical commodity forward — financing plus storage — and $y$ is netted against it because a holder of the physical good captures a convenience benefit that a holder of a paper future does not; $y$ is precisely the wedge that lets $F$ sit *below* what financing-plus-storage cost alone would demand, and, when $y$ is large relative to $r + u$, below $S$ itself.

Every term on the right of (10.4) except $S$ is unobservable directly and must be inferred rather than read off a quote screen: $r$ is a market financing rate, reasonably well pinned down, but $u$ and $y$ are almost never quoted separately, and in practice the curve is read the other way around — $F$, $S$ and $r$ are observed, and $u - y$ is backed out as whatever residual reconciles them, with $u$ and $y$ individually never separately identified from price data alone. Black (1976) later showed that once $F$ itself is taken as the state variable to be modelled, an option or any other derivative on the future can be priced without ever separating $u$ from $y$ at all — the curve already impounds both, wherever exactly it came from.

### Contango, backwardation, and the risk-premium reading

[Course Book §22.1] defines a curve as in *contango* at time $t$ if a later-dated contract trades above an earlier-dated one, and in *backwardation* if the reverse holds — a description of the curve's shape at a single instant, carrying no information on its own about which direction the underlying spot price will subsequently move. (10.4) shows mechanically what shape follows from what inputs on a commodity curve: $u > y$ (storage cost dominates convenience yield, the normal case when inventories are ample) produces contango; $y > u$ (convenience yield dominates, the case when inventories are tight and physical scarcity is being priced) produces backwardation, potentially even with $F < S$.

Keynes (1930) offered a second, independent reading of a downward-sloping curve, applying not to the curve's shape today but to where $F$ sits relative to the *expected future spot price* $\mathbb{E}[S_T]$. His theory of **normal backwardation** holds that commercial hedgers in a given commodity are, on net, more often short (producers hedging output) than long (consumers hedging input), so the futures price must sit *below* $\mathbb{E}[S_T]$ to compensate the speculators willing to take the other, net-long side of that hedging demand — a risk premium for bearing price risk the hedgers want to shed, conceptually distinct from, and not directly observable alongside, the storage/convenience decomposition above. The two readings do not conflict — one describes the curve's shape today, from quantities that are at least in principle observable; the other describes an expected-return relationship that today's shape alone cannot verify, since $\mathbb{E}[S_T]$ is never observed directly — and Section 10.3's citation to the empirical commodity-returns literature is precisely the record of testing whether Keynes's compensation is actually there once costs are accounted for, not merely a plausible story left untested.

Reading a curve two points at a time makes an implied carry rate directly recoverable without separating $r$, $q$, $u$ or $y$ individually. If a constant effective, continuously-compounded carry rate $c$ priced the curve between two expiries $T_1 < T_2$ quoted at the same instant, then $F_2 = F_1 e^{c(T_2 - T_1)}$,

so

$$c = \frac{\ln(F_2/F_1)}{T_2 - T_1}, \tag{10.5}$$

the **implied annualised carry** between the two legs — $r - q$ read directly off a financial curve via (10.3), or $r + u - y$ read directly off a commodity curve via (10.4) — without the individual pieces ever needing to be observed.

**Worked Example 10.3 — Reading three curves**

Three hypothetical curves, each given as two simultaneously quoted expiries, illustrative prices only.

**Curve P (hypothetical base-metal future).** $T_1 = 2$ months, $F_1 = 8{,}412.00$; $T_2 = 6$ months, $F_2 = 8{,}624.40$. Since $F_2 > F_1$, the curve is in **contango**. With $\Delta T = 4/12 = 0.333333$ years,

$$c = \frac{\ln(8{,}624.40/8{,}412.00)}{0.333333} = \mathbf{7.48}\% \text{ p.a.}$$

**Curve Q (hypothetical energy future).** $T_1 = 1$ month, $F_1 = 71.85$; $T_2 = 6$ months, $F_2 = 66.40$. Since $F_2 < F_1$, the curve is in **backwardation**. With $\Delta T = 5/12 = 0.416667$ years,

$$c = \frac{\ln(66.40/71.85)}{0.416667} = -\mathbf{18.93}\% \text{ p.a.}$$

**Curve R (hypothetical equity-index future).** $T_1 = 3$ months, $F_1 = 4{,}512.00$; $T_2 = 9$ months, $F_2 = 4{,}556.30$. Since $F_2 > F_1$, a mild **contango**. With $\Delta T = 0.5$ years,

$$c = \frac{\ln(4{,}556.30/4{,}512.00)}{0.5} = \mathbf{1.95}\% \text{ p.a.}$$

Curve P's near-7.5% implied carry is consistent with meaningful storage cost net of convenience yield on an amply stocked metal; Curve Q's sharply negative implied carry — the market paying, not charging, to hold the near contract relative to the far one — is the signature of a tight physical market, exactly Working's convenience-yield channel dominating; Curve R's modest, positive figure sits in the range a financing-rate-less-dividend-yield relationship typically produces for a broad equity index, an order of magnitude calmer than either commodity curve because nothing about holding a cash-settled index future involves a warehouse. None of the three numbers says anything about where any of these instruments' *spot* prices are headed; each says only what today's curve is currently charging, or paying, to hold the position from the near date to the far one.

## 10.5 The Roll

### Calendar mechanics

As one expiry approaches, a participant who wants to keep the same exposure past that date cannot simply wait: an expiring contract will deliver, cash-settle, or otherwise cease to exist on its fixed date (Section 10.1), so maintaining the exposure means closing the expiring position and opening an equivalent one in the next listed expiry — *rolling* the position forward. Liquidity migrates with this decision, not instantaneously but over a window in the days to weeks before expiry that is contract-specific and itself a matter of convention rather than physics: open interest and volume in the "front month" contract decay as holders roll or close out, while the next contract's OI and volume build, producing a crossover point before which market participants, and vendors constructing a "continuous" price series (below), conventionally redefine which contract counts as the front month. Exchanges, index providers and data vendors each publish

their own roll calendar or roll rule for a given contract; none of it is universal across contracts or stable enough to state here as a fixed day-count, and the operative rule for any specific series must be read off that provider's current documentation, never assumed from another contract's convention.

### Roll P&L

The roll itself is a real transaction, not a bookkeeping convention, with a real cash consequence identical in mechanism to the roll drag [Course Book §22.1] derived for a CFD tracking a rolling futures curve: closing the position in the expiring contract at its final relevant price $F_1$ and opening the equivalent position in the next contract at its prevailing price $F_2$, as near to the same instant as execution allows, so the position's economic exposure to the underlying is continuous across the roll even though the contract actually held is not.

**Worked Example 10.4 — Roll P&L in contango**

A hypothetical crude-oil-style future, "Contract W" — illustrative multiplier only, not a real exchange's contract size — has multiplier $m = 250$ (barrels-equivalent) per contract. A trader holds $N = 4$ contracts long in the expiring front-month contract, last relevant at $F_1 = \$82.40$, and rolls into the next contract, trading at $F_2 = \$83.75$: a contango curve.
Closing the four expiring contracts and opening four contracts in the next expiry costs, in cash,

$$N\,m\,(F_2 - F_1) = 4 \times 250 \times (83.75 - 82.40) = 4 \times 250 \times 1.35 = \$\mathbf{1{,}350.00},$$

paid to maintain identical notional exposure across the roll — cash that leaves the margin account exactly as any other variation margin debit would (§10.2), regardless of whether the underlying spot price moves at all that day.
**Reconciling with the roll-drag formula.** [Course Book §22.1] derived the per-unit roll drag $\delta = (F_2 - F_1)/F_1$ for a CFD replicating this same roll:

$$\delta = \frac{83.75 - 82.40}{82.40} = \frac{1.35}{82.40} = \mathbf{1.6383}\%.$$

The position's notional at $F_1$ is $N\,m\,F_1 = 4 \times 250 \times 82.40 = \$82{,}400.00$, and

$$\delta \times \text{notional} = 0.016383 \times 82{,}400.00 = \$\mathbf{1{,}350.00},$$

exactly the cash figure computed directly above — not a coincidence, since algebraically $\delta \times (NmF_1) = \big[(F_2 - F_1)/F_1\big] \times NmF_1 = Nm(F_2 - F_1)$ regardless of the numbers involved. The two arithmetics are the same statement at two levels of the same instrument: Chapter 22 gave the CFD-level percentage, this section gives the futures-level cash flow it is built from.

### Continuous-series construction: back-adjusted vs unadjusted

A backtest that wants one multi-year price history for "crude" or "the index future" cannot use any single contract's own history — no one contract trades for more than a few months to a year before expiring — so the series must be built by splicing successive front-month histories together at each roll, and the splicing method chosen determines whether the result is a faithful record of a continuously-held position's returns or a record contaminated by an artefact in the same family as Chapter 8's unadjusted-dividend trap.

**Definition — Unadjusted vs back-adjusted continuous series**

An **unadjusted (spliced) series** simply concatenates each contract's own raw settlement prices, switching at the roll date from the expiring contract's prices to the next contract's

raw prices with no further adjustment. Every historical price in the series is a genuine, real transaction price of *some* contract on *some* date — but the level jumps by $F_2 - F_1$ at every roll, a jump with nothing to do with that day's market action.

A **back-adjusted series** removes the jump by shifting an entire segment of history at each roll: at the roll date, compute the adjustment $F_2 - F_1$ and add it (an *additive* adjustment; an alternative convention instead multiplies by $F_2/F_1$, preserving percentage rather than absolute returns) to *every* price prior to that roll, cumulatively across every earlier roll already applied. The result is continuous at every roll by construction, and every day-over-day *difference* within a single contract's own history is preserved exactly, since adding a constant to two consecutive prices leaves their difference unchanged — but the *level* of any back-adjusted price more than one roll in the past is no longer any contract's real traded price; it is a synthetic level, correct for computing a continuously-held position's $\Delta P$, not for any point-in-time reading of what the market actually printed on that date.

Contract W's own roll, from Worked Example 10.4, illustrates both constructions against the same four days of prices.

| Day | Contract used | Unadj. level | Unadj. Δ | Back-adj. level | Back-adj. Δ |
|---|---|---|---|---|---|
| −2 | front | 81.90 | — | 83.25 | — |
| −1 | front (= $F_1$) | 82.40 | +0.50 | 83.75 | +0.50 |
| 0 | next (first use, $F_2 = 83.75$) | 84.20 | **+1.80** | 84.20 | **+0.45** |
| +1 | next | 83.95 | −0.25 | 83.95 | −0.25 |

On the roll day, the next contract itself moved from the reference price at which the roll leg executed, $F_2 = 83.75$, to its own Day 0 settlement of 84.20: a true, single-day move of $84.20 - 83.75 = +0.45$. The unadjusted series instead differences 84.20 against the *old* contract's 82.40 — an entirely different instrument's price level — and reports +1.80, overstated by exactly the roll gap $F_2 - F_1 = 1.35$, a jump with nothing to do with one day's trading. The back-adjusted series, having shifted the entire pre-roll segment up by 1.35, differences 84.20 against 83.75 instead and recovers the true +0.45 exactly, while leaving the two non-roll differences (Day $-2 \to -1$, Day $0 \to +1$) completely untouched, since every price on each of those pairs was shifted — or not shifted — by the identical constant.

**Trap — Unadjusted series have phantom jumps at rolls**

This is the futures cousin of Chapter 8's unadjusted-dividend trap, and the failure mode is identical in shape even though the mechanism differs. A backtest, volatility estimator, or cointegration test fed an unadjusted continuous series sees a price move on every roll date that no continuously-held position ever actually experienced: not a real one-day market shock, but the arithmetic difference between two different contracts' price levels, systematically dated to the roll calendar rather than to anything the underlying did that day. Run this across enough contracts and enough rolls and the effect compounds — a multi-year unadjusted crude series can carry a substantial directional drift that is entirely an artefact of a persistent contango or backwardation, nothing to do with the spot price's own path, exactly the kind of manufactured "signal" Chapter 8 warned an unadjusted index series can fabricate on a reporting-season clock. The fix is the same discipline in both cases: adjust the series for the mechanical, calendar-dated jump before a single statistic is computed from it, and never mistake a splicing artefact for the market speaking.

**Your project — The WTI CFD sits on exactly this machinery**

The platform's WTI CFD (v4) never touches a futures contract, a clearing house, or a roll calendar directly — Chapter 1's projbox already established that every SST order is a bilateral OTC contract with OANDA, and this chapter's machinery describes what OANDA, not the platform, must manage on the other side of that contract. OANDA prices the WTI CFD off the underlying futures curve this section built (front-month outright, or a blended construction across the roll window, mirroring Chapter 9's discussion of the analogous choice for index CFDs) and performs the roll on the book's behalf, passing its cash consequence through as a financing or swap line rather than as a price jump the client account ever sees directly. The figure that lands on a v4 account statement is, structurally, this section's $Nm(F_2 - F_1)$, scaled to the CFD's own notional and converted to a per-unit financing rate — not a fee the provider invented, and not evidence of anything the underlying spot price did that day. The accounting job v4 owns is narrow but non-negotiable: attribute that recurring charge to carry, at the rate this chapter's arithmetic predicts given the prevailing curve shape, and never let it leak into a backtest's estimate of the strategy's actual directional edge — exactly the leak this section's trapbox describes for an unadjusted series, now happening inside a live account rather than inside a spliced history.

## 10.6 Basis and Convergence

**Definition — Basis**

The **basis** of a futures contract at time $t$ is

$$b_t \;=\; F_t - S_t,$$

the future's price minus the spot (cash) price of the underlying at the same instant. Chapter 9 derived this quantity for a stock-index future specifically, $F - S = S(r - q)T$; the definition itself is generic to any futures market, financial or physical, and §10.4's cost-of-carry relationships are simply the specific value (10.3) or (10.4) predicts for it at each instant.

### Convergence at expiry

**Result — Convergence: why $F \to S$ as $T \to 0$**

Consider a physically deliverable contract at, or arbitrarily close to, its expiry, where $T \to 0$ so remaining financing and storage costs vanish. Suppose $F_T > S_T$ at that instant: an arbitrageur buys the underlying in the cash market at $S_T$, simultaneously sells (shorts) the future at $F_T$, and immediately delivers the underlying into the future, receiving $F_T$ for an asset that cost $S_T$ — a riskless profit of $F_T - S_T > 0$, locked in with zero net capital and no remaining time exposure. Buying pressure on the cash leg and selling pressure on the future close the gap. Symmetrically, if $F_T < S_T$, the reverse trade — buy the future, take delivery, sell the underlying received — is a riskless profit of $S_T - F_T > 0$, forcing convergence from the other side. Either direction of mispricing is arbitraged away by the same mechanism that priced the curve away from expiry (§10.4), so

$$\lim_{T \to 0} F_T = S_T, \qquad \text{equivalently} \qquad \lim_{T \to 0} b_t = 0.$$

A cash-settled contract enforces the identical limit definitionally rather than by arbitrage: its final settlement price is, by the contract's own terms, set equal to a spot-market reference (an

index close, a benchmark fixing) at expiry, so $b_T = 0$ by construction, with no delivery step required to force it.

### Basis risk for hedgers

A hedger who can hold the futures position to expiry, in a contract whose underlying exactly matches the physical exposure being hedged, locks in today's forward price with certainty — convergence guarantees the exit price. Almost no real hedge satisfies both conditions at once: the hedge is typically lifted before expiry, because the physical transaction it protects happens on its own calendar rather than the exchange's, or the available contract's underlying is only a close match for the exact exposure being hedged (a cross-hedge). Either way, the hedger is exposed not to the full price risk of the underlying but to **basis risk**: the risk that $b_t = F_t - S_t$ moves, unpredictably, between the hedge's inception and its lift.

**Worked Example 10.5 — A hedge with residual basis risk**

A business needs to buy 10,000 units of a commodity in six weeks and, wanting to lock in today's price, buys (goes long) the nearest futures contract today rather than waiting. At inception, spot $S_0 = 84.10$ and the future $F_0 = 85.60$, so $b_0 = F_0 - S_0 = +1.50$. Six weeks later, spot has risen to $S_1 = 88.30$ and the future to $F_1 = 89.40$, so $b_1 = F_1 - S_1 = +1.10$; the business now buys the physical commodity at $S_1$ and simultaneously closes the futures position.

**Effective price paid.** Cash outlay $S_1$, less the gain on the long futures leg, $F_1 - F_0$:

$$\text{effective price} = S_1 - (F_1 - F_0) = 88.30 - (89.40 - 85.60) = 88.30 - 3.80 = \mathbf{84.50},$$

equivalently $F_0 - b_1 = 85.60 - 1.10 = 84.50$. Had the basis been unchanged ($b_1 = b_0$), the effective price would have been exactly $F_0 - b_0 = S_0 = 84.10$: a hedge with no basis risk locks in today's spot price precisely, whichever way the market subsequently moves. Here it did not: the basis narrowed from 1.50 to 1.10 ($\Delta b = -0.40$), so the effective price overshoots the perfect-hedge benchmark by exactly $-\Delta b = +0.40$.

**What the hedge actually achieved.** Left unhedged, the business would have paid $S_1 = 88.30$, a full 4.20 above $S_0$. Hedged, it paid 84.50, only 0.40 above $S_0$: the futures position absorbed 3.80 of the 4.20 price move and left 0.40 as residual basis risk, a

$$\text{hedge effectiveness} = 1 - \frac{|\Delta b|}{|\Delta S|} = 1 - \frac{0.40}{4.20} = \mathbf{90.48}\%$$

reduction in price-risk exposure — not the full 100% a naive reading of "hedged" might assume. Scaled to the full 10,000 units (contract multiplier 100 per contract, 100 contracts): total cost hedged = \$845,000.00 against \$883,000.00 unhedged and an \$841,000.00 perfect-hedge benchmark — a basis-risk cost of exactly \$4,000.00 ($= 0.40 \times 10{,}000$), the price of a hedge that removed the overwhelming majority, but not all, of the exposure it was built to remove.

### Futures as price discovery

Where a future and its underlying both trade continuously, information gets impounded into both prices at close to the same rate, and neither structurally leads the other. Where the underlying's own market is constrained — closed outside limited hours, thinner, or simply slower to process new information than a deep, leveraged, centrally cleared future — the future routinely absorbs new information first, with the cash market catching up once it reopens or its own liquidity responds (Hasbrouck, 2007). Index futures are the cleanest illustration this book has already

built in full. Chapter 9's overnight-gap decomposition split a cash index's opening move into an overnight component — attributable to the near-continuous future (and the CFD tracking it) moving on news the cash market, closed, structurally could not yet price — and a smaller residual opening-auction component reflecting only what happened after the cash market itself reopened. The overnight component is exactly this section's price-discovery lead, made concrete and measured; the index CFD inherits it not because the provider engineers anything clever, but simply because it is priced off the future rather than off the cash index directly (Chapter 9), and the future is where the night's information was actually discovered first.

## 10.7 Exercises

**Exercise 10.1 — Tick and multiplier arithmetic on two hypothetical contracts**

Two further hypothetical contracts, unrelated to any real exchange's current listing. **Contract Gamma**, a hypothetical short-term-interest-rate-style future, has multiplier $m =$ \$2,500 per full point and tick size $\tau = 0.005$ points.

(a) Compute the tick value.

(b) Its price moves from 96.190 to 96.235. Compute the move in ticks, and the P&L on 8 contracts held long.

**Contract Delta**, a hypothetical soft-commodity future, has multiplier $m = 10{,}000$ units and tick size $\tau =$ \$0.0001 per unit.

(a) Compute the tick value.

(b) Its price moves from 0.6835 to 0.6791. Compute the move in ticks, and the P&L on 12 contracts held short.

▷ working space

**Solution — Exercise 10.1**

**Gamma.** (a) $v_\tau = 0.005 \times \$2{,}500 = \$\mathbf{12.50}$. (b) $\Delta P = 96.235 - 96.190 = 0.045 = 9$ ticks. Long 8 contracts: P&L $= 0.045 \times 2{,}500 \times 8 = \$\mathbf{900.00}$ (matches $9 \times \$12.50 \times 8 = \$900.00$).
**Delta.** (a) $v_\tau = 0.0001 \times 10{,}000 = \$\mathbf{1.00}$. (b) $\Delta P = 0.6791 - 0.6835 = -0.0044 = -44$ ticks (a fall). Short 12 contracts gain on the fall: P&L $= -(-0.0044) \times 10{,}000 \times 12 = \$\mathbf{528.00}$ (matches $44 \times \$1.00 \times 12 = \$528.00$).

**Exercise 10.2 — A full margin sequence: short position, a call, and a top-up decision**

A hypothetical energy future has multiplier $m = \$500$ per point, initial margin IM $= \$4{,}200$ per contract and maintenance margin MM $= \$3{,}150$ per contract. A trader sells (opens short) 2 contracts at $F_0 = 74.60$, posting total IM for the position, and holds through five subsequent settlements: 75.10, 75.85, 76.80, 76.35, 76.00.

(a) Compute total IM and total MM for the 2-contract position.
(b) Build the day-by-day variation margin and balance table, exactly as in Worked Example 10.2, and identify the day (if any) a margin call is triggered and the exact top-up required.
(c) Reconcile the total P&L two ways: sum of daily VM, and final balance minus total cash deposited.
(d) Compute the position's notional and leverage at entry.

▷ working space

**Solution — Exercise 10.2**

**(a)** Total IM $= 2 \times 4{,}200 =$ \$**8,400.00**; total MM $= 2 \times 3{,}150 =$ \$**6,300.00**.
**(b)** A short position's VM is $-(F_t - F_{t-1})\, m\, N$: it loses when price rises.

| Day | Settle | $\Delta F$ | VM (\$) | Balance (\$) | Action |
|---|---|---|---|---|---|
| 0 | 74.60 | — | — | 8,400.00 | post IM |
| 1 | 75.10 | +0.50 | −500.00 | 7,900.00 | none (>MM) |
| 2 | 75.85 | +0.75 | −750.00 | 7,150.00 | none (>MM) |
| 3 | 76.80 | +0.95 | −950.00 | 6,200.00 | **call**: top up +2,200.00 |
| | | | | 8,400.00 | (balance restored to IM) |
| 4 | 76.35 | −0.45 | +450.00 | 8,850.00 | none |
| 5 | 76.00 | −0.35 | +350.00 | 9,200.00 | none |

Day 3's settlement takes the balance to \$6,200.00, below MM = \$6,300.00: a call for 8,400.00 − 6,200.00 = \$**2,200.00**, restoring the balance to IM exactly.
**(c)** Sum of daily VM: $-500.00 - 750.00 - 950.00 + 450.00 + 350.00 = -\$\mathbf{1{,}400.00}$. Total deposited = 8,400.00 + 2,200.00 = \$10,600.00; final balance \$9,200.00; final balance minus total deposited = 9,200.00 − 10,600.00 = −\$**1,400.00**, matching, and matching the direct check $-(76.00 - 74.60) \times 500 \times 2 = -1.40 \times 1{,}000 = -\$1{,}400.00$: the short lost money as price rose over the five days, exactly as the position's direction predicts.
**(d)** Notional $= F_0\, m\, N = 74.60 \times 500 \times 2 =$ \$**74,600.00**; leverage = 74,600.00/8,400.00 = **8.88:1** (two decimal places) — lower than Worked Example 10.2's 37.5:1 purely because this hypothetical contract's IM is a much larger fraction of its notional, illustrating that leverage is a consequence of the margin schedule set for a given contract, never a fixed property of "a futures position" in general.

**Exercise 10.3 — Roll P&L under both curve shapes**

A hypothetical grain-style future, "Contract U", has multiplier $m = 600$ per contract. A trader holds $N = 10$ contracts long in the expiring contract, last relevant at $F_1 = 428.25$.

(a) The next contract trades at $F_2 = 436.10$ (contango). Compute the cash cost of the roll, the per-unit roll drag $\delta$, and confirm $\delta \times$ notional reconciles with the cash cost.
(b) Instead, the next contract trades at $F_2 = 421.90$ (backwardation, same $F_1$). Repeat (a).
(c) In one sentence, state which of the two curve shapes pays the long trader to roll, and why.

▷ working space

**Solution — Exercise 10.3**

Notional at $F_1$ in both parts: $N\,m\,F_1 = 10 \times 600 \times 428.25 = \$\mathbf{2{,}569{,}500.00}$.
**(a) Contango.** $\Delta F = 436.10 - 428.25 = +7.85$. Cash cost $= 10 \times 600 \times 7.85 = \$\mathbf{47{,}100.00}$ (paid). $\delta = (436.10 - 428.25)/428.25 = 0.018330 = \mathbf{1.8330}\%$. Check: $0.018330 \times 2{,}569{,}500.00 = \$47{,}100.00$, matching exactly.
**(b) Backwardation.** $\Delta F = 421.90 - 428.25 = -6.35$. Cash flow $= 10 \times 600 \times (-6.35) = -\$\mathbf{38{,}100.00}$, i.e. a *receipt* of \$38,100.00. $\delta = (421.90 - 428.25)/428.25 = -0.014828 = -\mathbf{1.4828}\%$. Check: $-0.014828 \times 2{,}569{,}500.00 = -\$38{,}100.00$, matching exactly.
**(c)** Backwardation pays the long: rolling from a dearer expiring contract into a cheaper next contract ($F_2 < F_1$) hands the long a cash receipt at every roll, the mirror image of contango's cash cost, exactly as [Course Book §22.1]'s sign convention predicts.

**Exercise 10.4 — Curve classification and implied annualised carry: three curves**

Classify each curve as contango or backwardation and compute its implied annualised carry $c$ via (10.5).

(a) Curve X (hypothetical precious-metal future): $T_1 = 1$ month, $F_1 = 2{,}384.50$; $T_2 = 7$ months, $F_2 = 2{,}431.80$.
(b) Curve Y (hypothetical livestock/soft future): $T_1 = 2$ months, $F_1 = 156.20$; $T_2 = 5$ months, $F_2 = 149.85$.
(c) Curve Z (hypothetical rates-style financial future): $T_1 = 3$ months, $F_1 = 95.640$; $T_2 = 12$ months, $F_2 = 96.050$.

▷ working space

**Solution — Exercise 10.4**

**(a)** $F_2 > F_1$: **contango**. $\Delta T = 6/12 = 0.5$ years. $c = \ln(2{,}431.80/2{,}384.50)/0.5 = \mathbf{3.93}\%$ p.a.
**(b)** $F_2 < F_1$: **backwardation**. $\Delta T = 3/12 = 0.25$ years. $c = \ln(149.85/156.20)/0.25 = -\mathbf{16.60}\%$ p.a.
**(c)** $F_2 > F_1$: **contango**. $\Delta T = 9/12 = 0.75$ years. $c = \ln(96.050/95.640)/0.75 = \mathbf{0.57}\%$ p.a.
Curve X's near-4% carry and Curve Z's near-0.6% carry both sit in ranges a financing-driven curve can plausibly produce; Curve Y's sharply negative, double-digit implied carry is the signature of a tight physical market pricing scarcity today more highly than scarcity three months out, the convenience-yield channel dominating storage cost by a wide margin.

**Exercise 10.5 — Continuous-series construction around a backwardation roll**

A hypothetical natural-gas-style future, "Contract V", rolls from front month ($F_1 = 3.220$) to next month ($F_2 = 3.110$) — a backwardation curve. Raw settlement prices around the roll: Day $-2$: 3.185 (front); Day $-1$: 3.220 (front, $= F_1$; the next contract is simultaneously quoted at $F_2 = 3.110$); Day 0: 3.145 (next contract, first day used); Day $+1$: 3.098 (next).

(a) Build the unadjusted (spliced) series and its day-over-day differences.
(b) Build the back-adjusted series (additive convention) and its day-over-day differences.
(c) State the true economic move on Day 0, the phantom move the unadjusted series reports, and confirm the two differ by exactly the roll gap.

▷ working space

**Solution — Exercise 10.5**

Roll gap $= F_2 - F_1 = 3.110 - 3.220 = -0.110$.
**(a)** Unadjusted series: 3.185, 3.220, 3.145, 3.098. Differences: Day $-2 \to -1$: $+0.035$; Day $-1 \to 0$: $3.145 - 3.220 = \mathbf{-0.075}$; Day $0 \to +1$: $-0.047$.
**(b)** Back-adjusted (shift the pre-roll segment by the roll gap, $-0.110$): Day $-2$: $3.185 - 0.110 = 3.075$; Day $-1$: $3.220 - 0.110 = 3.110$ (now equal to $F_2$, continuous); Day 0: 3.145; Day $+1$: 3.098 (both unchanged, the current segment). Differences: Day $-2 \to -1$: $+0.035$ (unchanged from (a)); Day $-1 \to 0$: $3.145 - 3.110 = \mathbf{+0.035}$; Day $0 \to +1$: $-0.047$ (unchanged).
**(c)** The true move on Day 0, within the new contract, from its own roll-reference price $F_2$: $3.145 - 3.110 = +0.035$, exactly what the back-adjusted series recovers. The unadjusted series instead reports $-0.075$: a phantom *fall*, in a session the new contract actually *rose*. The discrepancy is $-0.075 - (+0.035) = -0.110$, exactly the roll gap $F_2 - F_1$: a backwardated roll manufactures a phantom drop of the same magnitude a contango roll (§10.5's own table) manufactures a phantom rise, confirming that the direction of the artefact always follows the curve's shape at the roll, never the underlying's actual path that day.

**Key idea — What to remember**

A future is a forward with every non-price term standardised by an exchange, and a central counterparty that novates every trade converts Chapter 1's bilateral OTC credit risk into a single, margined claim on the CCP itself — a guarantee made credible only by initial, maintenance and variation margin collected in cash at every settlement, never taken on trust. Open interest counts the resulting stock of matched long/short pairs; volume counts the flow of trades that built or unwound it; and Chapter 7's COT report is simply that same open interest, sliced by who is holding each side. A futures curve prices carry directly for financial underlyings and prices storage against convenience yield for physical ones, and every position held past one contract's expiry pays or receives the roll's cash consequence — real money, reconciled exactly against Chapter 22's roll-drag percentage, and phantom-jump poison to any backtest built on an unadjusted spliced series rather than a properly back-adjusted one. Basis converges to zero at expiry, by arbitrage or by construction, but a hedge lifted early, or struck on an imperfect underlying, is exposed to whatever basis risk remains in between. The next chapter turns to options, built on precisely this exchange-and-clearing chassis with a payoff bolted on top of the linear claim studied here; the one after returns to commodities, where this chapter's storage-and-convenience machinery finally meets the physical markets it was built to describe.

# 11

# Options Markets in Practice

*Chapter 10 took a standardised contract, a clearing house and a margin ledger and showed how a futures market turns a forward price into a daily cash flow. An option is the same institutional apparatus wrapped around a different economic object: not a fixed forward price but a* right, *and therefore a price for uncertainty itself. The Course Book has already derived what that price is — the Black–Scholes–Merton value, the Greeks that are its derivatives, the implied volatility that inverts it, and the exact accounting by which a delta-hedged option pays realised minus implied variance [Course Book §20.1-20.5]. None of that is repeated here. This chapter is about the* market*: the chain a screen actually shows, the unit in which professionals think and quote, the market makers who are on the other side of every retail fill and what their hedging does mechanically to the underlying, the surface read as a live object rather than a fitted function, the cost stack in the only unit that makes option costs comparable across instruments, and the specific, structural, entirely predictable way that buying options into a scheduled event separates retail accounts from their money. The organising idea is a single change of units. Options are quoted in premium and traded in volatility, and almost every practical error in this market — mispricing a spread, mistaking a cheap wing for a bargain, buying an event that is already in the price — is a failure to convert between the two.*

## 11.1 The Listed Landscape

A futures market is one contract per expiry. An options market on the same underlying is a two-dimensional grid: for every listed expiry, a ladder of strikes, and at every strike a call and a put. That grid is the *chain*, and its size is the first structural fact about the market. A single underlying with a dozen listed expiries and forty strikes per expiry carries around a thousand separate contracts, each with its own bid, its own offer, its own open interest and its own liquidity — against one underlying with one bid and one offer. The consequences run through everything that follows: liquidity per contract is necessarily thin because it is spread across a thousand lines; a market maker cannot quote a thousand independent opinions, so the quotes must be generated from a small number of parameters; and the object the market genuinely trades is not any individual line but the surface those parameters describe.

**Definition — The chain**

The **chain** for an underlying is the full set of listed contracts, organised by expiry and, within each expiry, by strike. Each row conventionally shows, for one strike, the call's bid/ask/last/volume/open interest, the put's, and one or more derived columns — most importantly the **implied volatility** at which each side is quoted. **Open interest** is the number of contracts currently outstanding at that line, which is a stock, not a flow, and is the quantity dealer-positioning estimates are built from. The set of strikes and expiries is set by the exchange under published listing rules and expands as the underlying moves and as time passes.

### Expiry cycles

Expiries are listed on a schedule rather than continuously, and the schedule has a definite shape. A monthly cycle anchors the market: one expiry per calendar month, out to a year or more, with quarterly expiries carried furthest into the future because that is where hedgers' horizons and index-futures expiries line up. Onto that skeleton exchanges have progressively grafted shorter-dated contracts: weekly expiries on liquid underlyings, and on the most active index products expiries on multiple days of each week, which mechanically produces contracts that expire on the day they are traded — the *zero-days-to-expiry* or *0DTE* contracts. There is nothing conceptually new in a 0DTE option; it is an ordinary option with $T$ measured in hours. What is new is its market: because both gamma and theta of an at-the-money contract diverge like $1/\sqrt{T}$ as expiry approaches [Course Book §20.3], a 0DTE book carries convexity and decay per unit of time that no longer-dated book can match, and the hedging flow it generates is correspondingly concentrated. Section 11.3 makes that mechanical.

The number of expiries listed, the days of the week they fall on, the last trading day and the determination of the settlement value are exchange decisions that change, and have changed repeatedly during the 2020s. Treat the *shape* above as structure and every specific date, cycle or cut-off as something to read off the current contract specification.

### Multipliers, and why the unit matters twice

An option contract is not a claim on one unit of the underlying. It carries a *multiplier* $M$: the payoff and the premium are both scaled by it, so a quoted premium of 2.05 on a contract with $M = 100$ costs \$205, and a one-point move in the premium is \$100 of profit and loss. Equity option contracts have conventionally been written on a round lot of shares; index option contracts specify a currency amount per index point. Both conventions are contract-specific and both have exceptions — adjusted contracts after corporate actions, mini and micro variants, index products with different multipliers on the same index. Verify the current specification before sizing anything.

**Trap — The multiplier enters the position twice**

A wrong multiplier corrupts two independent calculations, and the second is easy to miss. It scales the *cash* — premium paid, margin required, profit and loss per point — which is the error everyone checks for. It also scales every *Greek*: a contract's delta, gamma and vega as reported by a pricing library are usually per unit of underlying, and the position's exposures are those numbers multiplied by $M$ and by the number of contracts. Getting the cash right and the Greeks wrong produces a book that appears correctly sized and is hedged by a factor of $M$ too little. Recompute both from the same stored multiplier, captured with a timestamp, and never from memory.

### Exercise style, settlement, and the clearing house

Two structural distinctions determine what actually happens to a position at the end of its life, and they are the ones most often confused.

**Exercise style** governs *when* the holder may exercise: an American option at any time up to expiry, a European option only at expiry. This matters to the short far more than to the long, because it is the short who receives an assignment without warning. **Settlement method** governs *what* changes hands: a physically settled contract delivers the underlying, so exercise creates or extinguishes a position in shares that must be paid for or borrowed; a cash-settled contract pays the difference between an official settlement value and the strike, so nothing is delivered and the position simply disappears. The two distinctions are logically independent, though in practice they are correlated: single-name equity options are typically American and physically settled, and broad index options are typically European and cash-settled (Hull, 2021). The Course Book states

the same landscape and the same instruction to confirm each contract's style from the exchange [Course Book §20.6]; the market consequence worth adding here is that an American, physically settled short is a *contingent share position*, and must be risk-managed as one from the day it is opened.

Behind every listed contract sits a clearing house that novates the trade: after clearing, the buyer's counterparty and the seller's counterparty are both the central counterparty, exactly as Chapter 10 described for futures. In the United States listed equity and index options clear through the Options Clearing Corporation; in Australia exchange-traded options clear through the ASX's clearing subsidiary. The economic content is identical to the futures case — bilateral credit risk is replaced by a mutualised exposure to the clearing house, supported by margin from members and a default fund — with one asymmetry specific to options. The buyer of an option pays the premium in full and thereafter has no further obligation, so the buyer's residual credit exposure is to the clearing house only. The seller has an open-ended obligation and is margined continuously against it, on a risk basis that rises as the position moves against them. Long option positions therefore consume capital once; short option positions consume it continuously and unpredictably, which is a funding fact before it is a risk fact and belongs in the same mental category as Chapter 10's variation-margin discipline.

### Australia and the United States, structurally

The reader's local venue and the deepest options market in the world differ in ways that are structural rather than numerical. The ASX lists exchange-traded options over individual listed companies, American-style and deliverable, alongside index options over the S&P/ASX 200, which are European and cash-settled. The US complex is far larger and layered differently: index options on the major benchmarks, an enormous ETF option market that gives index-like exposure in an American, physically deliverable wrapper, and single-name options across thousands of underlyings. Three differences matter to a systematic account.

*Depth.* The Australian market is materially smaller in both volume and quoted depth, which does not merely mean wider spreads: it means fewer strikes and expiries at which a two-sided market of usable size exists at all, and therefore a smaller set of structures that can be entered and, crucially, exited. *Wrapper.* An index-like exposure obtained through an American, deliverable ETF option carries assignment and delivery mechanics that the equivalent European, cash-settled index option does not; the payoff diagrams are near-identical and the operational risk is not. *Session.* Options trade during the underlying's session, so an Australian book on Australian underlyings and an Australian book on US underlyings are two different operational businesses, with different hedging windows and different overnight gap exposure — the same session arithmetic Chapter 9 built for index CFDs, now applied to a position whose delta changes while the market that could hedge it is shut.

## 11.2 How Options Are Quoted, and What Actually Trades

A screen shows premium. A dealer thinks in volatility. That sentence is the whole of this section, and unpacking it correctly is the single most useful change of habit an options novice can make.

The reason is not preference but comparability. A premium is a price in the units of the underlying, contaminated by everything the underlying is doing: the same contract's premium changes when spot moves, when a day passes, when a dividend is declared, and only lastly when the market's view of future variability changes. Two options on the same underlying at different strikes have premiums that cannot be compared at all, because most of the difference is moneyness. Implied volatility strips all of that out. It is the number that, put into the pricing formula together with the observable spot, strike, tenor and rate, reproduces the quoted premium [Course Book §20.4]; because it is unique — vega is strictly positive, so the map from volatility to price is strictly increasing — reading a premium and reading its implied volatility are the same act performed in two coordinate systems. In the volatility coordinate a 16-delta put in three months

and a 50-delta call in one week are directly comparable objects, and a dealer's inventory is a position in a single scalar field over strike and tenor rather than a heap of unrelated line items.

**Definition — The traded quantity**

A listed option's **implied volatility** $\sigma_I$ is the volatility input at which the model reproduces the observed premium. In an options market, $\sigma_I$ is the quantity that is negotiated, risk-managed and hedged; the premium is the settlement convention in which the agreed $\sigma_I$ is paid. Interdealer markets in currency options quote directly in volatility; listed equity and index markets quote in premium, and every participant with a screen converts.

### The conversion, and the only Greek needed for it

Vega, $\mathcal{V} = \partial V / \partial \sigma$, is the derivative of premium with respect to volatility [Course Book §20.3]. Quoted per *volatility point* — one percentage point of implied volatility, the market's working unit — it is the exchange rate between the two coordinate systems, and the conversion is a division.

**Result — Premium and volatility points**

For a position whose vega is $\mathcal{V}$ units of premium per volatility point, any quantity of premium $\Delta P$ converts to volatility points as

$$\Delta\sigma \;=\; \frac{\Delta P}{\mathcal{V}}, \qquad \text{and back as} \qquad \Delta P \;=\; \mathcal{V}\,\Delta\sigma. \tag{11.1}$$

Applied to a quoted market of width $W$ in premium, the width in volatility points is $w = W/\mathcal{V}$, and crossing it once costs $w/2$.

**The unit discipline.** Both sides of (11.1) must refer to the *same object*. Convert a single leg's premium with that leg's vega; convert a package's premium with the package's vega. Mixing them — converting a leg's spread with a two-leg straddle's vega, the commonest slip — misstates the cost by exactly the ratio of the vegas.

**Consequence (the round-trip identity).** Suppose every leg of a structure is quoted $w$ volatility points wide and the structure's vega is the sum of its legs' vegas in the same sign (a straddle, a strangle: all legs bought or all sold). Entering and exiting crosses each leg once in each direction. With $n$ legs, the round trip pays $2n$ crossings of $\frac{w}{2}\mathcal{V}_{\text{leg}}$ each, a total of $n\,w\,\mathcal{V}_{\text{leg}}$ in premium, against a package vega $n\mathcal{V}_{\text{leg}}$; so the round-trip cost, in volatility points of the package, is

$$c_{\text{rt}} \;=\; \frac{n\,w\,\mathcal{V}_{\text{leg}}}{n\,\mathcal{V}_{\text{leg}}} \;=\; w. \tag{11.2}$$

*A round trip costs one full quoted width, whatever the number of legs* — provided the legs' vegas add rather than cancel. Section 11.5 shows what happens when they cancel, which is the entire economics of defined-risk structures.

Equation (11.2) is the options analogue of the round-trip spread arithmetic Chapter 3 built for CFDs, and it carries the same warning: the number that matters is the width relative to the edge, not the width in isolation. A market 0.4 volatility points wide is tight if the forecast edge is three points and ruinous if the edge is half a point. The Course Book's cost table makes exactly this point in index points [Course Book §20.6]; the version in (11.2) is unit-free and therefore transfers between an index at 5,000 and a share at $62.50 without rescaling.

**Worked Example 11.1 — Two markets that look alike in dollars and are nothing alike in volatility**

Two hypothetical quotes, both of them ordinary-looking, illustrate why the conversion is not optional. All values are illustrative scenarios, not market data.

**Part A — a listed index option.** A hypothetical index stands at $S = 8{,}000.00$ points. Its 30-day at-the-money call is quoted 96.00 bid, 100.80 offered, in index points, and the contract's vega is given as $\mathcal{V} = 9.60$ index points per volatility point. The mid implied volatility is 18.00. Then

$$W = 100.80 - 96.00 = 4.80 \text{ index points}, \qquad w = \frac{4.80}{9.60} = \mathbf{0.50} \text{ volatility points},$$

so the market is **17.75** bid at **18.25** offered in the unit that matters, and a single crossing costs 0.25 volatility points, or 2.40 index points. On a hypothetical multiplier of \$10 per index point the quoted width is \$48.00 and the mid premium is \$984.00 per contract.

**Part B — a listed single-name option.** A share trades at \$62.50. Its 45-day call is quoted \$2.05 bid, \$2.25 offered, per share, with vega given as $\mathcal{V} = \$0.0875$ per share per volatility point. On a contract of 100 shares that is a vega of \$8.75 per volatility point and a quoted width of \$20.00. In volatility points,

$$w = \frac{0.20}{0.0875} = \mathbf{2.285714} \text{ volatility points},$$

a crossing costing 1.142857 volatility points.

**Part C — the comparison, and why the dollar view inverts it.** In premium the index option's market is 4.80 points wide and the single name's is twenty cents wide, which invites the conclusion that the share option is the tighter market. In the traded unit the ratio is

$$\frac{2.285714}{0.50} = \mathbf{4.5714}\times,$$

so the single-name market is more than four and a half times as expensive to cross. Nothing about the dollar figures could have told you that, because they are denominated in different things.

**Part D — the round trip.** Take the index strike as a straddle: two legs, each 0.50 volatility points wide, leg vega 9.60, package vega 19.20. Each crossing costs $4.80/2 = 2.40$ index points; a round trip crosses four times, for 9.60 index points; and

$$c_{\text{rt}} = \frac{9.60}{19.20} = \mathbf{0.50} \text{ volatility points } = w,$$

confirming (11.2). The single-name straddle's round trip is $4 \times (0.20/2) = \$0.40$ per share against a package vega of \$0.175, or 2.285714 volatility points — again exactly $w$. Read those two numbers as thresholds on a forecast: to break even before any hedging cost, slippage or forecast error, the index trade must beat the market's own volatility estimate by half a point, and the single-name trade must beat it by more than two and a quarter points. Chapter 12 of the Course Book's engine is a good forecaster; it is not routinely two and a quarter points better than the market [Course Book §12.6].

### Moneyness conventions: why currency desks quote strikes in delta

Listed equity and index markets identify a contract by its strike price, which is fixed for the contract's life while spot moves around it, so a strike that was at the money last month is a

different animal today. Over-the-counter currency option desks solve this differently: they identify strikes by *delta*. A "25-delta put" is whatever strike currently has a delta of $-0.25$, and it stays the 25-delta put as spot moves, which keeps quotes comparable across days and across pairs. On that convention a whole smile is transmitted as three numbers per tenor — the at-the-money volatility, the 25-delta risk reversal (its slope) and the 25-delta butterfly (its curvature) — from which the wing volatilities are reconstructed; the Course Book gives the inversion and the conventions it hides, including whether the delta is spot or forward and whether it is premium-adjusted [Course Book §20.4]. Two market consequences follow. First, delta-quoting means a currency vol book's positions are naturally expressed in smile parameters, so a currency desk's risk report reads as level, slope and curvature rather than as a list of strikes. Second, any pipeline ingesting currency vol quotes must record the convention alongside the number, because the same quoted "25-delta risk reversal" means different strikes under different conventions, and the discrepancy is systematic rather than noisy [Course Book §18.2].

## 11.3 The Market Makers and Their Hedging Flows

Chapter 2 ran the market maker's business in a single-instrument order book: quote two-sided, end flat, earn the spread on round trips, and lose it to adverse selection and to whatever inventory is still on the book when something happens. An options market maker runs the same business under two additional constraints that change its character completely.

The first is dimensionality. The maker is not quoting one instrument but the thousand-line chain of Section 11.1, and cannot form a thousand independent opinions. The quotes are therefore generated: a small number of surface parameters — level, slope, curvature, term structure — are fitted to whatever genuine two-way interest exists, and every line's bid and offer is computed from the surface plus a width that depends on the line's liquidity and the maker's inventory. This is why option quotes move in sheets. A trade in one strike updates the surface and repaints the neighbouring strikes without any of them having traded, which is a mechanism a naive backtest that treats each line as an independent time series will misread entirely.

The second is that the risk is multi-dimensional. A share market maker's inventory risk is one number, and Chapter 2's skewing rule moves one quote to shed it. An option maker who buys a call acquires delta, gamma, vega, theta and a set of higher-order exposures simultaneously, and these cannot be shed by skewing alone because they live on different axes. The response is a division of labour that defines the whole market's mechanics:

> **Key idea — The market maker hedges what he does not want and keeps what he is paid for**
>
> The delta is hedged *immediately*, in the underlying or its future, because delta is directional exposure the maker has no view on and is not compensated for carrying. The gamma, vega and theta are *retained* — they are the volatility position, and the spread the maker earns is compensation for warehousing exactly that. Every option trade in the listed market therefore generates a near-simultaneous, mechanically determined, same-day trade in the underlying. The options market is a machine for converting volatility opinions into delta-hedging order flow.

That last sentence is the reason this section exists. A retail account buying a call is not merely transferring premium to a dealer; it is causing a purchase of the underlying within seconds, of a size the account can compute. Aggregated across a whole chain and a whole street, the flow this generates is large, mechanical, and — crucially — *signed by the sign of the street's aggregate gamma*.

### Deriving the flow from the sign of gamma

The derivation is short and worth doing carefully, because the direction is easy to get backwards and the market commentary that discusses it frequently does.

**Result — Hedging-flow direction**

Let a dealer's aggregate position have delta $\Delta_{\text{pos}}$ and gamma $\Gamma_{\text{pos}} = \partial\Delta_{\text{pos}}/\partial S$, both measured in units of the underlying per unit move in $S$, and suppose the dealer is delta-neutral at spot $S$. After a move $\delta S$, the position's delta becomes

$$\Delta_{\text{pos}}(S + \delta S) \;\approx\; \Gamma_{\text{pos}}\,\delta S, \tag{11.3}$$

so to return to neutrality the dealer must trade

$$q_{\text{hedge}} \;=\; -\,\Gamma_{\text{pos}}\,\delta S \tag{11.4}$$

units of the underlying (positive $q$ meaning *buy*). Two regimes follow directly from the sign of $\Gamma_{\text{pos}}$:

**Long gamma** ($\Gamma_{\text{pos}} > 0$). $\delta S > 0 \Rightarrow q_{\text{hedge}} < 0$: the dealer *sells* into a rally. $\delta S < 0 \Rightarrow q_{\text{hedge}} > 0$: the dealer *buys* the dip. The hedging flow is *counter* to the move and therefore **dampening**; realised volatility in the underlying is mechanically suppressed, and the dealer is paid for this in gamma while paying for it in theta.

**Short gamma** ($\Gamma_{\text{pos}} < 0$). $\delta S > 0 \Rightarrow q_{\text{hedge}} > 0$: the dealer *buys* into a rally. $\delta S < 0 \Rightarrow q_{\text{hedge}} < 0$: the dealer *sells* into a decline. The hedging flow is *with* the move and therefore **amplifying**; the dealer collects theta and pays for it by being forced to trade at progressively worse prices in whichever direction the market travels.

The asymmetry in how these two regimes *feel* is worth stating. A long-gamma street produces a market that grinds: moves are absorbed, ranges hold, and realised volatility comes in under implied. A short-gamma street produces a market that trends within the day and gaps at the edges, because the hedging flow reinforces the move and the reinforcement is largest exactly where gamma is largest — near the big strikes, near expiry. This is a genuine positive-feedback loop of the same family as the mechanical selling Chapter 15 will examine in the 1987 crash and the funding-liquidity spiral of Brunnermeier and Pedersen (2009): a hedging rule that is individually rational and collectively destabilising.

**Worked Example 11.2 — The street's book, and what a one per cent move makes it do**

A hypothetical index stands at $S = 5{,}000.00$ points; its options carry a hypothetical multiplier of \$25 per index point. Three lines dominate the chain's open interest thirty days before expiry. The gamma per contract is the contract's $\Gamma$ in contracts of delta per index point of spot; the "dealer side" column is an *assumption* about who holds which side, discussed below.

| Strike | Type | Open interest | $\Gamma$ per contract | Dealer side | Contracts of $\Delta$ per point |
|---|---|---|---|---|---|
| 4,900 | put | 150,000 | 0.00121 | short | −181.50 |
| 5,000 | call | 48,000 | 0.00155 | long | +74.40 |
| 5,100 | call | 55,000 | 0.00138 | long | +75.90 |
| | | | | **Net** | **−31.20** |

**Step 1 — aggregate the gamma.** Each line contributes $\pm(\text{open interest}) \times \Gamma$, signed by the dealer's side:

$$-150{,}000(0.00121) + 48{,}000(0.00155) + 55{,}000(0.00138) = -181.50 + 74.40 + 75.90 = \mathbf{-31.20}$$

contracts of delta per index point. The street is **short gamma**. Expressed as dollar gamma — the practitioner's unit, $\Gamma S^2 M$, which is the same object as the Course Book's $\Gamma S^2$ scaled by

the multiplier [Course Book §20.5] —

$$-31.20 \times 5{,}000^2 \times 25 \;=\; -\$19.50 \text{ billion},$$

and the conventional "gamma per one per cent" figure is one hundredth of that, \$**195.0** million of delta per 1% move.

**Step 2 — the flow on a $+1\%$ move.** A 1% rally is $\delta S = +50.00$ points. By (11.3) the street's delta changes by $-31.20 \times 50 = -1{,}560$ contracts, so by (11.4) dealers must **buy** 1,560 **contracts** to restore neutrality. At the post-move level of 5,050 that is

$$1{,}560 \times 5{,}050 \times 25 \;=\; \$\mathbf{196.95} \text{ million}$$

of index bought *because* the index rose.

**Step 3 — the flow on a $-1\%$ move.** $\delta S = -50.00$ changes the delta by $+1{,}560$ contracts, so dealers must **sell** 1,560 **contracts**, worth $1{,}560 \times 4{,}950 \times 25 = \$193.05$ million at the post-move level. Short gamma sells declines and buys rallies: the flow is same-signed with price in both directions, which is what "amplifying" means operationally. Note that the \$195.0 million per-one-per-cent figure sits exactly between the two, and differs from each by \$1.95 million; the gap is the second-order effect of valuing the same 1,560 contracts at a different index level, and it is a reminder that the per-one-per-cent number is a local slope, not a constant.

**Step 4 — flip the sign of the book.** Had the same three lines been held the other way round — dealers long the put and short the calls — the net gamma would be $+31.20$, and every flow above reverses: sell 1,560 contracts into the $+1\%$ rally, buy 1,560 into the $-1\%$ decline. Same open interest, same gammas, opposite market impact. The entire sign of the effect rests on the "dealer side" column.

**Step 5 — what does not scale.** A 2% move gives $-31.20 \times 100 = -3{,}120$ contracts of delta, twice the flow, *if* $\Gamma$ *is unchanged*. It is not: as spot travels through 5,100 the 5,100 line's gamma peaks and falls away, the 4,900 line's collapses, and new strikes come into play. The aggregate gamma is a function of spot, not a number, and the honest version of this calculation is a profile over spot rather than a single figure.

### What is observable, and what is assumed

Open interest is published. Contract gammas are computable from observable quotes. The *dealer side* column is neither: exchanges publish how many contracts are outstanding, not who is long and who is short. Every published estimate of "street gamma" therefore rests on an assumption — typically that customers buy index puts for protection and sell calls for yield, so dealers are short puts and long calls — and Step 4 of Worked Example 11.2 shows that this assumption does not merely scale the answer, it determines its sign. The assumption is a reasonable central case for broad index options, where the protective-put demand is well documented and is the same demand that produces the skew of Section 11.4. It is much weaker for single names, weaker still for short-dated contracts where two-way speculative flow dominates hedging flow, and it fails outright in episodes where the customer base flips. Use the framework to reason about mechanism; do not treat a gamma-exposure number as data [Course Book §18.2].

### Pinning, honestly

A tendency has been documented for underlyings to close near strikes with large open interest on expiry days, and the hedging mechanism above supplies a candidate explanation: a long-gamma dealer near a large strike sells the underlying above it and buys it below it, and that is

exactly a restoring force towards the strike. Three caveats have to travel with the claim. The mechanism requires the street to be *long* gamma at that strike, which by the previous subsection is an assumption and not an observation, and short-gamma positioning at a large strike predicts the opposite — repulsion, not attraction. The effect that has been documented is a statistical tendency in the distribution of closing prices, of a magnitude that is small relative to a typical day's range, not a reliable attractor a position can be sized against. And expiry-day price formation involves several simultaneous mechanisms — hedge unwinds, the official settlement procedure, index-rebalance flow — so attributing an observed close to pinning is an identification problem, not a reading. State it as a tendency with a plausible mechanism; do not build a strategy on it without the full Chapter 15 apparatus of the Course Book applied to however many strike-and-date variants were tried [Course Book §15.4].

### Why 0DTE concentrates all of this

Everything above scales with gamma, and at-the-money gamma scales like $1/\sqrt{T}$ [Course Book §20.3]. A contract with hours to live therefore carries, per unit of open interest, gamma an order of magnitude larger than a monthly contract on the same strike, and its hedging flow arrives over hours rather than weeks. Three structural consequences follow, none of which requires a single number to state. The hedging flow generated by same-day expiries is concentrated into the session rather than spread across the contract's life. Because those contracts expire the same day, the flow does not accumulate into a standing position that persists overnight — it is created and extinguished within the session, which is why its imprint is intraday rather than multi-day. And because gamma is largest precisely at the money, the flow's intensity is highest when the index is sitting on a heavily traded strike, which is where an intraday range most often either holds or breaks decisively. The Course Book's warning about the retail short-0DTE "income" claim stands unchanged and is the other half of the same picture [Course Book §20.6].

## 11.4 The Surface as a Market Object

The Course Book treats the volatility surface as a mathematical object: a function $\sigma_{\text{IV}}(K, T)$ constrained by no-arbitrage, fitted rather than interpolated, with SVI as the workhorse parameterisation [Course Book §20.4]. A trading floor treats it as a market object with a price, a bid-ask, an inventory and a set of participants whose demand explains its shape. This section reads it that way.

### Reading the implied volatility column

The one column of a chain a professional looks at first is the implied volatility column, and the reading has three layers, taken in order.

**The level.** The at-the-money volatility of the front expiry is the market's price for near-term variability. It is the number to compare against a realised-volatility forecast, and comparing anything else against a realised-volatility forecast is a units error.

**The slope across strikes — the skew.** On equity indices the implied volatility column falls monotonically as strike rises: downside puts are quoted at materially higher volatility than upside calls. The shape is not a modelling artefact. It is a demand curve. Portfolios that are structurally long equities want protection against the downside and do not want protection against the upside, so there is persistent one-way demand for out-of-the-money puts and persistent supply of out-of-the-money calls (from the same portfolios, selling upside for income). A dealer absorbing that flow ends up short downside options and long upside options, and prices the imbalance exactly as Chapter 2's inventory-skewing argument says a dealer must: by moving the price of the thing he is accumulating. The skew is the standing inventory imbalance of the whole street, converted into a price. Gatheral (2006) treats the shape's mathematics; the reason it is a *persistent* feature rather than a fluctuating one is that the demand producing it is persistent.

**The slope across expiries — the term structure.** At a fixed moneyness, implied volatility varies with tenor. Two mechanisms dominate. Volatility mean-reverts, so when near-term realised volatility is unusually high the front expiries price above the back ones and the term structure inverts, and when markets are calm the front prices below the back and the term structure is upward-sloping — the same mean reversion the Course Book's GARCH and HAR machinery estimates from the other side [Course Book §12.3]. Superimposed on that smooth shape are *bumps*: expiries that contain a scheduled event price above their neighbours by an amount that is nothing to do with mean reversion and everything to do with a known date.

### Event volatility, isolated

The bump is the single most tradeable-looking and most misunderstood feature of the surface, and it can be extracted arithmetically. The reasoning is entirely additive-in-variance, which is the right coordinate: variances of independent contributions add, volatilities do not.

**Result — Decomposing an expiry's variance into diffusion and event**

Let an expiry have $T$ years of life and quoted at-the-money implied volatility $\sigma_f$, and suppose the market's view of ordinary day-to-day (diffusive) volatility over the same window, absent any scheduled event, is $\sigma_b$. Treating the event's contribution as an independent one-off jump of variance $V_e$ added to the diffusive variance,

$$\underbrace{\sigma_f^2\,T}_{\text{total variance}} \;=\; \underbrace{\sigma_b^2\,T}_{\text{diffusive}} \;+\; \underbrace{V_e}_{\text{event}}, \qquad \text{so} \qquad \sigma_e \;:=\; \sqrt{V_e} \;=\; \sqrt{T\left(\sigma_f^2-\sigma_b^2\right)}. \tag{11.5}$$

$\sigma_e$ is a *one-off* standard deviation of proportional move, not an annualised rate: it is the size of the move the market is pricing for the event itself. The corresponding expected *absolute* move, which is the quantity the straddle price pays for, is

$$\mathbb{E}\left|\text{event move}\right| \;=\; \sqrt{\tfrac{2}{\pi}}\,\sigma_e \;\approx\; 0.7979\,\sigma_e \tag{11.6}$$

under a Gaussian jump, since $\mathbb{E}|Z| = \sqrt{2/\pi}$ for $Z \sim \mathcal{N}(0,1)$. The event's share of the expiry's total variance is $V_e/(\sigma_f^2 T)$.

**Convention warning.** Practitioners differ over whether the event day also contributes its ordinary diffusive variance or replaces it, and over whether $T$ counts calendar or trading days. The choice moves $\sigma_e$ by a few per cent of itself; record which convention a stored surface used [Course Book §18.2].

Equation (11.5) and the straddle price are two readings of the same number, which is the check that makes the framework trustworthy. An at-the-money straddle costs approximately $\sqrt{2/\pi}\,S\sigma\sqrt{T}$, so *straddle premium divided by spot* is the market's expected absolute move over the option's life — the "implied move" quoted before earnings. Extract $\sigma_e$ from the term structure by (11.5), extract the total expected move from the straddle price, and they must reconcile through the diffusive component. If they do not, one of the inputs is stale.

**Worked Example 11.3 — The event volatility crush**

A hypothetical single name trades at $S_0 = \$84.00$ and reports earnings after today's close. The front expiry has eight calendar days to run and its at-the-money implied volatility is 45.0; the market's diffusive baseline for the same window, read from the expiries either side of the event, is 28.0. Contracts are on 100 shares. Financing is taken as zero, which at eight days moves the premia below by less than 0.04%. This is an illustrative scenario computed

exactly, not market data.

**Step 1 — what the market is charging for the event.** With $T = 8/365 = 0.0219178$, equation (11.5) gives

$$\sigma_f^2 T = 0.45^2(0.0219178) = 0.004438356, \qquad \sigma_b^2 T = 0.28^2(0.0219178) = 0.001718356,$$

$$V_e = 0.002720000, \qquad \sigma_e = \sqrt{0.002720000} = 0.052154 = \mathbf{5.2154}\%,$$

a diffusive contribution of 4.1453% and a total of $\sqrt{0.052154^2 + 0.041453^2} = 6.6621\%$ over the option's life. The event alone accounts for **61.28**% of the expiry's total variance. The expected absolute event move is $0.7979 \times 5.2154\% = \mathbf{4.1613}\%$.

**Step 2 — price the straddle, and read the same number off it.** At $S = K = 84.00$, $T = 8/365$, $\sigma = 45.0$, the call and the put are each \$2.232133 per share, so the straddle is

$$V_0 = \$4.464267 \text{ per share } = \$\mathbf{446.43} \text{ per contract},$$

with straddle vega \$0.0991692 per share per volatility point, i.e. \$9.9169 per contract. The "implied move" quoted from the straddle is $4.464267/84.00 = \mathbf{5.3146}\%$, which is $0.7979 \times 6.6621\% = 5.3156\%$ to rounding — the expected absolute move over the whole eight days, event plus diffusion. The two readings agree.

**Step 3 — the morning after.** Earnings are released. The shares gap *up* 3.00% to $S_1 = \$86.52$ — a large overnight move by any ordinary standard — and the front expiry's implied volatility falls to 28.0, the diffusive baseline, because the event it was pricing has now happened. Seven days remain. Revaluing exactly,

$$\text{call} = \$2.944407, \quad \text{put} = \$0.424407, \quad V_1 = \$3.368815 \text{ per share} = \$\mathbf{336.88} \text{ per contract},$$

$$\text{P\&L } = \ 336.88 - 446.43 \ = \ -\$\mathbf{109.55} \text{ per contract } = \ -\mathbf{24.54}\%.$$

The underlying moved three per cent overnight and the position that needed only *movement* lost a quarter of its value.

**Step 4 — attribute it, exactly.** Revalue in sequence, changing one input at a time; the three steps then sum to the total with no residual.

| Step | Change applied | Straddle (per share) | Effect (per contract) |
|---|---|---|---|
| Start | $S = 84.00$, $\sigma = 45.0$, $T = 8/365$ | 4.464267 | — |
| Time | $T : 8 \to 7$ days | 4.176035 | −\$28.82 |
| Vol | $\sigma : 45.0 \to 28.0$ | 2.598680 | −\$157.74 |
| Spot | $S : 84.00 \to 86.52$ | 3.368815 | +\$77.01 |
| | | **Total** | −\$**109.55** |

The crush costs \$157.74; the three-per-cent gap earns back \$77.01; one day of decay costs \$28.82. The volatility term is roughly twice the gain from a move most traders would have called a good one. (A naive first-order Greek attribution — vega × 17 points, plus $\frac{1}{2}\Gamma(\delta S)^2$, plus one day of theta — gives −\$151.23, leaving a \$41.69 residual, because a 17-point change in volatility and a \$2.52 change in spot are far too large for local derivatives. Sequential revaluation is exact; Greek attribution is a story.)

**Step 5 — what the shares actually had to do.** Solve for the spot at which the post-event straddle is worth what was paid, under the post-event surface ($\sigma = 28.0$, seven days):

$$S^{\uparrow} = \$88.1149 \ (+4.8987\%), \qquad S^{\downarrow} = \$79.8137 \ (-4.9837\%).$$

The break-evens are almost five per cent away. The delivered move of three per cent was not close, and — the point of the whole example — it was never going to be enough on average, because the market's own expected absolute event move was 4.1613% and the buyer paid for that. To profit, the buyer needed not a big move but a move bigger than the one already in the price.

**Step 6 — the counterfactual that isolates the crush.** Had implied volatility stayed at 45.0 with everything else identical, the same straddle would be worth \$4.706548 per share and the position would have *made* \$24.23 per contract, +5.43%. The entire difference between a modest gain and a 24.54% loss is the volatility term. Buying this straddle was never a bet on the earnings move; it was a bet that the market's price for the earnings move was too low.

**Trap — The event is already in the price**

The retail strategy of buying options "because a catalyst is coming" has a structural, not a statistical, defect: the catalyst has a date, the date is public, and the option that spans it is priced with the event's variance already added. Worked Example 11.3 shows the consequence with the arithmetic done: an expiry whose implied volatility is 45 against a 28 baseline is charging 5.22% for the event and expects an absolute move of 4.16%; a buyer who is right that "the stock will move" and delivers 3% still loses a quarter of the premium, and even a 4.9% move only returns the stake. The crush is not a manipulation, a liquidity failure, or something that happens to unlucky traders — it is the term structure resolving on schedule, and it happens every time. Three defences follow, and only three. Have a view on the *size* of the event move relative to the priced $\sigma_e$ of (11.5), not on its direction or its existence. Compare that priced move against a properly constructed historical distribution of the same event type on the same underlying, with the sample size stated. And subject any apparent edge to the same deflation as every other strategy in the programme, because "buy the straddle before earnings" has been tested by everyone and the number of variants tried is enormous [Course Book §15.4].

### The volatility index, and what it is not

The best-known summary of an equity index's surface is a volatility index, of which the VIX is the archetype. Its construction is worth stating precisely because it is routinely described wrongly. It is *not* an average of Black–Scholes implied volatilities, and it is not the implied volatility of any traded option. It is a model-free estimate of the risk-neutral expected variance of the index over the next 30 days, computed from a strip of out-of-the-money option prices — puts below the forward, calls above it — weighted in proportion to $1/K^2$, and interpolated between the two expiries that bracket 30 days (Cboe, 2019). The $1/K^2$ weighting is not a convention: it is precisely the weighting that makes a static portfolio of options have constant dollar gamma, and therefore replicate a variance swap, which is the same construction the Course Book cites as the fix for the path-dependence of a single option's dollar-gamma weighting [Course Book §20.5].

Three consequences for a systematic user. Because it aggregates the whole out-of-the-money strip, a volatility index responds to the *wings* as well as the level, so it can rise when at-the-money volatility is flat if the skew steepens — it is not a clean proxy for at-the-money implied volatility and should not be regressed as if it were. Because it is a risk-neutral expectation it embeds the variance risk premium and is therefore a biased forecast of realised volatility, high on average [Course Book §20.5]. And because it is an index, it is not tradeable: exposure is obtained through futures and options on the index, which price the *forward* value of the index at their own expiry and therefore have their own term structure, their own roll and their own carry — the futures machinery of Chapter 10 applied to an underlying that cannot be delivered or arbitraged against a cash position.

## 11.5 Costs and Liquidity

Chapter 3 built a cost taxonomy for CFDs and insisted that no strategy is evaluated except net of a measured cost surface. Options obey the same rule with two aggravating factors: the spread is a far larger fraction of the available edge than it is in liquid FX, and the number of distinct lines means the surface has to be measured per strike and per tenor rather than per instrument.

### The shape of the width surface

Quoted width in volatility points varies systematically along three axes, and the pattern is structural even though the numbers are not.

*Moneyness.* Width is narrowest at the money and widens into the wings, for two independent reasons that compound. A wing option has small vega, so the same premium width divided by (11.1) is a larger number of volatility points; and wing options trade less, so the premium width is larger too. *Tenor.* Width in volatility points is narrowest in the front expiries where the two-way flow is, and widens into the back months. *Underlying.* Broad index options are the tightest markets, large single names materially wider, small and mid-caps wider again — the ratio between an index leg and a single-name leg computed in Worked Example 11.1 was more than four and a half to one, and larger ratios are ordinary. On top of all three, width is state-dependent in exactly the way Chapter 2's inventory-risk argument predicts: it widens at the open, into scheduled events, and in stress, which is to say precisely when a systematic strategy most wants to trade.

| **Tenor** | **At the money** | 25-**delta** | 10-**delta** |
|---|---|---|---|
| Weekly (≈ 7 days) | narrowest | wider | widest |
| Front month (≈ 30 days) | narrow | wider | widest |
| Three months (≈ 90 days) | wider than front | wider still | widest |

The table above is deliberately ordinal, because the cardinal version drifts. As an indication of magnitude only, and as of the mid-2020s: at-the-money width on the most liquid listed index options has commonly been of the order of a few tenths of a volatility point, wing width several times that, and single-name width a multiple of the index figure again. Those are orders of magnitude for sanity-checking a data feed, not inputs to a model. The instruction is unchanged from the Course Book's version of the same table [Course Book §20.6]: build the width surface from your own venue's recorded quotes, at the times you would actually trade, and verify current specifications.

**Trap — An options mid-price backtest is worse than an FX mid-price backtest, twice over**

The platform's mid-only candle problem is already known: a backtest run on mid prices earns an edge no account can capture, and the fix is a measured bid/ask cost surface. In options the same error is compounded. First by magnitude — half a volatility point of width against a two-point forecast edge is a quarter of the edge per crossing, four crossings per straddle round trip, so the mid-price version of an options backtest can be profitable when the executable version is not, at cost levels that are entirely ordinary. Second, and more insidiously, by *existence*: on an illiquid line the "mid" may not correspond to any price at which anything could trade, because one or both sides are quotes a market maker posts to satisfy an obligation rather than to transact, and a stale bid on a wing strike produces a mid that moves when nothing happened. An options research database must therefore store both sides of the quote with a capture timestamp, never a derived mid alone, and any line whose width exceeds a pre-registered threshold should be excluded from the tradeable universe before signals are computed rather than after [Course Book §18.3].

### Where the liquidity is, and what that does to structure choice

Volume and quoted depth concentrate near the money and in the front expiries, and thin out sharply in the wings and the back months — the same concentration the Course Book notes for both the Australian and US markets [Course Book §20.6]. The consequence that catches systematic traders is asymmetric in time. A position is entered when the trader chooses, which is when the line is liquid because that is why it was chosen. It is exited when the strategy says, which may be after spot has drifted and the once-at-the-money strike has become a wing. Liquidity risk in options is therefore a *path* risk: the exit's cost is a random variable correlated with the very move that triggered the exit.

This is also the honest frame for defined-risk structures. Retail accounts are steered towards verticals, iron condors and butterflies because those structures bound the loss, and bounding the loss is genuinely necessary — a naked short straddle has unbounded downside and belongs nowhere near an account sized by the Course Book's CVaR constraint [Course Book §14.5]. But the truncation is bought, and the price is paid in exactly the unit this chapter has been developing.

**Worked Example 11.4 — What a defined-risk structure costs, in volatility points**

A hypothetical index stands at 5,000.00 with a 30-day at-the-money implied volatility of 18.0. Quoted leg vegas, in index points per volatility point, are 5.72 at the money, 4.55 at the 25-delta strikes and 2.52 at the 10-delta strikes; quoted widths are 0.40 volatility points at the money, 0.60 at the 25-deltas and 1.20 at the 10-deltas. All figures are illustrative.

**Part A — the straddle.** Two legs, both sold, so the vegas add: package vega $2 \times 5.72 = 11.44$. Each crossing costs $(0.40/2) \times 5.72 = 1.1440$ index points; four crossings cost 4.5760 index points; and

$$c_{\text{rt}} = \frac{4.5760}{11.44} = \mathbf{0.4000} \text{ volatility points},$$

which is $w$, exactly as (11.2) requires. Priced at 18.0 implied, the straddle itself is worth 205.85 index points, so the round trip is **2.2230**% of the premium.

**Part B — the iron condor.** Now sell the 25-delta strangle (4,850 put, 5,175 call) and buy the 10-delta strangle (4,700 put, 5,325 call) against it, 150 index points of wing on each side. Four legs, and now the vegas *cancel*:

$$\text{net vega} \;=\; 2(2.52) - 2(4.55) \;=\; 5.04 - 9.10 \;=\; -4.06$$

index points per volatility point — a short vega position of 4.06, less than half the straddle's 11.44, obtained by paying for eight leg crossings instead of four:

$$\underbrace{4 \times \tfrac{0.60}{2}(4.55)}_{\text{25-delta legs} \,=\, 5.4600} + \underbrace{4 \times \tfrac{1.20}{2}(2.52)}_{\text{10-delta legs} \,=\, 6.0480} \;=\; \mathbf{11.5080} \text{ index points.}$$

Converting with the *net* vega, as Result 11.2's unit discipline requires,

$$c_{\text{rt}} = \frac{11.5080}{4.06} = \mathbf{2.8345} \text{ volatility points}, \qquad \frac{2.8345}{0.4000} = \mathbf{7.0862} \times \text{ the straddle's round trip.}$$

**Part C — against the money at stake.** Suppose the short strangle collects 58.40 index points and the long strangle costs 17.90, for a net credit of 40.50. With 150 points of wing the maximum loss is $150.00 - 40.50 = 109.50$ index points. Then

$$\frac{11.5080}{40.50} = \mathbf{28.41}\% \text{ of the credit}, \qquad \frac{11.5080}{109.50} = \mathbf{10.51}\% \text{ of the maximum loss.}$$

More than a quarter of everything the structure can possibly earn is consumed by the eight crossings, against 2.22% for the straddle — a ratio of 12.78 to one. Two conclusions, both structural. The wings are not free insurance: they cost roughly seven times as much per unit of retained volatility exposure, because the long legs cancel most of the vega while adding all of their own spread. And a defined-risk structure changes the shape of the tail without changing the expectancy in the trader's favour — it truncates a loss that was going to be estimated badly anyway, at a cost that must be counted before the strategy is evaluated, not after [Course Book §20.5].

### Expiry: exercise, assignment, and pin risk

The last day of a contract's life has mechanics of its own, and they are operational rather than mathematical.

**Exercise and assignment.** A long holder exercises; a short holder is *assigned*, usually by a random allocation run by the clearing house among the short positions at that strike. The short does not choose, is not consulted, and typically learns after the market has closed. In-the-money contracts are commonly exercised automatically under an exchange or broker threshold rule, which means a position the holder intended to abandon can be exercised for them, and a short they believed safe can be assigned on a contract that finished barely in the money.

**Pin risk.** The dangerous case is a physically settled contract that finishes almost exactly at the strike, because then the short does not know whether they will be assigned, and the two outcomes differ by a full share position. Concretely: short 20 contracts of a \$60.00 call on 100 shares each, and the shares settle at \$60.02 on the Friday. The residual value of the position is $0.02 \times 2{,}000 = \$40.00$. If assigned, the short is delivered into a short position of 2,000 shares over the weekend. Should the shares open 3% higher on Monday at \$61.8206, the loss on that unwanted share position is

$$2{,}000 \times (61.8206 - 60.02) \;=\; \$\mathbf{3{,}601.20},$$

which is **90.03** times the residual value of the option position that created it. Pin risk is not a pricing problem; it is the risk of acquiring, without consent and without warning, a directional position an order of magnitude larger than the one being closed. Cash-settled European index options have no pin risk at all — there is nothing to deliver — which is a concrete operational reason an index book is simpler to run than a single-name book, and one the payoff diagrams do not show.

**Settlement determination.** For cash-settled index options the payoff is computed against an official settlement value that is frequently *not* the last traded index level: it may be derived from opening prices of the constituents on the expiry morning, which can differ materially from any level the index printed. A position marked against the wrong reference for its last few hours is mismarked, and a backtest that settles index options at the closing index level is modelling a contract that does not exist. Read the settlement procedure from the contract specification.

## 11.6 The Volatility Trader's Workflow

A volatility trader's day has three phases, and they are the same three whether the book is a dealer's or a systematic account's: *screen*, *structure*, *manage*. Setting them out in order is what turns the preceding five sections from description into procedure.

### Screen: the surface against a forecast

The screen is a comparison, expiry by expiry, between what the market charges for variance and what a volatility model expects to be delivered. The market's side is the at-the-money implied volatility from the chain; the model's side is a realised-volatility forecast over the matching

horizon, which is precisely what the Course Book's Stage-1 engine produces [Course Book §12.6]. The edge is their difference in volatility points, $e = \sigma_I - \hat{\sigma}_r$, and it is tradeable only if it exceeds the round-trip cost $c_{\text{rt}} = w$ of (11.2).

| Expiry | Days | $\sigma_I$ | $\hat{\sigma}_r$ | $e$ | $w$ | $\lvert e\rvert - w$ | Verdict |
|---|---|---|---|---|---|---|---|
| Front weekly | 7 | 24.60 | 20.10 | +4.50 | 0.55 | +3.95 | sell volatility |
| Front month | 30 | 20.80 | 19.40 | +1.40 | 0.40 | +1.00 | sell volatility |
| Three month | 91 | 19.20 | 19.45 | −0.25 | 0.35 | −0.10 | no trade |

The screen above is a hypothetical snapshot, and its most important row is the one that looks best. A front-weekly edge of 4.50 volatility points against a 0.55-point market does not occur in an efficient front expiry for no reason, and the reason is almost always a scheduled event inside the window. Apply Result 11.4 with $T = 7/365$:

$$\sigma_f^2 T = 0.2460^2(0.0191781) = 0.001160581, \qquad \sigma_b^2 T = 0.2010^2(0.0191781) = 0.000774814,$$

$$V_e = 0.000385767, \qquad \sigma_e = \mathbf{1.9641}\%, \qquad \mathbb{E}|\text{move}| = 0.7979 \times 1.9641\% = \mathbf{1.5671}\%,$$

and the event accounts for **33.24**% of the expiry's total variance. Strip it out and the residual implied volatility is 20.10 — exactly the engine's forecast, and therefore exactly zero residual edge. That equality is by construction, and saying so is the point: the engine forecasts *diffusive* volatility, so the entire apparent edge is attributed to the event, which is the correct accounting and the reason the number cannot be traded as an inefficiency. What remains is a genuine but different question — is 1.5671% a fair price for this event? — answerable only against a distribution of the same event type on the same underlying, with the sample size stated. If the historical mean absolute move on this event has been 1.35%, the market is charging 0.2171 points of move too much, and *that* is the trade, with all the caveats a comparison of a market price to a small-sample historical mean deserves.

**Key idea — Screen in volatility, decide in net volatility, never in premium**

Three columns make the screen honest and no fewer will do. The market's implied volatility, the model's realised-volatility forecast for the *same* window, and the quoted width in volatility points at the moment of the comparison. The decision variable is $|e| - w$, not $e$, and never the premium. A screen that ranks candidates by premium ranks them by moneyness and tenor, which the trader already knows; a screen that ranks by $e$ alone ranks illiquid wings and event-laden front expiries to the top, which is worse than useless because both look like opportunities and neither is.

### Structure: choosing the instrument for the edge

The edge is a scalar; the instrument is not. Four considerations pick it.

*Purity.* The at-the-money straddle is the cleanest expression of a view on volatility alone: the legs' deltas nearly cancel at a strike near the forward while their gammas and vegas add, so the package begins direction-free and carries twice a single leg's convexity [Course Book §20.5]. If the view is about level, the straddle is the instrument.

*Shape of the view.* A strangle moves the strikes out of the money, which cheapens the position, flattens the gamma profile and — importantly — makes the trade partly a view on the *skew* rather than the level, because the wings' implied volatilities are what is being bought or sold. A calendar spread trades the *term structure*, buying one expiry against another. Matching the structure to the shape of the disagreement is what stops a level view from being contaminated by a skew position that was never intended.

*Cost.* Worked Example 11.4 priced this: adding wings to bound the loss cost roughly seven times as much per unit of retained vega. The correct reading is not that defined-risk structures are

wrong but that their cost must be paid out of the edge before the trade is judged, and an edge of one volatility point cannot fund a 2.83-point round trip whatever the payoff diagram looks like.

*Tail.* The Course Book's trap stands: short volatility is the high-hit-rate, negative-expectancy shape, and it must be sized under a fat-tailed conditional distribution by a CVaR constraint, never by volatility alone [Course Book §14.5]. That constraint, not the trader's preference, decides whether the naked or the defined-risk version is admissible.

### Manage: delta bands and the gamma/theta budget

Once on, a volatility position is managed on two dials.

*The delta band.* The position is delta-hedged, but not continuously — continuous hedging is infinitely expensive. The practical rule is a band: rebalance when the position's delta drifts beyond a stated fraction of its initial delta-neutral notional, plus a fixed cadence to catch slow drift. The band's width is a genuine trade-off with no closed-form answer at retail scale: narrow bands track the theoretical hedge and pay spread constantly; wide bands save spread and accumulate hedging error, which the dollar-gamma accounting shows is pure loss for a short-gamma book [Course Book §20.5]. Choose the band from measured hedging costs at your venue, log every rebalance, and treat the band as a pre-registered parameter rather than a discretionary one.

*The gamma/theta budget.* A long-volatility position earns from movement and pays for stillness; a short one is the reverse. Because these are two halves of one trade, the position should be monitored as a running comparison of realised variance delivered against implied variance paid, not as a profit-and-loss number, which mixes the volatility result with hedging error and residual delta. A book that is losing money while realised variance is running above implied has a hedging problem, not a forecasting problem, and the two require opposite responses.

**Your project — v2, as this workflow made systematic**

The Playbook's Chapter 10 sets out the v2 options overlay as a sequence of build steps; read against this section it is exactly the three phases above with artifacts and gates attached, and reading them side by side is the cleanest way to see what the platform is actually building.

- **Screen → capture and signal.** The chain snapshot table stores bid, ask, quoted implied volatility, spot at capture and *two* timestamps per line, because a settlement reference and a capture instant are different times; the volatility-signal table stores $\hat{\sigma}_{\mathrm{IV}}$, $\hat{\sigma}_r$ and the edge in volatility points. The event-strip computation above belongs here as a stored column, not a mental adjustment: any expiry containing a scheduled event carries its own $\sigma_e$, and the screen's ranking must be run on the residual edge, never the raw one.
- **Structure → the delta-hedged straddle protocol.** At-the-money straddle, sold when the edge is positive and bought when negative, with the defined-risk substitute available for accounts that cannot carry naked short gamma — priced, per Worked Example 11.4, at roughly seven times the round-trip cost per unit of vega, which is a number the gate must see rather than a footnote.
- **Manage → the hedge log.** One row per rebalance, with the delta before and after, so that the hedging error is measurable separately from the volatility result. Without this the two failure modes above cannot be distinguished.
- **The cost surface.** Built exactly as the FX one was: from recorded quotes at the times trades would have been placed, per strike and per tenor, never an average and never a mid. The cost-doubling control applies unchanged — if out-of-sample performance does not degrade when every cost term is doubled, the cost surface is under-measured rather than the edge being large.
- **The gate.** v2 is gated on v1's Stage-1 volatility engine having already passed, because the engine is the whole premise; on the engine beating $\sigma_I$-as-forecast on the Course

Book's own forecast-comparison test [Course Book §17.1]; on cost-net performance surviving cross-validation purged by *option life* rather than by calendar, since two contracts whose lives overlap share one realised path; and on the combined book's tail staying inside budget [Course Book §14.5]. Every strike rule and expiry rule tried counts in the trial ledger that prices the deflation correction [Course Book §15.4].

The single sentence that connects this chapter to the platform: the options market is the one venue where a good conditional-variance forecast is paid *directly*, with no directional view required — and it is also the venue where the cost of being wrong about the cost is largest.

## 11.7 Exercises

**Exercise 11.1 — Reading a chain**

A hypothetical index stands at 5,000.00 with 30 days to the expiry below. The at-the-money 5,000 line prints an implied volatility of 18.60, a vega of 5.70 index points per volatility point, and a quoted width of 2.28 index points. Six further lines are quoted; "width" is the quoted bid–ask width in *index points*, and vega is per volatility point.

| Strike | Type | Implied vol | Vega | Width (index pts) |
|---|---|---|---|---|
| 4,700 | put | 23.90 | 2.60 | 2.30 |
| 4,800 | put | 22.10 | 3.55 | 2.05 |
| 4,900 | put | 20.10 | 4.85 | 2.20 |
| 5,100 | call | 17.60 | 4.75 | 2.15 |
| 5,200 | call | 16.70 | 3.45 | 2.05 |
| 5,300 | call | 16.20 | 2.55 | 2.25 |

(a) Convert every quoted width into volatility points. Which line is tightest and which is widest in the traded unit, and what is the ratio? Compare that ratio with the ratio of the premium widths, and say what the comparison shows.

(b) Compute the risk-reversal and butterfly readings from the three equidistant strike pairs ($\pm 100$, $\pm 200$, $\pm 300$), using $\mathrm{RR} = \sigma_{\text{put}} - \sigma_{\text{call}}$ and $\mathrm{BF} = \frac{1}{2}(\sigma_{\text{put}} + \sigma_{\text{call}}) - \sigma_{\text{ATM}}$. State the direction of the skew and what the pattern of the butterflies says.

(c) Which line offers the cheapest volatility, which the richest, and what is the ratio?

(d) Your engine forecasts 18.40 realised volatility over the same 30 days. Buying pays the mid plus half the width in volatility terms and selling receives the mid minus half; using that rule, classify each of the seven lines as buy, sell or no trade.

(e) Look at the pattern of your answer to (d) and explain, in two or three sentences, why it is not a signal.

▷ working space

**Solution — Exercise 11.1**

**(a)** Divide each width by its vega:

| Strike | Width (index pts) | Vega | Width (vol pts) |
|---|---|---|---|
| 4,700 | 2.30 | 2.60 | 0.884615 |
| 4,800 | 2.05 | 3.55 | 0.577465 |
| 4,900 | 2.20 | 4.85 | 0.453608 |
| 5,000 | 2.28 | 5.70 | 0.400000 |
| 5,100 | 2.15 | 4.75 | 0.452632 |
| 5,200 | 2.05 | 3.45 | 0.594203 |
| 5,300 | 2.25 | 2.55 | 0.882353 |

Tightest is the 5,000 line at 0.400000; widest is the 4,700 put at 0.884615; the ratio is **2.2115**. Among the six quoted lines alone the range is 0.452632 to 0.884615, a ratio of 1.9544. The premium widths span only 2.05 to 2.30, a ratio of **1.1220**. **A chain that looks almost uniformly quoted in index points is not remotely uniform in the unit that matters**: the twelve-per-cent dispersion in premium terms is a 121-per-cent dispersion in volatility terms, entirely because vega falls away from the money. Any cost model built on premium widths will systematically under-price the wings.

**(b)**

$$\pm 100:\ \mathrm{RR} = 20.10 - 17.60 = +\mathbf{2.50}, \qquad \mathrm{BF} = \tfrac{1}{2}(20.10 + 17.60) - 18.60 = +\mathbf{0.25};$$

$$\pm 200:\ \mathrm{RR} = 22.10 - 16.70 = +\mathbf{5.40}, \qquad \mathrm{BF} = +\mathbf{0.80};$$

$$\pm 300:\ \mathrm{RR} = 23.90 - 16.20 = +\mathbf{7.70}, \qquad \mathrm{BF} = +\mathbf{1.45}.$$

Every risk reversal is positive, so **puts are bid over calls at every distance**: the classic equity-index downside skew, whose economic content is persistent one-way demand for crash protection and matching supply of upside (Section 11.4). Across the whole quoted range the slope is $(23.90 - 16.20)/600 = 1.28333$ volatility points per 100 index points. The butterflies increase monotonically with distance from the money, $0.25 \to 0.80 \to 1.45$, which is the market pricing progressively fatter tails than lognormality allows — the curvature is a tail-risk gauge, distinct from the slope.

**(c)** Cheapest volatility: the 5,300 call at **16.20**. Richest: the 4,700 put at **23.90**. Ratio $23.90/16.20 = \mathbf{1.4753}$. The same underlying, the same expiry, and a fifty-per-cent difference in the price of variance — which is the whole content of the statement that the surface, not the level, is the market's object.

**(d)** Half-widths in volatility points are half the (a) column. Buy if $\sigma_I + w/2 < 18.40$; sell if $\sigma_I - w/2 > 18.40$.

| Strike | $\sigma_I$ | Buy at | Sell at | Verdict |
|---|---|---|---|---|
| 4,700 | 23.90 | 24.3423 | 23.4577 | sell |
| 4,800 | 22.10 | 22.3887 | 21.8113 | sell |
| 4,900 | 20.10 | 20.3268 | 19.8732 | sell |
| 5,000 | 18.60 | 18.8000 | 18.4000 | **no trade** (sell price equals the forecast exactly) |
| 5,100 | 17.60 | 17.8263 | 17.3737 | buy |
| 5,200 | 16.70 | 16.9971 | 16.4029 | buy |
| 5,300 | 16.20 | 16.6412 | 15.7588 | buy |

The 5,000 line is the instructive edge case: selling receives exactly 18.40, equal to the forecast, so the expected edge net of cost is exactly zero and the rule declines it.

**(e)** The pattern is: sell every put, buy every call — a risk reversal, produced without any view on the skew ever being formed. It is an **artefact of comparing a scalar to a surface**. The engine emits one number, a forecast of realised volatility, and the chain quotes a whole curve across strikes; mechanically, any horizontal line drawn through a monotonically falling curve is above it on one side and below it on the other, so the rule will always prescribe selling the high-volatility wing and buying the low-volatility wing whatever the forecast is. The skew it is trading against is a risk premium — payment for crash insurance that has repeatedly had to be paid out — not a mispricing, and a strategy that harvests it is a short-crash-insurance strategy in disguise, with exactly the tail that implies. The honest fix is to compare like with like: forecast the volatility *at the moneyness being traded*, or trade the level only (at the money) and treat any wing position as a separate, separately justified view on skew.

**Exercise 11.2 — Conversion drills**

Two quotes, one on a hypothetical index and one on a hypothetical single name. All figures illustrative.

- **Index leg.** Quoted 41.20 bid, 43.60 offered, in index points; vega 4.85 index points per volatility point; mid implied volatility 20.10.
- **Single-name leg.** Quoted $1.28 bid, $1.44 offered, per share; vega $0.0615 per share per volatility point; contracts on 100 shares.

(a) For the index leg: the width in index points and in volatility points, and the market restated as a two-sided volatility quote.

(b) For the single-name leg: the width per share, per contract, and in volatility points, and the vega per contract.

(c) Which is the more expensive market to cross, by what factor?

(d) A strategy needs 1.50 volatility points of edge to be worth doing. Express that requirement in premium for each leg.

(e) Build an at-the-money straddle from two legs like each of the above, and compute the round-trip cost in volatility points of the package both by the premium route and by (11.2). State the break-even forecast accuracy each market demands.

▷ working space

**Solution — Exercise 11.2**

**(a)** $W = 43.60 - 41.20 = \mathbf{2.40}$ index points; $w = 2.40/4.85 = \mathbf{0.494845}$ volatility points; half-width 0.247423. The mid premium is 42.40, so in the traded unit the market is

$$\mathbf{19.852577} \text{ bid at } \mathbf{20.347423} \text{ offered.}$$

**(b)** $W = 1.44 - 1.28 = \$\mathbf{0.16}$ per share $= \$\mathbf{16.00}$ per contract; vega per contract $= \$0.0615 \times 100 = \$\mathbf{6.15}$ per volatility point;

$$w = \frac{0.16}{0.0615} = \mathbf{2.601626} \text{ volatility points} \qquad \left(\text{identically } \frac{16.00}{6.15}\right),$$

which is the check that the conversion is scale-free: per share or per contract, the volatility width is the same number, because the multiplier cancels.

**(c)** $2.601626/0.494845 = \mathbf{5.2575}\times$. The single-name market costs more than five times as much to cross, a fact wholly invisible in the premium quotes (2.40 index points against sixteen cents).

**(d)** Index leg: $1.50 \times 4.85 = \mathbf{7.2750}$ index points. Single-name leg: $1.50 \times \$0.0615 = \$0.092250$ per share $= \$\mathbf{9.2250}$ per contract. Note how differently the same requirement reads in the two premium currencies, and how identically it reads in volatility points — which is the argument for the unit.

**(e)** *Index straddle.* Package vega $2 \times 4.85 = 9.70$. Each crossing costs $2.40/2 = 1.20$ index points; four crossings $= 4.80$ index points; $4.80/9.70 = \mathbf{0.494845}$ volatility points, equal to $w$ as (11.2) requires. *Single-name straddle.* Package vega $2 \times \$0.0615 = \$0.1230$ per share; four crossings of \$0.08 each $= \$0.32$; $0.32/0.1230 = \mathbf{2.601626}$ volatility points, again equal to $w$. Break-even forecast accuracy: the index trade must beat the market's own volatility estimate by **0.4948** volatility points, the single-name trade by **2.6016**, before hedging costs, slippage or any forecast error. The second number is larger than the typical dispersion between two competent volatility forecasts, which is the quantitative reason a retail volatility programme belongs in index options rather than single names.

**Exercise 11.3 — Dealer hedging flow from an aggregated chain**

A hypothetical index stands at 4,200.00; its options carry a hypothetical multiplier of \$30 per index point. Four lines carry the chain's open interest. $\Gamma$ is in contracts of delta per index point; the dealer-side column is the conventional assumption that customers buy index puts and sell index calls.

| Strike | Type | Open interest | $\Gamma$ per contract | Dealer side |
|---|---|---|---|---|
| 4,000 | put | 128,000 | 0.00142 | short |
| 4,100 | put | 96,000 | 0.00168 | short |
| 4,200 | call | 84,000 | 0.00185 | long |
| 4,300 | call | 63,000 | 0.00171 | long |

(a) Aggregate the street's net gamma in contracts of delta per index point, and state whether it is long or short.

(b) Compute the hedging flow required by a $+1\%$ move: direction, contracts, and notional value at the post-move index level.

(c) Do the same for a $-1\%$ move, and describe the effect on the market in one sentence.

(d) Compute the "dollar gamma per one per cent" figure at 4,200 and reconcile it with your answers to (b) and (c).

(e) State precisely which input to this calculation is not observable, and what happens to every answer above if it is wrong.

▷ working space

**Solution — Exercise 11.3**

**(a)** Sign each line by the dealer's side and multiply by open interest:

$$-128{,}000(0.00142) = -181.76, \quad -96{,}000(0.00168) = -161.28,$$

$$+84{,}000(0.00185) = +155.40, \quad +63{,}000(0.00171) = +107.73,$$

$$\text{net } \Gamma_{\text{pos}} = -181.76 - 161.28 + 155.40 + 107.73 = \mathbf{-79.91}$$

contracts of delta per index point. The street is **short gamma**.

**(b)** $+1\%$ is $\delta S = +42.00$ points, so the index moves to 4,242.00 and by (11.3) the position's delta changes by $-79.91 \times 42.00 = -3{,}356.22$ contracts. By (11.4) dealers must **buy** 3,356.22 **contracts**, worth

$$3{,}356.22 \times 4{,}242.00 \times 30 = \$\mathbf{427.11} \text{ million.}$$

**(c)** $\delta S = -42.00$ takes the index to 4,158.00 and changes the delta by $+3{,}356.22$ contracts, so dealers must **sell** 3,356.22 **contracts**, worth $3{,}356.22 \times 4{,}158.00 \times 30 = \$\mathbf{418.65}$ million. In one sentence: a short-gamma street buys strength and sells weakness, so its hedging is same-signed with price and therefore amplifies whatever move is under way, most strongly where gamma is largest.

**(d)** Dollar gamma per one per cent is $|\Gamma_{\text{pos}}| \times 0.01S \times S \times M$:

$$79.91 \times 42.00 \times 4{,}200.00 \times 30 = \$\mathbf{422.88} \text{ million.}$$

It sits between the two figures in (b) and (c) because it values the same 3,356.22 contracts at 4,200 rather than at the post-move level; the \$4.23 million above it and \$4.23 million below it are the second-order effect of the index level itself moving. The per-one-per-cent number is a local slope, and quoting it as though it were a constant overstates the down-move flow and understates the up-move flow.

**(e)** The **dealer-side column**. Exchanges publish open interest — how many contracts exist — not who holds which side. If the assumption is reversed (dealers long the puts, short the calls), the net gamma becomes $+\mathbf{79.91}$ and every flow above flips direction: sell 3,356.22 contracts into the rally, buy 3,356.22 into the decline, and the street dampens rather than amplifies. Same open interest, same gammas, opposite market impact. The assumption is defensible as a central case for broad index options because protective-put demand is the same demand that produces the skew of Exercise 11.1; it is much weaker for single names and for short-dated contracts where speculative two-way flow dominates. Any published gamma-exposure figure is therefore a model output, not data, and should be stored with its assumption attached.

**Exercise 11.4 — An event crush with different numbers**

A hypothetical single name trades at \$52.00 and reports after today's close. The front expiry has 11 calendar days to run at an implied volatility of 58.0; the diffusive baseline read from the neighbouring expiries is 34.0. Contracts are on 100 shares; financing is zero. The at-the-money straddle is worth \$4.175783 per share, with straddle vega \$0.0719354 per share per volatility point.

(a) Decompose the expiry's variance: the event's one-off standard deviation, the diffusive standard deviation over the life, the total, the event's share of total variance, and the expected absolute event move.

(b) Check the straddle price against the decomposition by computing the "implied move" $V_0 / S_0$ and comparing.

(c) The shares fall 6.00% overnight to \$48.88 and the implied volatility falls to 34.0, with 10 days left; the straddle is then worth \$3.509686 per share. Compute the P&L per contract and as a percentage.

(d) Attribute the result by sequential revaluation, given the intermediate values \$3.981605 (time only) and \$2.334633 (time then volatility).

(e) The break-even spot levels under the post-event surface are \$55.8885 and \$48.0245. State what the buyer needed, compare it with what happened, and explain in one sentence why a 6% move was insufficient.

(f) Had implied volatility not moved, the straddle would have been worth \$4.636779. What does that tell you about what the trade actually was?

▷ working space

**Solution — Exercise 11.4**

**(a)** $T = 11/365 = 0.0301370$. By (11.5),

$$\sigma_f^2 T = 0.58^2(0.0301370) = 0.010138082, \qquad \sigma_b^2 T = 0.34^2(0.0301370) = 0.003483836,$$

$$V_e = 0.006654247, \qquad \sigma_e = \sqrt{0.006654247} = \mathbf{8.1574}\%, \qquad \sigma_{\text{diffusive}} = \mathbf{5.9024}\%,$$

$$\sigma_{\text{total}} = \sqrt{8.1574^2 + 5.9024^2} = \mathbf{10.0688}\%,$$

the event accounting for **65.64**% of the expiry's total variance. The expected absolute event move is $0.7979 \times 8.1574\% = \mathbf{6.5086}\%$.

**(b)** $V_0/S_0 = 4.175783/52.00 = \mathbf{8.0304}\%$, and $0.7979 \times 10.0688\% = 8.0337\%$. The straddle price and the term-structure decomposition are two readings of one number, and they agree to rounding.

**(c)**

$$\text{P\&L} = (3.509686 - 4.175783) \times 100 = -\$\mathbf{66.61} \text{ per contract } = -\mathbf{15.95}\%.$$

**(d)** Sequential revaluation, one input at a time:

$$\text{time } (11 \to 10 \text{ days}): \quad 3.981605 - 4.175783 = -0.194178 \text{ /share} = -\$\mathbf{19.42} \text{ /contract},$$

$$\text{volatility } (58.0 \to 34.0): \quad 2.334633 - 3.981605 = -1.646972 = -\$\mathbf{164.70},$$

$$\text{spot } (52.00 \to 48.88): \quad 3.509686 - 2.334633 = +1.175053 = +\$\mathbf{117.51},$$

summing to −$66.61 with no residual. The crush cost $164.70; the six-per-cent gap earned back $117.51; a day of decay cost $19.42.

**(e)** The buyer needed $48.0245 on the downside, a move of $-\mathbf{7.6453}\%$ (or $+7.4779\%$ on the upside). The delivered $-6.00\%$ was well short. In one sentence: the market had already priced an expected absolute event move of 6.51% and charged for it in the 58.0 implied volatility, so a 6% move was slightly *smaller* than what the buyer had paid for — the position needed the event to be bigger than the price of the event, not merely to happen.

**(f)** Holding volatility at 58.0, the same gap would have produced $4.636779 - 4.175783 = +\$0.460996$ per share, or $+\$\mathbf{46.10}$ per contract, $+11.04\%$. The entire difference between an eleven-per-cent gain and a sixteen-per-cent loss is the volatility term. The trade was therefore never a bet on the earnings move at all; it was a bet that the market's price for the earnings move was too low, and that bet lost even though the directional call was, in magnitude terms, not far off.

**Exercise 11.5 — Why "buy cheap options before events" fails structurally**

Write a structured argument, of roughly a page, explaining why buying options into a scheduled event is not a repeatable edge — and be precise about what would have to be true for it to become one. Your answer should (i) state the mechanism using the variance decomposition of Result 11.4; (ii) explain what "cheap" can and cannot mean on an option whose expiry spans a known event; (iii) connect the argument to the general no-free-lunch reasoning of Chapter 1 of this book — who is on the other side, and why would they trade with you; and (iv) state the evidential standard any claimed edge would have to clear.

▷ working space

**Solution — Exercise 11.5**

**(i) The mechanism.** An expiry spanning a scheduled event carries total variance $\sigma_f^2 T = \sigma_b^2 T + V_e$. The event's contribution $V_e$ is added to the price *because the date is public*. The implied volatility of that expiry is therefore not a forecast of ordinary volatility at all; it is a blend of a diffusive rate and a one-off jump size, and it must fall when the jump resolves, because $V_e$ leaves the remaining life whatever the outcome. The "crush" is not a market failure, a liquidity event or a manipulation; it is the arithmetic of (11.5) running forwards in time. A long option position held through the event is short that arithmetic with certainty and long the realised jump with uncertainty. Worked Example 11.3 shows the resulting asymmetry concretely: a three-per-cent overnight move — large by any ordinary standard — returned a 24.5% loss, because the priced expected absolute move was 4.16% and the break-even was 4.90%.

**(ii) What "cheap" can mean.** A pre-event option is never cheap in the sense the phrase usually intends, namely low premium: its premium is high, which is exactly the complaint. Nor is it cheap in implied volatility: that number is elevated precisely because of the event. There is exactly one coherent sense in which such an option can be cheap, and it is narrow: the event's priced move $\sigma_e$, extracted by (11.5), may be small relative to what this event, on this underlying, has historically delivered. That is a comparison between a market price and a conditional distribution, and it is a statement about *one number*, not about the option. Any argument for buying that does not produce $\sigma_e$ and compare it to something is not an argument about cheapness; it is a restatement of the belief that the stock will move, which the market already holds and has already charged for.

**(iii) Who is on the other side.** Chapter 1's discipline applies without modification: a trade exists only if someone will take the other side, and their reason is the trade's real risk. The seller of a pre-event straddle is a dealer or a systematic volatility seller who has priced the event from a large historical sample of comparable events, who will delta-hedge the position through the release, who has a far better estimate of $\sigma_e$ than a retail account does, and who is being paid the variance risk premium on top — the compensation that exists because option sellers lose money precisely when everything else is losing money [Course Book §20.5]. The buyer's counterparty is therefore better informed about the specific quantity being traded, better hedged, and paid a premium for the residual risk. That is the adverse-selection structure of Chapter 2 in its purest form, and the null hypothesis must be that the buyer's expectancy is negative by the amount of the premium plus the round trip. Costs make it worse: by (11.2) the round trip costs one full quoted width, and single-name pre-event markets are among the widest listed markets there are — Exercise 11.2 put a single-name straddle's round trip at 2.60 volatility points, which is a large fraction of any plausible edge in $\sigma_e$.

**(iv) The evidential standard.** "Buy straddles before earnings" is among the most heavily tested propositions in retail finance, which is precisely what makes an encouraging backtest of it worth so little. The strategy space is large — which underlyings, which expiry, how many days before, how many days after, at the money or which delta, hold through or exit on the print, filter by implied-versus-historical rank, filter by sector — so the best of a few hundred variants is guaranteed to look good in-sample. Any claim therefore needs: the event-move distribution estimated point-in-time, with no use of data published after the trade date [Course Book §18.2]; cost-net evaluation on a measured width surface at pre-event widths, not average widths; cross-validation purged by option life, since overlapping contracts share one realised path; a deflation correction that counts *every* variant tried, not the one reported [Course Book §15.4]; and a sample containing enough events per underlying for the tail to be represented at all. The honest prior, before any of that, is the one this book has repeated in every market it has examined: the structural features of a market are priced, and a mechanism visible to

everyone with a screen is not an edge. The edge, if there is one, lives in a better estimate of $\sigma_e$ than the market's — which is a forecasting claim, testable by the same machinery as every other forecasting claim in the programme, and subject to the same gate.

**Key idea — What to remember**

The listed option market is a thousand-line grid quoted in premium and traded in volatility, and the conversion between the two units — divide by vega, and make certain the vega and the premium refer to the same object — is the single habit that makes everything else legible: it turns incomparable dollar spreads into comparable widths, it shows that a round trip costs one full quoted width whenever a structure's legs add rather than cancel, and it prices the wings of a defined-risk structure at seven times the straddle's cost per unit of retained exposure. Every option trade generates a delta hedge in the underlying within seconds, so the street's aggregate gamma is a mechanical flow generator whose sign determines whether hedging dampens moves or amplifies them — and that sign rests on an assumption about who holds which side of the open interest, which is not published and must never be mistaken for data. The surface itself is a market object rather than a fitted function: its skew is the standing inventory imbalance of a world that is structurally long equities and wants protection, and its term-structure bumps are scheduled events whose variance can be extracted arithmetically and whose resolution — the crush — is the most reliably profitable trade in retail finance for the person on the other side. What the platform's volatility engine buys here is not a directional view but a direct market for a conditional-variance forecast, gated on the engine first beating the market's own implied volatility as a forecast, and costed in the same unit the market quotes. The next chapter leaves the derivatives complex for commodities, where the curve rather than the surface is the market's central object, and where storage and convenience yield replace vega as the thing that has to be priced.

# 12

# Commodities

*Chapter 11 closed this book's excursion through mechanism-agnostic market structure — a listed contract, a clearing house, a margin ledger, a surface quoted in one unit and traded in another, all built by Chapter 10 and Chapter 11 to work identically whichever underlying sits beneath the ticker. Commodities are where that agnosticism stops paying its way. A stock index has no harvest and no storage tank; a currency pair has no delivery point and no convenience yield; a barrel of crude, a bushel of wheat and an ounce of gold all do, and the physical reality underneath the price — how much of the thing exists, where it sits, what it costs to hold, what it is worth to have on hand rather than on paper — finally does most of the analytical work. Chapter 10 flagged this in passing, generalising its cost-of-carry relationship to a commodity curve and noting that commodities, two chapters further on, are where that curve's storage-and-convenience-yield term finally earns its keep; this is that chapter. Four questions organise it. What physically determines a commodity's price, and why is almost none of the "spot" trading this book's reader will ever see actually a spot transaction (§12.1)? What does a futures curve's shape say about that physical reality, and how far can the answer be trusted before it slides from description into forecast (§12.2–§12.3)? What, specifically, is gold — priced like a currency, behaving like neither a currency nor an ordinary commodity — and what does the platform's own WTI contract inherit from crude's peculiar delivery geography (§12.4–§12.5)? And once roll drag, financing and the academic risk-premium debate are all laid on the table together, what does a leveraged CFD account actually keep (§12.6)? None of the arithmetic here is new in kind — it is Chapter 10's curve, Chapter 4's covered-interest logic, and Chapter 7's positioning categories, applied to instruments whose underlying can be stored, spoiled, mined, or, for one week in April 2020, refused at any price.*

## 12.1 The Physical Layer

Every commodity price is, underneath whatever curve or CFD quote represents it on a screen, the resolution of three physical flows: how much of the thing is being produced, how much is being consumed, and how much sits in storage bridging the gap between the two.

**Definition — The price-formation triangle**

A commodity's price is set, at any horizon longer than a single session's order-flow noise, by the resolution of three physical quantities.

**Production.** The flow rate at which the physical good is extracted, grown or refined — mine output, well flow, harvest volume, refinery throughput — set by decisions (drilling, planting, capacity utilisation) taken well before the commodity reaches a market at all, and typically slow to adjust: a mine or a well can rarely double output on a quarter's notice, and a crop planted in spring cannot be re-planted once a summer drought has already cut its yield.

**Consumption.** The flow rate at which the good is used up — burned, eaten, manufactured into something else — set by an entirely separate set of decisions (industrial activity, weather-driven heating and cooling demand, dietary habits) that shares no calendar with production at all.

**Storage.** The stock that absorbs the difference between the two flows at every point in time. When production exceeds consumption, the surplus is added to storage (a *build*); when consumption exceeds production, the shortfall is drawn from storage (a *draw*). Storage is what makes it possible for production and consumption to run on entirely different calendars — a harvest concentrated in one month, consumption spread across the year — without price gyrating without bound between the two.
The single state variable tying all three together, and that the rest of this chapter returns to repeatedly, is the *level of storage relative to what the market considers normal for the season*: ample storage means the marginal unit of the commodity is cheap to hold and easy to obtain on short notice, tight storage means the opposite, and §12.2 shows exactly how that single fact is what a futures curve's shape is actually pricing.

Three broad groups populate this triangle, and the group a commodity belongs to sets, to first order, how strong a candidate it is for the seasonality §12.3 treats carefully.

**Energy.** Crude oil, refined products (petrol, diesel, jet fuel, heating oil — refined from crude at a margin the industry calls the *crack spread*, itself a tradeable spread in its own right, priced off the same production/consumption/storage logic one refining step downstream of crude) and natural gas together form the group with the largest, most immediate consumption response to weather: heating and cooling demand move on a calendar the physical layer cannot ignore, and §12.3 treats natural gas and heating oil as the cleanest seasonality case this chapter has.

**Metals.** Precious metals — gold, silver, and the platinum-group metals — sit closer to a monetary asset than to an industrial input, discussed on their own terms in §12.4; industrial metals — copper prominent among them, used throughout this chapter's exercises as an illustrative case even though it is not one of the platform's own instruments — track the global industrial-production and construction cycle far more closely than any weather calendar, and mined output, unlike a harvest, runs close to continuously through the year, exactly why §12.3 finds almost nothing seasonal in this group.

**Agriculture.** Grains, softs and livestock are where the harvest calendar is not a metaphor: a single annual (or, for some crops, twice-yearly) harvest concentrates a year's production into a period of weeks, storage must bridge the entire gap to the next one, and the resulting scarcity-into-abundance cycle is this chapter's other clean seasonality case.

### Spot is a convenient fiction for most of this list

Chapter 10 defined the **basis** $b_t = F_t - S_t$ and showed it converges to zero at expiry, by arbitrage for a physically deliverable contract or by construction for a cash-settled one — machinery that presumes a genuine, independently observable spot price $S_t$ exists to converge toward. For crude oil and natural gas specifically, that presumption is close to false for a retail participant. There is no centralised cash market in either: the actual "spot" transaction is a bespoke, bilaterally negotiated physical delivery between a producer, a refiner, a pipeline operator or a utility, struck at a specific delivery point — Cushing, Oklahoma for WTI; a North Sea loading terminal for Brent; a specific pipeline hub for natural gas, Henry Hub in Louisiana the most widely referenced US benchmark point — on terms no public screen ever shows in full. What a retail platform quotes as "spot WTI" or "spot natural gas" is, functionally, the nearest listed futures contract's price, or a short blended window across a roll (Chapter 10; [Course Book §22.1]) — not a transaction any retail account could actually enter directly. Gold and silver sit closer to the textbook case: a deep, continuous OTC bullion market really does exist and really does trade physical metal — but even there, arbitrage ties that OTC price so tightly to the futures curve (§12.2) that the distinction rarely matters in practice, and the platform's own XAU/USD and XAG/USD CFDs are, structurally, no different from its WTI CFD in this respect (§12.4). The practical consequence for everything that follows: whenever this chapter, or the reader's own platform statements, say "spot," read it as "the nearest tradeable proxy for spot," not as evidence a genuine cash market was actually observed.

## 12.2 Storage and the Curve

Chapter 10 derived, from the same no-arbitrage logic used to price a financial future, the fair-value relationship a commodity's futures curve must satisfy once physical storage is admitted as a cost and physical possession as a benefit. Restated here without rederivation:

$$F = S\,e^{(r+u-y)T}, \tag{12.1}$$

where $u$ is the proportional annual cost of storing and insuring the physical commodity and $y$ is Working's (1949) **convenience yield**: the implicit, non-cash benefit a holder of the physical good captures — avoiding a stock-out, meeting a sudden order, keeping a production line running — that a holder of a paper future does not. Chapter 10 also flagged (12.1)'s central practical limitation: with only $F$, $S$ and $r$ routinely observable, $u$ and $y$ are never separately identified from the curve alone, only their difference $u - y$, backed out as whatever residual reconciles the equation.

This chapter needs the two terms separated, not merely their difference, because $u$ and $y$ answer different questions: $u$ is close to an engineering fact (a warehousing quote, an insurance premium — in principle obtainable independently of the futures market), while $y$ is a statement about how tight the physical market is judged to be right now, exactly the price-formation triangle's storage-relative-to-normal state variable made numeric. The identification problem Chapter 10 flagged dissolves the moment $u$ is supplied from outside the curve itself — from a warehousing and insurance quote for the specific commodity and location in question — because (12.1) then has only one unknown left.

**Result — Backing out the convenience yield**

Given $S$, $F$, $r$, $u$ and $T$, rearranging (12.1) for $y$ is one line of algebra:

$$\ln F = \ln S + (r+u-y)T \quad\Longrightarrow\quad y \;=\; r + u \;-\; \frac{1}{T}\ln\!\left(\frac{F}{S}\right).$$

The result is a genuine convenience yield — a rate directly comparable across commodities and dates — only if $u$ is itself a trustworthy, independently sourced storage-cost estimate; feed the formula a wrong $u$ and it returns a wrong $y$ with exactly the same confident precision, since the arithmetic cannot tell a mis-specified input from a correct one. The two classification thresholds Chapter 10 established carry over unchanged: $y > r + u$ produces **backwardation**, potentially with $F < S$ outright when $y$ is large enough; $y < r + u$ produces **contango**.

**Worked Example 12.1 — Convenience yield, backed out from two wheat curves**

Two hypothetical wheat curves, six months apart in the crop calendar — illustrative figures only, not a quote from any real wheat market — with $u$ supplied as an independent storage-and-insurance estimate in each case.

**Part A — pre-harvest, tight stocks.** $S = 612.00$, $F = 601.75$ (note $F < S$ already: outright backwardation before any decomposition), $r = 5.25\%$, $u = 3.20\%$, $T = 60/365 = 0.1643836$ years.

$$\ln(F/S) = \ln(601.75/612.00) = -0.01689021, \qquad r + u = 8.4500\%,$$

$$y = 8.4500\% - \frac{-0.01689021}{0.1643836} = 8.4500\% + 10.2749\% = \mathbf{18.7249}\%.$$

Check: $S\,e^{(r+u-y)T} = 612.00 \times e^{(0.0845-0.187249)\times 0.1643836} = \mathbf{601.75}$, exactly $F$. With $y = 18.7249\% \gg r + u = 8.4500\%$, the classification is unambiguous: **backwardation**, and a large one — pre-harvest stocks tight enough that the market is pricing a substantial scarcity premium for physical wheat on hand right now, over and above whatever it would cost to simply finance and store it.

**Part B — post-harvest, ample stocks.** $S = 480.00$, $F = 495.60$ (now $F > S$), $r = 4.60\%$, $u = 2.50\%$, $T = 150/365 = 0.4109589$ years.

$$\ln(F/S) = \ln(495.60/480.00) = 0.03198305, \qquad r + u = 7.1000\%,$$

$$y = 7.1000\% - \frac{0.03198305}{0.4109589} = 7.1000\% - 7.7825\% = -\mathbf{0.6825}\%.$$

Check: $S\,e^{(r+u-y)T} = 480.00 \times e^{(0.046+0.025-(-0.006825))\times 0.4109589} = \mathbf{495.60}$, exactly $F$. Here $y < r + u$: **contango**, as ample post-harvest stocks predict — but notice the sign of $y$ itself: a *negative* implied convenience yield, not merely a small positive one. This is not nonsensical — $y$ is whatever residual the equation returns given the assumed $u$ — but it is worth reading honestly: it says the curve is pricing *more* carry cost than the assumed 2.50% storage rate alone supplies, so either the true marginal cost of storing wheat this deep into a glut genuinely exceeds the average rate used for $u$ (a full bin costs more at the margin than an empty one), or $u$ was simply underestimated for this specific case. A negative $y$ is a flag to check the input, not a licence to report a below-zero convenience yield as a stable fact about wheat.

Keynes (1930) read a downward-sloping curve differently: not as a statement about $F$ versus $S$ today at all, but about where $F$ sits relative to the market's own expectation of the future spot price, $\mathbb{E}[S_T]$. His theory of **normal backwardation** holds that commercial hedgers in a given commodity are, on net, more often short than long — producers hedging output outnumber consumers hedging input — so $F$ must sit below $\mathbb{E}[S_T]$ to compensate the speculators willing to take the other, net-long side of that hedging pressure: a risk premium for absorbing price risk the hedging side wants to shed, not a statement about storage economics at all. The two readings do not contradict each other — (12.1)'s decomposition describes the curve's shape *today*, built from quantities that are at least in principle observable; Keynes's describes an expected-return relationship that today's curve alone cannot verify, because $\mathbb{E}[S_T]$ is never directly observed — and §12.6 below is precisely where this book takes up whether Keynes's compensation is actually there, once realistic costs are subtracted from it, rather than assumed.

**Trap — A backwardated curve is a fact about today, not a forecast**

Worked Example 12.1's Part A is a striking number — an 18.7% backed-out convenience yield is a genuinely tight physical market — and it is tempting to read tight-and-backwardated as "buy the physical, or buy the future, because scarcity like this must resolve upward." Nothing derived above licenses that step. (12.1) is a statement about the relationship between $F$ and $S$ *today*, given $r$, $u$ and $y$ as they stand today; it says nothing about where $S$ itself is headed, in exactly the same way Chapter 10 showed contango and backwardation say nothing about a financial future's underlying direction. Keynes's normal-backwardation reading *does* make a directional claim — that $F$ sits below $\mathbb{E}[S_T]$ — but that claim is a hypothesis about a risk premium, not a fact the curve's shape hands over for free, and §12.6 treats it exactly as carefully as Chapter 7 treated an extreme COT $z$-score: a candidate reading, not a conclusion, until it clears an honest test.

## 12.3 Seasonality, Where It Is Real

The price-formation triangle (§12.1) makes a sharp, falsifiable prediction about which commodities should show genuine calendar-driven return patterns: wherever production or consumption is itself tied to a fixed point in the calendar, storage should show a predictable seasonal build-and-draw cycle, and the curve — pricing exactly that storage state through $y$ (§12.2) — should show a matching seasonal pattern in its shape. Wherever neither flow has a calendar, no such prediction

exists, and any seasonal pattern found is, absent a specific mechanism, more likely to be Chapter 1's phantom FX Septembers wearing a commodity ticker.

**Where it is real.** Natural gas and heating oil face a demand flow — heating, and, for gas, a smaller cooling-driven summer bump — pinned to the northern-hemisphere calendar with genuine physical force behind it: storage is drawn down through the winter withdrawal season and rebuilt through the spring-to-autumn injection season, a cycle with nothing arbitrary about its timing. Agricultural futures answer to an equally real calendar on the production side: a harvest concentrates a year's supply into weeks, and the storage drawn down to bridge to the next one predicts, mechanically, a scarcity build into the pre-harvest period and a release immediately after — the same shape Worked Example 12.1's two curves were constructed to illustrate.

**Where it is weak to absent.** Mined metals — both industrial and precious — have no harvest and, for the most part, no heating-driven demand cycle either: extraction runs close to continuously through the year, and whatever demand seasonality exists (cultural gold-buying calendars concentrated in parts of the year are the standard candidate raised for precious metals specifically) is a much weaker, less mechanically forced effect than a genuine storage-drawdown cycle. §12.4 returns to gold's own case without pretending the testing standard below is any softer for having a plausible-sounding mechanism attached.

[Course Book §22.2] sets out this book's full seasonality-testing discipline in detail — the joint test across all twelve calendar months rather than a scan for the single best-looking one, the Bonferroni (or better, deflated and correlation-aware) correction once the search moves from "is there an effect" to "which month," and the insistence on an economic mechanism proposed *before* the data is examined, not fitted to whichever cell happened to survive — and none of it is re-derived here. The one point worth restating for this chapter specifically: a plausible mechanism, of exactly the kind natural gas and agriculture genuinely have, makes a false-positive finding *more* persuasive to a reader, not less, which is precisely why the discipline has to be applied with no less rigour to a commodity than to an FX pair with no mechanism at all.

**Worked Example 12.2 — A seasonal natural-gas curve, read two ways**

A hypothetical natural-gas futures curve, four expiries quoted at one instant — illustrative units only, not a live market's current level: January (deep winter) 4.85, April (early injection season) 3.55, July (mid-injection, summer) 3.10, October (end of injection season, just before the next withdrawal begins) 3.95.

**The winter premium.** January trades $4.85 - 3.10 = \mathbf{1.75}$ above July, or 56.4516% of July's level (1.75/3.10) and 36.0825% of January's own level (1.75/4.85) — a large, one-sided premium for gas delivered at the point of maximum heating demand and minimum storage, exactly the shape §12.2's framework predicts when $y$ (here, the convenience of having gas in hand as the coldest weeks arrive) is running far above whatever $r + u$ alone would charge for financing and storing it that long.

**The restocking premium.** October, priced at the very end of the injection season, trades $3.95 - 3.10 = \mathbf{0.85}$ above July: storage is being rebuilt ahead of the next winter, and the market is already charging a premium for gas available at exactly the point the withdrawal season is about to reopen, months before January's own premium is at its widest.

**How fast the winter premium decays.** January trades $4.85 - 3.55 = \mathbf{1.30}$ above April: over three months, roughly three-quarters of the full January-to-July gap (1.30 of 1.75) has already closed, consistent with a genuine physical demand cycle that peaks sharply in the coldest weeks and fades quickly once they pass, rather than a level shift that would hold roughly constant across the shoulder months.

None of these three numbers is a forecast — Trapbox 12.2 applies here exactly as it did to wheat — they are simply what a curve shaped by a genuine, physically-forced seasonal storage cycle looks like at one instant, and the reason natural gas is this chapter's cleanest

illustration of §12.2's machinery actually operating as advertised.

## 12.4 Gold, Silver, and the Platform's Metals

Gold is the one commodity this book treats as closer to a monetary asset than to a physical input consumed by anyone. It is barely used up: essentially all gold ever mined still exists, in a vault, a piece of jewellery or a central bank reserve, so the production/consumption/storage triangle (§12.1) collapses for gold into overwhelmingly one thing — an enormous, slowly growing stock, with annual mine output a small fraction of that stock rather than a flow that clears against a comparably sized annual consumption the way a harvest does. What actually moves gold's price is not a physical scarcity story at all, but three channels closer to a rates or FX desk's dashboard than to a commodities desk's.

**Definition — Gold's three channels**

**Real-yield channel.** Gold pays no coupon and costs something to store, so its opportunity cost of holding is the real, inflation-adjusted yield available on a safe asset instead — conventionally read off the market for inflation-protected government debt (TIPS, in the US case). When real yields rise, the income foregone by holding a zero-yielding asset rises with them and gold tends to fall; when real yields fall, especially through zero, that opportunity cost shrinks or reverses and gold tends to rise. This is a monetary-economics stylised fact, stated here without a fresh derivation, and one of the more reliable macro correlations in commodities at a multi-week-and-longer cadence — not, and this matters for a platform trading a fifty-bar horizon, bar by bar.

**USD leg.** Gold is priced and traded globally in US dollars, so a broad-dollar move mechanically reprices gold in every other currency with no change in gold's own dollar price at all; because the same hawkish policy stance that lifts real yields frequently lifts the dollar too, the two channels often reinforce rather than offset each other.

**Haven regime.** In acute risk-off episodes, demand for gold as a store of value — an asset that is, distinctively among the platform's instruments, nobody's liability — can dominate the first two channels entirely for days at a time. Chapter 7's risk-on/risk-off treatment flagged gold as *conditional* precisely for this reason: its sign in a canonical risk-off week is not fixed the way the yen's or the Swiss franc's is, because safe-haven buying and a real-yield-driven or USD-driven sell-off can be present in the same episode, pulling in opposite directions. Chapter 14 works through exactly this conditionality as a two-phase pattern — an early liquidation phase in which gold is sold alongside everything else to raise cash, followed by a haven phase in which it decouples and rallies once the immediate cash-raising need has passed — and that two-phase structure is the concrete version of "conditional" this defbox otherwise leaves abstract.

For a first-order sense of the real-yield channel's rough scale — not a valuation model gold actually satisfies, and Chapter 13 is where this book treats real yields and duration-style sensitivities on their own terms — a simple sketch makes the direction and rough magnitude concrete.

**Result — A first-order real-yield sensitivity sketch**

Model gold's fair value, holding every other determinant fixed, as inversely related to a discount factor applied over a stylised sensitivity horizon $T$ against the prevailing real yield $\rho$: $S \approx \bar{S}\, e^{-\rho T}$, where $\bar{S}$ absorbs every other determinant of gold's level. Then $\ln S = \ln \bar{S} - \rho T$,

so $d(\ln S)/d\rho = -T$, and for a small change $d\rho$,

$$\frac{dS}{S} \approx -T\, d\rho.$$

$T$ here is not gold's duration in any bond-pricing sense — it is an assumed sensitivity scale for the scenario at hand, not an estimate this book derives from data; an empirical sensitivity would come from a regression on realised gold and real-yield data, not from a formula of this shape.

Take, for concreteness, $T = 10$ years and a starting gold price $S_0 = \$2{,}340.00$; real yields rise 30 bp ($d\rho = +0.30\%$):

$$\frac{dS}{S} \approx -10 \times 0.30\% = -\mathbf{3.0000}\%, \qquad dS = -3.0000\% \times 2{,}340.00 = -\$70.20, \qquad S_1 = \$\mathbf{2{,}269.80}.$$

A non-linearised check, holding $\bar{S}$ fixed at the value implied by $S_0$ and an assumed starting real yield $\rho_0 = 1.80\%$ (so $\bar{S} = S_0\, e^{\rho_0 T}$): the exact post-shock level is $S_1^{\text{exact}} = \bar{S}\, e^{-(\rho_0 + d\rho)T} = \$\mathbf{2{,}270.8425}$, against the linear approximation's \$2,269.80 — a difference of only \$1.0425 on a 30 bp move. The linear sketch is a reasonable stand-in for a modest shock; the gap widens, as it must for any first-order approximation, once $d\rho$ grows large enough that the ignored $\frac{1}{2}(T\, d\rho)^2$ term stops being negligible.

### The lease rate, and why gold's financing looks like an FX swap charge

Physical gold, like a currency, can be borrowed and lent: central banks and large bullion holders lend metal to bullion banks, who use it to finance market-making and other short positions, at a **lease rate** — the closest thing gold has to its own interest rate, quoted in the interdealer bullion market rather than set by any central bank. Chapter 4 showed that an FX swap prices a collateralised loan of one currency against another at an implicit rate equal to the two currencies' forward points, and that the nightly rollover or swap charge on every FX CFD position on the platform is exactly that same covered-interest-parity mechanism, renewed daily. Once gold is treated as a currency with its own interest rate — the lease rate standing in for what Chapter 4 called $r_f$ — the financing charge on a long XAU/USD position is structurally identical: an investor long gold forgoes the USD interest they could otherwise have earned, and forgoes, or in principle could earn, gold's own lease rate on the metal itself.

**Worked Example 12.3 — XAU/USD overnight financing: the lease-rate CIP mechanic**

A hypothetical XAU/USD CFD position, 100 oz long at $S = \$2{,}385.40$; illustrative rates only, to be verified against the platform's current published financing schedule before being used for anything beyond illustration: USD reference rate $r_{\text{usd}} = 5.10\%$, gold lease rate $= 0.35\%$, provider administrative spread $= 0.75\%$.
The core differential, mirroring Chapter 4's $r_d - r_f$ with gold's lease rate playing the foreign-currency rate:

$$r_{\text{usd}} - \text{lease} = 5.10\% - 0.35\% = \mathbf{4.7500}\%.$$

Financing rate charged on the long position, core differential plus the provider's spread:

$$4.7500\% + 0.75\% = \mathbf{5.5000}\%\ \text{p.a.}$$

Notional: 100 oz $\times$ \$2,385.40 = \$**238,540.00**. One night's financing charge, and a representative 30-day charge:

$$\$238{,}540.00 \times 5.5000\% \times \frac{1}{365} = \$\mathbf{35.9444}, \qquad \$238{,}540.00 \times 5.5000\% \times \frac{30}{365} = \$\mathbf{1{,}078.3315},$$

the latter 0.4521% of notional over the month. **The short side is not simply the mirror.** A short earns the core differential *less* the same admin spread, not plus it — $4.7500\% - 0.75\% = 4.0000\%$ p.a., a 30-day credit of \$784.2411 — so the spread is a genuine two-sided cost to the client on both legs, exactly a bid/ask around the core CIP-style rate rather than a single number applied with a sign flip; only that piece is negotiable in the ordinary sense. The core differential itself is not a fee at all, merely the arithmetic difference between two market-set rates, converted into cash — the same distinction [Course Book §22.1] draws for a WTI roll's cash cost, applied here to a financing rate instead of a roll gap.

### Silver: the same channels, a higher beta, and a channel gold lacks

Silver tracks gold directionally through the same real-yield and haven channels above, but with materially higher beta: a given move in real yields or in risk sentiment tends to produce a larger percentage move in silver than in gold, in both directions. The extra volatility is not free-floating noise — silver carries a substantial industrial-demand component (electronics and solar-panel manufacture prominent among current uses) that gold, used only marginally in industry, does not, so silver also answers to the global manufacturing cycle in a way gold structurally cannot. Stated as an order of magnitude, indicative and to be verified against current data before being relied on for sizing: gold's realised volatility typically runs in the low-to-mid teens per cent annualised (Chapter 14's own portfolio illustration uses 14.00% for exactly this reason), and silver's commonly runs close to double that, from the high teens into the twenties.

### XAU/XAG on the platform: quoted like FX, financed like FX

The platform's XAU/USD and XAG/USD CFDs are quoted in the same bid/ask-around-a-numeric-price convention as any FX pair — gold and silver behave, in the quote window, exactly like a currency versus the US dollar, the concrete sense in which §12.1 called them closer to a monetary asset than a physical input. The daily financing line item on both embeds precisely the lease-rate CIP mechanism Worked Example 12.3 derived — structural, in the sense that some such charge must exist as long as gold and silver have their own lease rates and USD has its own reference rate, but with every rate involved (USD reference, lease, admin spread) a drifting figure that must be verified against the platform's current published schedule before being used for anything beyond the illustrative arithmetic above.

## 12.5 Crude Oil Markets

**Definition — WTI and Brent: two benchmarks, two geographies**

**West Texas Intermediate (WTI).** A land-locked, physically-settled US benchmark grade, deliverable at Cushing, Oklahoma — a major pipeline and storage hub rather than a port, chosen historically for its pipeline connectivity across the US mid-continent. A long who has not closed the expiring contract must take delivery there; Chapter 10's settlement-method defbox and convergence result both apply to WTI without modification, and §12.5 below is what happens when the delivery point itself becomes the binding constraint.

**Brent.** A seaborne benchmark priced off North Sea crude loaded for cargo delivery, without a single fixed inland delivery point in the way Cushing is one — cargoes can, in principle, be redirected to whichever buyer values them most, a flexibility WTI's pipeline geography does not share. Brent is the more widely referenced benchmark for pricing crude internationally — most of the world's traded crude is priced at some contractual differential to Brent rather than to WTI — precisely because a waterborne grade travels; a land-locked one, however large the underlying market, prices first and foremost off its own regional balance.

Neither description carries a current WTI-Brent spread, a current output level, or any other drifting figure: the geography is structural and stable, the price relationship between the two benchmarks is not, and must be read off a current quote before being used for anything beyond the structural point made here.

### Cushing as a physical chokepoint

§12.2's storage-cost term $u$ was treated throughout as a smooth, finite proportional rate — a reasonable assumption whenever storage capacity itself is not the binding constraint. Cushing's tank capacity is finite, and finite in a way that matters specifically for the expiring front-month WTI contract: once regional storage is close to full, the marginal barrel arriving with nowhere to go does not face a slightly higher storage *rate*, it faces the prospect of no available storage at any rate. (12.1)'s $u$ implicitly assumes a seller of storage willing to take the marginal barrel at some finite price; when that willingness genuinely disappears, the formula's own assumption — not merely its calibration — stops holding.

**Trap — Physical delivery breaks the usual zero floor**

Ordinary price intuition assumes an asset is worth at least zero, on the reasoning that an owner can always dispose of it for free (*free disposal*) rather than pay to hold something worthless. A contract that obliges its holder to take physical delivery of a good that is costly, or unlawful, to simply discard has no such guarantee: if nobody wants the barrels and there is nowhere left to lawfully put them, the rational price for the obligation is whatever it costs to induce someone else to take it off your hands — a price below zero. On 20 April 2020, the CME's expiring May WTI contract traded and settled at a negative price for the first time in the contract's history — the publicly reported instance of exactly this mechanism, Cushing's storage constraint biting on a contract nobody wanted to physically hold, with global demand having collapsed under the first wave of pandemic lockdowns. Chapter 15 examines that episode in full, as one of six crisis case studies, including what it did to a systematic pipeline's own price-transform assumptions; the mechanics above are what that chapter's narrative sits on top of, not a preview of it, and this chapter goes no further into the specific day's trading.

### OPEC as a supply regime-setter

Agricultural supply (§12.1) is set by nature's own calendar; industrial-metals supply is close to continuously adjusting mine output. Crude's marginal supply is, at the OPEC and OPEC+ coalition's discretion, neither: a group of major exporters coordinates production quotas among its members, which makes a meaningful share of global supply a discrete policy decision taken at scheduled meetings rather than a continuous, price-taking response to the prevailing price — a structural fact about how the market's supply curve is actually set, stable enough to state here, though the coalition's membership, quotas and compliance are themselves drifting facts requiring current verification before any of them enter an actual trading decision. The practical consequence for §12.2's machinery: a curve move driven by an OPEC+ meeting outcome is a supply-regime shock hitting $S$, and via expectations the whole curve at once, not a change in $u$ or $y$ that a storage-and-convenience reading alone would explain.

### The lesson for a CFD holder: roll policy and margin around expiries

Chapter 10 already established that the platform's own WTI CFD never touches a futures contract, a clearing house or a roll calendar directly — OANDA performs the roll on the client's behalf and passes its cash consequence through as a financing line, exactly as §12.6 below quantifies. What this chapter's chokepoint logic adds is why the days immediately around a physically-settled

contract's expiry are not just another ordinary roll: providers routinely widen spreads and raise margin requirements in the roll window specifically, and in some cases force an early roll well before expiry rather than carry a client position into the window where delivery risk, and the pricing distortion Trapbox 12.5 describes, could in principle bite. None of this is a fixed rule this book can state as a day-count — it is provider policy, verify-current by construction — but the direction is structural and worth internalising before it is ever needed: a cost model built from an ordinary week's spread and margin data will not see an expiry-window regime shift coming, because it was never built to look at the calendar in the first place.

## 12.6 Commodity Returns, Honestly

### Two camps, a shared dataset, and a disagreement that never fully resolved

Gorton, Rouwenhorst (2006) constructed a broad, equally-weighted index of fully-collateralised commodity futures — long the front future, the freed-up capital earning a T-bill rate as collateral — and reported historical returns comparable to equities over their sample, with low correlation to stocks and bonds and, notably, periods of backwardation contributing meaningfully to that return: a result read, fairly, as historical support for Keynes's (1930) normal-backwardation story, applied now to a diversified index rather than a single curve. Erb, Harvey (2006) pushed back on the same broad dataset from several directions at once: index-level historical returns are highly sensitive to which commodities are included and how they are weighted, so a different reasonable index construction on the same underlying markets can tell a noticeably different story; the roll-yield contribution specifically is time-varying, tracking whether the average constituent happened to sit in backwardation or contango over the sample period studied, a fact about that period's curve shapes rather than a stable structural premium in the way an equity risk premium is normally understood; and a meaningful share of a diversified commodity index's apparent historical outperformance is attributable to a *rebalancing* or *diversification return* — an artefact of periodically rebalancing a basket of volatile, imperfectly correlated assets back to fixed weights — rather than to any single constituent's own risk premium.

§12.2's reconciliation of Working and Keynes applies again here, one level up: both readings can be simultaneously true of the same historical data — genuine periods of normal backwardation *and* a return decomposition sensitive enough to index construction and rebalancing mechanics that headline historical figures should never be taken as a stable, forward-looking risk premium available for the taking. The honest summary this book gives is the one Chapter 10 already modelled for a single roll: roll yield is a property of the curve the position inherits, not a fee anyone invented, but whether that property has historically summed to a genuine risk premium, a rebalancing artefact, both, or neither, is exactly the kind of claim this book's evaluation discipline ([Course Book §15.4]) requires clearing a test for, not a fact a citation alone settles.

### What a leveraged CFD account actually earns

**Result — The CFD-level return decomposition**

For a long position of $N$ units entered at $P_0$ and marked at $P_1$, held through one or more rolls each with cash drag $F_{2,i} - F_{1,i}$ per unit ([Course Book §22.1]), and charged financing at annualised rate $f$ on the entry notional $NP_0$ over horizon $T$:

$$\text{net P\&L} = \underbrace{N(P_1 - P_0)}_{\text{spot P\&L}} - \underbrace{N\sum_i (F_{2,i} - F_{1,i})}_{\text{roll drag}} - \underbrace{NP_0\, f\, T}_{\text{financing}}.$$

A short position flips the sign of every term. Each roll's dollar drag is exactly $\delta_i \times (NF_{1,i})$ for that roll's own $\delta_i = (F_{2,i} - F_{1,i})/F_{1,i}$ — [Course Book §22.1]'s own per-roll formula, applied

once per roll rather than extrapolated across a year at one assumed constant $\delta$; summing several genuinely different $\delta_i$ over a real holding period needs none of [Course Book §22.1]'s simple-versus-compounded distinction, which exists only to extrapolate a single assumed roll charge across many repetitions, not to add up already-realised, already-different cash flows.

**Worked Example 12.4 — A WTI CFD's return, decomposed over a quarter**

A hypothetical long WTI CFD position, $N = 100$ barrels-equivalent, entered at $P_0 = \$78.40$ and marked at quarter-end $P_1 = \$82.15$; illustrative prices only. Three rolls occur during the quarter, at $(F_1, F_2)$ pairs $(78.75, 79.40)$, $(80.10, 80.55)$ and $(81.60, 81.95)$ — all three in mild contango — and the position is charged financing throughout at $f = 6.00\%$ p.a. on the entry notional, $T = 91/365 = 0.249315$ years for the quarter.

**Spot P&L.** $N(P_1 - P_0) = 100 \times (82.15 - 78.40) = \$\mathbf{375.00}$.

**Roll drag, roll by roll** (per-roll $\delta_i$, [Course Book §22.1]'s formula, alongside the dollar drag):

| Roll | $F_1 \to F_2$ | $\delta_i = (F_2 - F_1)/F_1$ | Dollar drag $= 100(F_2 - F_1)$ |
|---|---|---|---|
| 1 | $78.75 \to 79.40$ | 0.8254% | \$65.00 |
| 2 | $80.10 \to 80.55$ | 0.5618% | \$45.00 |
| 3 | $81.60 \to 81.95$ | 0.4289% | \$35.00 |

Total roll drag $= \$65.00 + \$45.00 + \$35.00 = \$\mathbf{145.00}$.

**Financing.** Entry notional $= 100 \times 78.40 = \$7{,}840.00$; charge $= \$7{,}840.00 \times 6.00\% \times 0.249315 = \$\mathbf{117.2778}$.

**Net P&L.**

$$\$375.00 - \$145.00 - \$117.2778 = \$\mathbf{112.7222}.$$

Roll drag plus financing together consume 69.9407% of the gross spot P&L, leaving only 30.0593% of it as the position's actual net result — and expressed against the capital the position actually committed rather than against its own gross P&L, the spot move alone was a 4.7832% return on the \$7,840.00 notional, while the net result was 1.4378%: the all-in cost of holding the position for the quarter, roll and financing together, ran to 3.3454% of notional, on a curve that never left mild contango throughout.

**Reconciling against the academic construction.** Gorton and Rouwenhorst's (2006) collateralised index earns a T-bill-style yield on the *full* notional as compensation for the capital the strategy is assumed to post; a leveraged CFD account posts only margin, not full notional, and earns no such collateral yield at all — it is simply not credited. At an illustrative collateral rate of 5.00% p.a., the same \$7,840.00 notional would have earned $\$7{,}840.00 \times 5.00\% \times 0.249315 = \$97.7315$ over the quarter in the academic construction; added to the \$117.2778 the CFD account was actually charged (a fully-collateralised position would not be charged financing at all, since it is not borrowing anything), the total swing between the two constructions is $\$97.7315 + \$117.2778 = \$\mathbf{215.0093}$ over one quarter, on this one hypothetical position — the concrete size of the gap between what a Gorton-Rouwenhorst-style historical risk premium describes and what a real, leveraged CFD account is structurally able to capture of it, even if every word of the risk-premium story were true.

### Positioning data: at home in commodities in a way it is not in FX

Chapter 7 built the CFTC's Disaggregated report in full — Producer/Merchant/Processor/User (PMPU, this section's shorthand from here on), Swap Dealers, Managed Money and Other Reportables — and flagged that it exists only for contracts traded on a US, CFTC-regulated futures exchange, exactly why Chapter 7 needed a proxy construction to say anything about currency

pairs that list no such future of their own. WTI and the platform's metals face no such gap: each has a genuine, CFTC-regulated futures market underneath it, so the Disaggregated report's categories apply directly, with no proxy step required, and every reading Chapter 7 built — the $z$-score, the trend-confirmation-versus-contrarian ambiguity, the Managed-Money-against-PMPU divergence reading — carries over to a commodity unmodified. Positioning data is, in this narrow but real sense, more naturally at home in commodities than in almost any other corner of this book's four pillars.

**Your project — XAU, XAG and WTI on the platform: v4's accounting job**

v4 of the SST roadmap is commodities: the platform's XAU/USD, XAG/USD and WTI CFDs, sitting alongside the other eighteen OANDA instruments the platform already trades. Each of the three inherits a carry mechanism this chapter has now made explicit, and none of the three inherits the same one: WTI's carry is Working's storage-and-convenience curve (§12.2), realised at each roll exactly as Worked Example 12.4 decomposed it; gold's and silver's carry is the lease-rate CIP mechanism (§12.4), realised daily rather than at discrete rolls. v4's ledger inherits Chapter 10's non-negotiable separation — the *forecast* component (bar-by-bar mark-to-market on the tracked price, the only piece TimesFM's quantile head has any opinion about) kept explicitly apart from the *carry* component (roll drag or lease-rate financing, set by the market before the position is ever opened) — and this chapter adds one further discipline the separation must not silently violate: nowhere does a real v4 account earn the collateral yield Worked Example 12.4's reconciliation showed a fully-collateralised academic index assumes. A v4 backtest that credits the strategy with a risk-free return on notional it never actually held as posted collateral is manufacturing exactly the kind of edge Chapter 10's own trapbox warned an unaccounted roll term can manufacture — now from the other side of the ledger, crediting income the account structurally never receives rather than omitting a cost it structurally does pay.

## 12.7 Exercises

**Exercise 12.1 — Convenience-yield back-out: copper, two curves**

Two hypothetical copper curves, illustrative figures only, each with $u$ supplied as an independent storage-and-insurance estimate.
**(a) Tight market.** $S = 4.15$, $F = 4.09$, $r = 4.80\%$, $u = 1.80\%$, $T = 45/365$ years.
**(b) Ample market.** $S = 3.85$, $F = 3.93$, $r = 4.50\%$, $u = 1.60\%$, $T = 120/365$ years.
For each part: compute $\ln(F/S)$, $r + u$, the implied convenience yield $y$ via Result 12.2, and classify the curve as contango or backwardation, stating whether $y$ exceeds or falls short of $r + u$. Confirm each answer by substituting back into (12.1).

▷ working space

**Solution — Exercise 12.1**

**(a)** $T = 45/365 = 0.1232877$ years. $\ln(4.09/4.15) = -0.01456336$. $r + u = 6.6000\%$.

$$y = 6.6000\% - \frac{-0.01456336}{0.1232877} = 6.6000\% + 11.8125\% = \mathbf{18.4125}\%.$$

Check: $4.15 \times e^{(0.048+0.018-0.184125)\times 0.1232877} = \mathbf{4.09}$, matching $F$ exactly. Since $F < S$ and $y = 18.4125\% \gg r + u = 6.6000\%$: **backwardation**, a tight physical copper market with a substantial scarcity premium, comparable in scale to Worked Example 12.1's pre-harvest wheat case.

**(b)** $T = 120/365 = 0.3287671$ years. $\ln(3.93/3.85) = 0.02056628$. $r + u = 6.1000\%$.

$$y = 6.1000\% - \frac{0.02056628}{0.3287671} = 6.1000\% - 6.2556\% = -\mathbf{0.1556}\%.$$

Check: $3.85 \times e^{(0.045+0.016-(-0.001556))\times 0.3287671} = \mathbf{3.93}$, matching $F$ exactly. Since $F > S$ and $y = -0.1556\% < r + u = 6.1000\%$: **contango**. As in Worked Example 12.1's Part B, the backed-out $y$ is again slightly negative — read honestly as the curve pricing marginally more carry cost than the assumed 1.60% storage rate supplies, not as evidence copper's convenience yield is a stable negative number worth trading on.

**Exercise 12.2 — The full CFD return decomposition: short natural gas**

A hypothetical short natural-gas CFD position, $N = 2{,}000$ units, entered at $P_0 = 3.20$ and marked at quarter-end $P_1 = 2.95$. Three rolls occur, all in contango, at $(F_1, F_2)$ pairs $(3.22, 3.30)$, $(3.15, 3.19)$ and $(3.05, 3.11)$ — recall a short's roll drag flips sign against Result 12.6's long-position convention, so a short is *credited* in contango. Financing on the short side nets a reference rate of 5.00% p.a. against a 3.25% administrative spread (the short earns the reference rate less the spread, mirroring Worked Example 12.3's long/short asymmetry), over $T = 91/365$ years.

(a) Compute the spot P&L.

(b) Compute each roll's per-unit $\delta_i$ and the total roll credit received by the short.

(c) Compute the financing credit.

(d) Assemble the net P&L, and express both the spot P&L and the net P&L as a percentage of entry notional.

▷ working space

**Solution — Exercise 12.2**

**(a)** Short gains when price falls: spot P&L $= N(P_0 - P_1) = 2{,}000 \times (3.20 - 2.95) =$ **\$500.00**.
**(b)**

| Roll | $F_1 \to F_2$ | $\delta_i$ | Long drag = short credit |
|---|---|---|---|
| 1 | $3.22 \to 3.30$ | 2.4845% | \$160.00 |
| 2 | $3.15 \to 3.19$ | 1.2698% | \$80.00 |
| 3 | $3.05 \to 3.11$ | 1.9672% | \$120.00 |

Total roll credit $=$ \$160.00 $+$ \$80.00 $+$ \$120.00 $=$ **\$360.00**.
**(c)** Notional $= 2{,}000 \times 3.20 =$ \$6,400.00. Credit rate $= 5.00\% - 3.25\% = 1.7500\%$ p.a. Financing credit $=$ \$6,400.00 $\times$ 1.7500% $\times$ (91/365) $=$ **\$27.9233**.
**(d)** Net P&L $=$ \$500.00 $+$ \$360.00 $+$ \$27.9233 $=$ **\$887.9233**. As a percentage of the \$6,400.00 notional: spot P&L alone is 7.8125%, and the net result, boosted rather than eroded by every non-spot component, is 13.8738% — the mirror image of Worked Example 12.4's long WTI case, where roll and financing *consumed* most of the spot gain. This short is paid on every leg of the decomposition by construction of the scenario — a falling price the position is short, a contango curve the short is rolling through profitably, and a credit rather than a charge on financing — not a general property of short commodity positions, which face exactly the reverse in a backwardated market with rates working against them.

**Exercise 12.3 — Gold scenario: real yields fall 50 bp**

Real yields fall 50 bp from their current level; gold currently trades at $G_0 =$ \$2,410.00. Using Result 12.4's linear sketch with the same illustrative sensitivity horizon $T = 10$ years used in the main text:

(a) Walk through the mechanism, in words, that connects a falling real yield to a rising gold price.
(b) Compute $dG/G$, the implied $dG$, and $G_1 = G_0 + dG$.
(c) State at least two conditions under which this mechanism's clean prediction should *not* be taken at face value, drawing on this chapter's own caveats about the sketch and on Chapters 7 and 14's treatment of gold's other channels.

▷ working space

**Solution — Exercise 12.3**

**(a)** Gold pays no coupon, so holding it costs whatever a safe real-yielding alternative would otherwise have paid. When real yields fall, that foregone income shrinks, which lowers the opportunity cost of holding a zero-yielding asset and, all else equal, makes gold more attractive to hold at the margin — pushing its price up until the market once again prices gold as fairly compensating for a now-smaller foregone yield.

**(b)** $d\rho = -0.50\%$:

$$\frac{dG}{G} \approx -T\, d\rho = -10 \times (-0.50\%) = +\mathbf{5.0000}\%.$$

$$dG = 5.0000\% \times 2{,}410.00 = \$\mathbf{120.50}, \qquad G_1 = 2{,}410.00 + 120.50 = \$\mathbf{2{,}530.50}.$$

**(c)** Three caveats, none of which the linear sketch can see. First, the sketch is a first-order approximation with an assumed, not estimated, sensitivity $T$ — Result 12.4 flagged this explicitly, and a different assumed $T$ would scale $dG/G$ proportionally with no way to prefer one $T$ over another from the formula alone. Second, *why* real yields fell matters and the sketch cannot distinguish the reasons: a fall driven by disinflation with growth intact is the clean case the mechanism describes, but a fall driven by a growth scare severe enough to also trigger a broad flight to the US dollar could see gold's USD leg (§12.4) partly offset the real-yield tailwind, exactly the ambiguity Chapter 7 flagged in classifying gold's sign within a risk-off week. Third, and most concretely, if the same episode is severe enough to trigger acute risk-off conditions, Chapter 14's two-phase pattern applies: an initial liquidation phase in which gold is sold alongside risk assets to raise cash can temporarily push gold *down* even as real yields fall, with the real-yield-driven rally only asserting itself once that cash-raising phase has passed — so the sign in part (b) may be right in direction over a multi-week horizon while being badly wrong, for days at a time, about the path taken to get there.

**Exercise 12.4 — Reading a disaggregated commodity snapshot: producers short into a rally**

A hypothetical Disaggregated-style weekly snapshot for a copper future — illustrative figures only — over four consecutive weeks:

| Week | Price | PMPU net | Managed Money net |
|---|---|---|---|
| 1 | 4.20 | −18,000 | +22,000 |
| 2 | 4.35 | −24,500 | +28,000 |
| 3 | 4.52 | −31,000 | +35,500 |
| 4 | 4.68 | −39,000 | +44,000 |

(a) Compute the percentage change from Week 1 to Week 4 in price, in the magnitude of PMPU's net position, and in Managed Money's net position.

(b) PMPU and Managed Money sit on opposite sides throughout, and both positions are *growing* more extreme as price rallies. State the two readings Chapter 7's divergence framework offers for this pattern, applied to this specific snapshot, without adjudicating between them.

(c) Is the disaggregated view available for this instrument by the ordinary CFTC reporting route, or does it require Chapter 7's futures-as-proxy construction? Justify briefly.

▷ working space

**Solution — Exercise 12.4**

**(a)** Price: $(4.68 - 4.20)/4.20 = \mathbf{11.4286}\%$. |PMPU net|: $(39{,}000 - 18{,}000)/18{,}000 = \mathbf{116.6667}\%$. Managed Money net: $(44{,}000 - 22{,}000)/22{,}000 = \mathbf{100.0000}\%$. Both positioning categories have grown by materially more, in percentage terms, than price itself has moved — the week-on-week detail confirms it is not a one-week spike: PMPU's net short widens by 6,500, 6,500 and 8,000 across the three weekly steps and Managed Money's net long widens by 6,000, 7,500 and 8,500, a steady build on both sides through all four weeks.
**(b) Trend confirmation.** Managed Money, the speculative bucket, is reading a genuine bullish move correctly and adding to it every week; PMPU's mirror-image growing short is not bearish information at all, but ordinary commercial hedging — producers locking in increasingly favourable forward sale prices as the rally extends, exactly what a hedger is supposed to do as price rises, corroborating rather than contradicting the move. **Contrarian at extremes.** Managed Money's net long is becoming steadily more crowded, and by Week 4 a large share of the willing marginal speculative buyer base may already be positioned, leaving progressively less fresh demand to extend the rally further and an increasingly one-directional exposure to any shock (Brunnermeier, Nagel, Pedersen, 2008); on this reading, PMPU's widening short is not passive hedging but the commercial side — typically closer to the physical market's own supply-and-demand information — expressing a view that the rally has overrun the physical fundamentals, selling forward specifically *because* they judge the price unsustainable. Chapter 7 states both readings without resolving the ambiguity, and this snapshot is constructed, like Chapter 7's own example, to admit both at once.
**(c)** Directly available: copper is a physical commodity with a genuine CFTC-regulated futures market, so the Disaggregated report applies without Chapter 7's proxy construction, which exists specifically to cover instruments — FX pairs prominent among them — with no CFTC-regulated future of their own. §12.6 made exactly this point: positioning data is structurally more directly available for a commodity like copper than for almost any currency pair on the platform.

**Key idea — What to remember**

A commodity's price is the resolution of three physical flows — production, consumption and the storage that bridges them — and that same storage state is what a futures curve prices through Working's (1949) convenience yield $y$ against the financing-and-storage cost $r + u$: back the two apart, given an independent storage estimate, and a single curve reading becomes a genuine statement about physical scarcity, never a forecast of where spot is headed, Keynes's (1930) normal-backwardation risk-premium story notwithstanding until it clears an honest test. That test is exactly what Gorton, Rouwenhorst (2006) and Erb, Harvey (2006) ran against each other on historical commodity-index returns, and the honest reading is that both roll yield and rebalancing mechanics are real, overlapping, sample-sensitive contributors to what a diversified index has earned — and, separately again, what a real leveraged CFD account keeps of any of it, after roll drag and financing and without the collateral yield an academic index construction assumes, is smaller still, exactly as [Course Book §22.1] predicted and this chapter's own WTI decomposition confirmed to the cent. Seasonality is genuinely real wherever the physical calendar forces it — natural gas, heating oil, the harvest cycle — and genuinely absent almost everywhere else, gold's own case included, with no exemption from [Course Book §22.2]'s testing discipline for having a plausible mechanism attached. Gold and silver behave less like physical commodities than like a currency with its own lease-rate interest rate, financed on the platform exactly as an FX position is; WTI inherits crude's own delivery geography, Cushing's storage constraint chief among it, closely enough that a CFD holder's roll-window caution is not superstition but a direct consequence of the mechanics

this chapter derived. Every rate this chapter treated as a given input — the financing rate, the real yield, the lease rate — arrives from outside the commodity market itself; the next chapter turns to where it comes from.

PART V

# The System View



LUNA LABS

# 13

# Rates, Money, and Macro Transmission

*Every chapter so far has treated an interest rate as an input arriving from outside the frame: a differential to be quoted in a covariate table, a financing rate applied at rollover, a discount factor multiplying a dividend stream. This chapter is the outside. Central banks set a short rate deliberately, for macroeconomic reasons that have nothing to do with any one currency pair, equity index, or gold price; that short rate propagates through an entire term structure by nothing more exotic than the same no-arbitrage arithmetic Chapter 5 used to price a forward; and the calendar on which it moves — the meeting dates, the data releases — is, uniquely among the objects this book has catalogued, a piece of the future the market is handed in advance, in writing, months ahead of the event. None of this is a bond desk's brief, and this chapter does not attempt one: no convexity trading, no swaption volatility, no repo specialness. What it owes the reader is narrower and more useful — enough of the curve to read a differential correctly, enough of duration to see why a rate move sized in bond points and a rate move sized in FX pips are frequently the same fact measured twice, and a disciplined account of how a single short-rate surprise propagates, with sign, through FX, equities, gold, and the funding cost of every levered position this platform, or its reader, might ever hold.*

## 13.1 Policy rates and the short end

A central bank does not set bond yields, exchange rates, or bank lending rates directly; what it sets is an *operational target* for the overnight rate at which banks lend reserves to one another — the US federal funds rate, the Reserve Bank of Australia's cash rate, and their analogues elsewhere. Almost everything else this chapter treats as a market object is, at some remove, that one administered number propagating through a chain of arbitrage and expectation.

**Definition — The policy rate, and the corridor that anchors it**

The usual mechanism enforcing an operational target is a *corridor*. The central bank offers a deposit rate to banks that park reserves with it overnight, which no bank will lend below — why accept less from another bank than the central bank itself pays for the same riskless overnight placement; and a lending rate at which banks short of reserves can borrow directly from the central bank, which no bank will pay above — why pay another bank more than the central bank itself charges. The overnight market rate is squeezed between the two, and the announced "target" or "target range" sits inside, or at one edge of, that band — a rate the central bank *enforces* through its own balance sheet, not merely announces. Which facility a given central bank currently emphasises, and the exact width of its corridor, are matters of each institution's own operational framework, revised periodically, and should be verified against its current published framework rather than assumed.

The overnight rate is the base of a chain, not the whole curve. A term money-market rate — a 90-day bank bill, a term deposit, the rate a CFD provider itself pays to fund its own book for longer than one night — is priced, to a first approximation, as the market's expectation of the average

overnight rate over that term, plus a small term or liquidity premium for the inconvenience of committing funds rather than rolling them nightly. This is the expectations hypothesis of the term structure in miniature, one rung below the government curve Section 13.2 treats formally, and it is why a change in the policy rate propagates to every floating-rate loan, every term deposit, and every CFD provider's own funding cost within days, without the central bank needing to touch any of those rates directly: it moves the anchor, and the chain of expectation does the rest.

Central banks do not reset the policy rate continuously; they meet on a published schedule and change it, if at all, only at those meetings — an unscheduled inter-meeting move is rare enough to be newsworthy in itself. The US Federal Open Market Committee holds eight scheduled meetings a year, the cadence Chapter 5 already used, indicatively, to size how often a genuine policy surprise can occur at all. The Reserve Bank of Australia has met on the same eight-a-year cadence since February 2024, having reduced its own schedule from eleven meetings a year following a 2023 review of its operational framework — a concrete illustration, not a hypothetical one, of why this book insists meeting cadences, and even institutional structures themselves, drift and must be checked against each central bank's current published calendar rather than assumed fixed. What is stable is the mechanism, not the count: a scheduled decision, communicated with a statement and, for the larger central banks, a press conference and, for some, published economic projections.

What the market expects a given meeting to deliver is not a matter of survey opinion; it is priced, and that machinery is Chapter 5's to derive, not this chapter's to repeat. Chapter 5 showed that a meeting-dated overnight index swap prices an implied probability $p = m/\delta$ for a policy increment $\delta$ against $m$ basis points priced, under the simplifying restriction that the meeting either delivers $\delta$ or holds — a market *price*, not a survey probability, and contaminated by whatever term premium the instrument itself carries. The US market runs the equivalent construction on Fed funds futures, exchange-traded and centrally cleared rather than an OTC swap, which settle to the realised average daily effective rate over the contract month: a Fed funds future's price relative to the current policy rate backs out a meeting-implied probability by the same $p = m/\delta$ arithmetic once a single meeting's increment is isolated from a contract month that may straddle more than one decision. Every caveat Chapter 5 attached to the OIS version — the binary restriction, the ignored term premium, the fact that $p$ is a price and not a forecaster's probability — applies unchanged to this instrument too.

## 13.2 The yield curve

Section 13.1 pinned one point on the curve: the overnight rate, and the market's priced expectation of it out to the next meeting or two. Everything from there out to two years, ten years, or thirty is a separate, related price, and it is worth having language for the whole object before asking what its shape means.

**Definition — The zero-coupon (spot) rate**

The *zero-coupon rate*, or *spot rate*, $z_n$ for maturity $n$ years is the annualised return on a hypothetical instrument that pays nothing until $n$ years from now and then pays exactly its principal: one unit invested today at $z_n$ becomes $(1 + z_n)^n$ at maturity, with no coupon along the way to reinvest at some other, uncertain rate. The associated *discount factor* $\mathrm{DF}(n) = (1 + z_n)^{-n}$ is the price today of one unit of currency certain at $n$ years, and a full *yield curve* is nothing more than the function $n \mapsto z_n$ read off market prices at every traded maturity. Real government bonds mostly pay coupons rather than a single redemption at maturity, so a zero curve is not usually observed directly; it is *stripped* out of a set of coupon-bond prices by exactly the no-arbitrage logic Result 13.2 uses one step further on — a coupon bond is only a bundle of zero-coupon cash flows, one per coupon date plus the

final redemption, and its price must equal the sum of those flows each discounted at the zero rate for its own maturity. The *par yield* for a maturity is the single coupon rate that would price a bond of that maturity at exactly 100 given the stripped zero curve; this book states the fact and moves on, since the bootstrap itself is a bond-desk exercise this chapter deliberately does not set.

A curve built this way prices money at every maturity *today*. The next question is what, if anything, it says about rates the market expects *between* two future dates — and the answer is pinned by an argument this book has run once already.

**Result — The forward rate, from no-arbitrage**

Fix two horizons $a$ and $a+b$ years ahead, with zero rates $z_a$ and $z_{a+b}$. The *forward rate* $f_{a,b}$ is the interest rate, agreed *today*, for a loan that begins in $a$ years and runs a further $b$ years. It is pinned by an argument identical in spirit to Chapter 5's proof of covered interest parity, with one currency doing the work of two.

*Route 1 (go long, direct).* Invest one unit of currency today at the $(a+b)$-year zero rate. At $a+b$ years it is worth $(1+z_{a+b})^{a+b}$.

*Route 2 (go short, then roll at a rate fixed today).* Invest the same one unit today at the $a$-year zero rate, maturing at $a$ years to $(1+z_a)^a$ — a quantity known today. Simultaneously, today, enter a forward-rate agreement committing both parties, with no cash changing hands until year $a$, to reinvest whatever matures at $a$ years for a further $b$ years at the rate $f_{a,b}$. At $a+b$ years, Route 2 is worth $(1+z_a)^a(1+f_{a,b})^b$.

Both routes start from one unit of currency, require no further capital, and carry no reinvestment-rate risk — the forward-rate agreement removes exactly the uncertainty that would otherwise attach to reinvesting at whatever the $a$-to-$(a+b)$-year rate turns out to be. If the two routes disagreed at maturity, borrowing on the cheaper one and lending on the dearer one would be a riskless, self-financing profit, exactly as in Chapter 5's arbitrage argument for a mispriced FX forward. No arbitrage therefore forces

$$\boxed{(1+z_a)^a(1+f_{a,b})^b \;=\; (1+z_{a+b})^{a+b}} \qquad \Longleftrightarrow \qquad f_{a,b} \;=\; \left[\frac{(1+z_{a+b})^{a+b}}{(1+z_a)^a}\right]^{1/b} - 1.$$

**Worked Example 13.1 — A 1-year rate, 1 year forward, from the 1y and 2y zeros**

A hypothetical curve, not market data. The 1-year zero rate is $z_1 = 4.2000\%$ and the 2-year zero rate is $z_2 = 4.5500\%$. Find the 1-year rate implied for the period from year 1 to year 2 — market shorthand "1y1y" — and say what its position relative to $z_1$ and $z_2$ reveals.

Specialising Result 13.2 to $a = b = 1$,

$$f_{1,1} = \frac{(1+z_2)^2}{1+z_1} - 1.$$

$$1+z_1 = 1.042000, \qquad (1+z_2)^2 = 1.045500^2 = 1.093070,$$

$$f_{1,1} = \frac{1.093070}{1.042000} - 1 = 0.049012 = 4.9012\%.$$

The forward rate, 4.9012%, exceeds both spot rates, $z_2 = 4.5500\% > z_1 = 4.2000\%$: the general pattern on an upward-sloping curve, and it is arithmetic, not opinion — a two-year zero can only average a higher return than a one-year zero if the *second* year's implied rate more than compensates for the first, exactly as a class average sitting above every already-recorded score requires the next score to be higher still. Whether $f_{1,1}$ is best read as the

market's genuine expectation of the 1y1y spot rate in a year's time, or as today's expectation contaminated by a term premium, is Chapter 5's forward-premium question asked of a single currency's own curve rather than of two currencies' spot and forward exchange rates — and the honest answer is the same: some of both, in a proportion this book does not attempt to estimate. (All figures verified in `verify_mk13.py`, §1.)

**Trap — Two different objects share the name "forward"**

Chapter 5 also has a "forward rate": the differential embedded in an FX forward's swap points, extracted by inverting the covered-interest-parity identity. That object and $f_{a,b}$ above are priced by the same style of no-arbitrage argument, but they are not the same number and do not answer the same question. Chapter 5's forward rate is a statement about *two currencies at one maturity* — how much one currency's money-market rate exceeds the other's, embedded in a single outright forward price. This section's forward rate is a statement about *one currency at two maturities* — how the market prices reinvestment risk along its own curve. Confusing them is an easy and costly error: reading a steep domestic yield curve as if it revealed something about a currency forward, or a rich currency forward as if it revealed something about the domestic curve's slope. Keep the question each answers straight — which market, and how many maturities.

The single most quoted curve statistic is the *2s10s spread*, $z_{10} - z_2$: positive (the ten-year yields more than the two-year) is called *normal* or *upward-sloping*; a spread compressing toward zero is *flattening*; a negative spread is *inverted*. The conventional reading treats a positive 2s10s as the market discounting further growth and, eventually, further tightening, and an inversion as the market pricing cuts — the curve, on this reading, encoding a recession forecast the way any expectations-hypothesis argument would predict: if short rates are expected to fall, a long-maturity bond that locks in today's decent yield for years is worth more than a short-maturity one that will only get to reinvest, in a year, at whatever the (by then lower) prevailing rate has become, so long yields sit below short ones.

| Shape | 2s10s sign | Conventional reading (untested) |
|---|---|---|
| Normal / upward | $z_{10} - z_2 > 0$ | Growth and/or further tightening priced ahead |
| Flat | $z_{10} - z_2 \approx 0$ | Transition; policy near a perceived peak or trough |
| Inverted | $z_{10} - z_2 < 0$ | Cuts, and often a slowdown, priced ahead |
| Humped | sign changes along the curve | Near-term tightening, longer-term normalisation |

The honest caveat belongs immediately next to that story, not in a footnote. Chapter 5 already showed that the expectations hypothesis of the term structure is itself empirically rejected — Campbell and Shiller (1991) found yield spreads predicting subsequent short-rate changes with the *wrong* sign at short horizons, the same structural failure Fama (1984) documented for the FX forward premium, and for the identical reason: a forward-looking price decomposes into an expectation and a time-varying risk premium — here, a term premium — and the premium is the more variable component. An inverted curve is therefore not a clean forecast of future cuts; it is a forecast contaminated by whatever is moving the term premium at the time, which historically has often itself been recession risk, which is exactly why the correlation between inversions and subsequent recessions looks as strong as it does in hindsight while remaining a poor, thin basis for a real-time trading rule — any such rule this platform ever considered would need to clear the deflation and multiple-testing discipline [Course Book §15.4] before a single basis point were sized against it. A curve shape is a fact about prices, observed with certainty; what it forecasts is a hypothesis, tested like any other.

## 13.3 Duration in one section

A yield curve is a set of prices for money at different maturities; *duration* answers the question a position actually asks: if the curve, or the one point on it a given bond sits at, moves by a small amount $dy$, how much does the bond's price move? The relationship is linear to a first approximation and needs exactly one number to state.

**Definition — Modified duration**

For a bond priced at $P$ as a function of its own yield $y$, the *modified duration* is

$$D = -\frac{1}{P}\frac{dP}{dy}, \qquad \text{so that} \qquad \frac{dP}{P} \approx -D\,dy$$

to first order in a small yield change $dy$. $D$ is measured in years, is positive for an ordinary (option-free, long) bond, and the minus sign is the whole content of the relationship: yields up, price down. Under continuous compounding, $D$ equals the bond's *Macaulay duration* — its cash-flow-weighted average time to payment — exactly; under the market's usual periodic compounding, modified duration is Macaulay duration divided by $(1 + y/m)$ for $m$ compounding periods a year, a distinction Hull (2021) treats in full and this chapter states without re-deriving.

**Result — Duration and convexity of a zero-coupon bond**

The general coupon-bond case is Hull's (2021) to derive; the zero-coupon case is three lines and is worth doing, because it is exactly the instrument Result 13.2 priced, and the tie later in this section needs it. Under continuous compounding at yield $y$, a zero paying one unit at $T$ years is worth $P(y) = e^{-yT}$. Differentiating,

$$\frac{dP}{dy} = -T\,e^{-yT} = -T\,P \quad \Longrightarrow \quad D = -\frac{1}{P}\frac{dP}{dy} = T,$$

so a $T$-year zero's modified duration is exactly $T$ — no approximation, no compounding-frequency adjustment, because there is only one cash flow and continuous compounding leaves no residual conversion factor. The second derivative gives the *convexity*,

$$C = \frac{1}{P}\frac{d^2P}{dy^2} = T^2,$$

and the second-order Taylor expansion of $\ln P(y + dy) - \ln P(y)$ around $dy = 0$ is

$$\frac{dP}{P} \approx -D\,dy + \tfrac{1}{2}C\,dy^2,$$

with the correction term non-negative by construction whenever $C > 0$: convexity always adds back, on either side of a yield move, never subtracts, for any option-free bond, zero-coupon or coupon-bearing alike — it is the curvature of a convex function, and $e^{-yT}$, or any positively weighted sum of such terms, is convex in $y$ by inspection.

**Worked Example 13.2 — Duration and convexity P&L on a rate shock**

A hypothetical bond position, not market data. A book holds \$2,000,000 notional of a bond with modified duration $D = 7.00$ and convexity $C = 75.0$. Yields rise $+50$bp. Find the price change with and without the convexity correction, in percentage and in dollars.

**Duration only.**

$$\frac{dP}{P} = -D\,dy = -7.00 \times 0.0050 = -0.035000 = -3.5000\%.$$

**With convexity.**

$$\tfrac{1}{2}C\,dy^2 = 0.5 \times 75.0 \times 0.0050^2 = 0.000938 = 0.0938\%.$$

$$\frac{dP}{P} \approx -3.5000\% + 0.0938\% = -3.4063\%.$$

**Dollars.** On \$2,000,000, duration-only gives −\$70,000.00; with the convexity correction, −\$68,125.00 — convexity gives back \$1,875.00 of the loss, precisely because the second-order term is positive regardless of which way yields moved, exactly as Result 13.3 showed it must be. (All figures verified in `verify_mk13.py`, §2.)

### Why this matters beyond the bond desk: the bond move *is* the differential move

Chapter 5's central worked illustration priced a policy surprise through to spot with one piece of arithmetic: a persistent shift $\Delta$ in the 2-year interest-rate differential moves the exchange rate, under the UIP-benchmark logic Chapter 5 derived, by $\Delta \times H$, where $H$ is the number of years the market expects the shift to persist. Result 13.3 just derived a formally identical piece of arithmetic for a bond: a $T$-year zero's price moves by $-T\,dy$ for a yield shift $dy$. Set $T = H$ and $dy = \Delta$ and the two expressions are the same multiplication, executed on two different instruments.

**Worked Example 13.3 — The bond move is the differential move**

Chapter 5's own worked example took an illustrative pass-through of a policy surprise to the 2-year differential of $\Delta = 8$bp, sustained over $H = 2$ years, and found a UIP-benchmark spot response of $\Delta \times H = 16.00\text{bp} = 0.1600\%$. Take a 2-year zero-coupon government bond, $T = 2$, and apply the identical 8bp yield shift.

**Duration-only.**

$$\frac{dP}{P} = -T\,dy = -2 \times 0.0008 = -0.001600 = -0.1600\% \;\; (-16.00\text{bp}).$$

**Exact, continuous compounding.** $P(y+dy)/P(y) - 1 = e^{-T\,dy} - 1 = -0.159872\%$ ($-15.9872$bp), against a convexity-term prediction of $\tfrac{1}{2}T^2dy^2 = 0.000128\%$ (0.0128bp) — the linear approximation is accurate to three decimal places of a basis point at this size of shock, because $dy$ is small and convexity is a genuinely second-order effect.

The duration-only bond move, −16.00bp, and Chapter 5's spot move, +16.00bp, agree exactly in magnitude, and the opposite sign is exactly what the economics predicts, not a discrepancy: a rate rise that lowers a bond's price is, via Chapter 5's UIP arithmetic, the same rate rise that strengthens the currency it is denominated in. Neither fact is a coincidence to marvel at; both are the same first-order arithmetic, $-T\,dy$ against $\Delta \times H$, applied to two instruments driven by the identical underlying number — the 2-year differential. The practical consequence is for risk management, not direction: a trader who holds a rates position sized off a 2-year duration and, separately, an FX position sized off a 2-year differential view, is not running two independently diversifying bets. Both P&Ls move off the same shock, in a fixed, computable ratio set by $T$ and $H$, and a portfolio or a risk model that prices them as uncorrelated is missing the shared factor that Result 13.3 and Chapter 5's own iterated UIP relation both derive from the same identity. (All figures verified in `verify_mk13.py`, §2, block S2b.)

The same mechanics reach further than bonds. Any asset priced as a stream of future cash flows discounted at a rate is duration-like in exactly Result 13.3's sense, and an equity is the extreme case: a claim on a growing, risky, and technically unbounded cash-flow stream, discounted at a rate that is itself a sum of a risk-free component and a risk premium. Under the simplest constant-growth pricing $P = \mathrm{CF}_1/(r-g)$ — next period's cash flow $\mathrm{CF}_1$, discount rate $r$, perpetual growth $g < r$ — differentiating with respect to $r$ gives

$$\frac{d\ln P}{dr} = -\frac{1}{r-g},$$

an "equity duration" of $1/(r-g)$ years, precisely analogous to Result 13.3's $D = T$ for a zero, with $r-g$ playing the role a zero's flat maturity plays for a bond. At $r = 7.00\%$ and $g = 4.00\%$ — illustrative, not exotic, figures for a mature-market growth index — that duration is $1/0.0300 = 33.33$ years, roughly $4.76\times$ the $D = 7.00$ bond of Worked Example 13.2, and the identical $+50$bp shock that cost that bond 3.5000% (duration-only) costs this stylised equity 16.6667% — a move $4.76\times$ larger for the same rate shock, entirely on account of how much further out its cash flows sit and how thin the cushion $r-g$ is between its growth rate and its discount rate. Section 13.5 takes this mechanism, adds the risk-premium channel sitting alongside the pure discount-rate one, and puts a measured empirical magnitude on it.

## 13.4 The macro calendar

Sections 13.1–13.3 describe a curve as a static object with a knowable shape; almost none of that shape is static in practice, because it moves in discrete jumps around a small number of scheduled events. This section is about the anatomy of those events, and about one property they share that almost nothing else this book has catalogued shares: the market knows, months in advance, exactly when they will happen.

**Definition — Consensus, actual, surprise**

Ahead of a scheduled data release, a poll of professional economists (or, for a policy decision, the market-implied pricing Section 13.1 already derived) produces a *consensus* — typically the median or mean forecast. The release itself is the *actual*. The *surprise* is

$$\text{surprise} = \text{actual} - \text{consensus},$$

and the logic Chapter 5 already established for policy surprises applies without modification here, for data releases generally: only the surprise component can move a price at the instant of release, because the consensus was already the market's best guess and was therefore already in the price. Comparing surprises across different series requires standardising them first, since a raw surprise of "+75" means nothing on its own for a series reported in thousands of jobs against one reported in percentage points of an index: the *standardised surprise* divides the raw surprise by an estimate of the series' own historical surprise dispersion, $\hat{\sigma}$,

$$z_{\text{surprise}} = \frac{\text{actual} - \text{consensus}}{\hat{\sigma}},$$

putting non-farm payrolls, CPI, and every other release on a comparable footing before any covariate built from several of them is combined.

Two structural features of the release calendar recur across every series a systematic book actually consumes.

**The releases themselves.** In the United States, the Employment Situation report — headlined by *non-farm payrolls* (NFP), the month-on-month change in payrolled employment — is released

on a fixed monthly schedule (conventionally the first Friday of the month, though the exact date shifts with the calendar and should be checked against the current release schedule rather than assumed) and is, by reputation and by the volume of market activity it moves, the single most-watched routine data release in this book's asset universe. The Consumer Price Index (CPI) follows monthly, and the Personal Consumption Expenditures price index (PCE) — the Federal Reserve's own stated preferred inflation gauge — follows again, later in the month, drawn from the same underlying spending data. The Federal Open Market Committee's eight scheduled decisions a year, already treated in Section 13.1, complete the core US calendar. In Australia, the Australian Bureau of Statistics publishes a quarterly Consumer Price Index — the long-standing, full-basket official measure — alongside a more frequently updated monthly CPI indicator introduced to narrow the gap with economies that already publish a monthly headline figure; the monthly indicator's basket coverage and its role relative to the quarterly release have themselves evolved since introduction and should be verified against the ABS's current methodology rather than assumed fixed. The monthly Labour Force Survey and the Reserve Bank of Australia's own eight-a-year decisions (Section 13.1) complete the domestic side. All of this is release-calendar *structure* — stable in form, though the exact dates and the fine detail of any one country's statistical methodology drift and must be checked against each agency's own published schedule rather than assumed fixed.

**Revisions.** Almost every headline release above is revised, sometimes materially, in subsequent months' reports — NFP's own initial print is well known for carrying meaningful revision risk two and three months after first publication. This is not a defect to route around; it is a permanent feature of how statistical agencies build these series from incomplete initial survey response, and it is precisely why [Course Book §18.2]'s point-in-time discipline exists: a backtest that uses the *current*, fully revised value of a historical NFP print at a date before that revision had actually happened is using information the market never had — an easy and common way a plausible-looking macro covariate becomes a leak. The US Federal Reserve Bank of St. Louis's ALFRED database — an archival companion to the more familiar FRED — exists precisely to address this, storing each data series *as it stood on every past date*, vintage by vintage, so a researcher can reconstruct exactly what a model built on that date could actually have known; this book's own point-in-time discipline is the same discipline ALFRED is built to enforce, applied to every macro series the platform ever ingests, not only the US ones with a dedicated archival source.

### The one thing the future genuinely publishes in advance

Nearly everything this book has treated as a covariate is either *static* (an instrument's identity, its asset class, its typical session profile) or *observed*, known only up to the present moment and no further: a rolling volatility estimate, a positioning composite, yesterday's close. The Course Book's Temporal Fusion Transformer draws a third category, *known-future* inputs [Course Book §8.4] — variables whose value for every step of the forecast horizon is already knowable today, not estimated, not forecast, simply read off a calendar. Day-of-week and session bucket, Chapter 4's own construction and already this library's standard known-future example [Course Book §8.4], qualify by construction: tomorrow is definitely a Tuesday. The economic release calendar earns the same classification for a subtler and more valuable reason. Almost no other genuinely informative fact about markets is published in advance — tomorrow's order flow, tomorrow's volatility shock, tomorrow's headline are none of them knowable today even in principle — but a central bank's meeting schedule and a statistical agency's release calendar are published, in writing, months or a year ahead, precisely so that markets can prepare for them. A covariate that flags "NFP releases in 18 hours" is not a forecast of anything; it is a fact, known with certainty today, about the shape of the distribution the model should be forecasting eighteen hours from now.

**Trap — Knowing *when* is not knowing *what***

The release date and time of next month's NFP print are known-future covariates in the fullest sense: certain, calendar-derived, available today for every step of the forecast horizon. The *value* that print will carry is the opposite of known-future — it is, definitionally, the least predictable number in the whole covariate set, because if it were predictable the consensus forecast would already equal it and there would be no surprise left to move anything (Definition 13.4). Do not let the first fact contaminate the second. A known-future release-timing dummy belongs in a model exactly as Chapter 4's day-of-week and session-bucket covariates do: as a flag that conditions the *width* of the forecast distribution, telling the model to widen its uncertainty into a release rather than pretending the next few hours look like any other few hours. It has no business being fed in as if it also disclosed the release's direction.

**Your project — Two different jobs for the same calendar**

The release calendar feeds the platform's TimesFM envelope two separate ways, and conflating them wastes half its value. First, as a known-future *covariate* [Course Book §8.4]: hours-to-next-release and a release-type dummy, computed exactly as Exercise 13.3 below constructs them, entering the model the same way a session bucket does, telling it to widen the forecast distribution into a scheduled event rather than treating the next few hours as routine. Second, and separately, as a *cost-model* input: OANDA's own quoted spreads widen predictably in the minutes around a major release, exactly as they widen at rollover and in thin Asian hours (Chapter 4) — and a news-window cost cell, keyed off the same release-type taxonomy this section built, belongs in the cost surface Chapter 4 introduced alongside its session buckets, priced from measured pre- and post-release spreads rather than an average-day figure. A backtest that lets a strategy trade through a scheduled release at its average, all-day spread makes exactly the execution-realism error [Course Book §18.3] warns against, in its most literal form: pretending the calendar it just used as a covariate had no effect on the market it is about to trade.

**Worked Example 13.4 — An NFP surprise, standardised, and priced through to spot**

A hypothetical scenario, not a historical print. Consensus for non-farm payrolls is 180k; the actual print is 255k. The series' own illustrative historical surprise dispersion is $\hat{\sigma} = 45$k. EURUSD trades at $S_0 = 1.0850$.

**Surprise, and standardised surprise.**

$$\text{surprise} = 255 - 180 = 75\text{k}, \qquad z_{\text{surprise}} = \frac{75}{45} = 1.6667.$$

A print 1.6667 historical standard deviations above consensus is a genuinely large upside surprise — strong enough, on its own, to move the policy-expectation path Section 13.1 priced, without any change to the FOMC's own calendar or communication.

**Pass-through to the differential, and the spot recap.** Take an illustrative pass-through of $\Delta(\text{2y EUR-USD differential}) = -3.0$bp — the differential narrows because the USD side of it firmed on the stronger labour-market reading — sustained over Chapter 5's $H = 2$-year benchmark horizon:

$$\text{spot change} = \Delta \times H = -3.0 \times 2.0 = -6.0\text{bp} = -0.0600\%,$$

$$\Delta S = 1.0850 \times (-0.000600) = -0.000651, \qquad S_1 = 1.084349 \ \ (-6.51 \text{ pips}).$$

The mechanics from here are entirely Chapter 5's — Section 13.1 and this section supply only the two pieces Chapter 5 takes as given: which surprise, standardised against which series'

own history, and which differential it moves. (All figures verified in `verify_mk13.py`, §3.)

## 13.5 Transmission to the four pillars

Sections 13.1–13.4 built the machinery: a corridor, a curve, a duration, a calendar. This section spends it, on the question this whole chapter exists to answer — when a rate surprise hits, what happens, with what sign, in each of this book's four pillars.

| Pillar | Sign | Mechanism | Developed in |
|---|---|---|---|
| FX | + for the hiking currency | UIP-benchmark differential pass-through: a widening rate advantage appreciates the currency, immediately, in proportion to the horizon the market expects it to persist | Chapter 5; Sections 13.1–13.4 above |
| Equities | − | Higher discount rate lowers the present value of any future cash-flow stream; higher required risk premium compounds it; effect concentrated in long-duration (growth) names | Section 13.3 above; Bernanke and Kuttner (2005) |
| Gold | −, conditional | Zero-yield asset: a higher *real* yield raises the opportunity cost of holding it; fires only if expected inflation does not rise in step with the nominal rate | Section 13.6 below; Chapter 12 |
| Carry / funding, everywhere | − for levered risk positions | Rising rates raise the funding cost of every levered book; if constraints bind, orderly cost becomes forced deleveraging | Chapter 14; Chapter 15 |

**Table 13.1.** A hawkish rate surprise, walked through the four pillars, sign by mechanism. "Sign" is the typical direction for the currency or market whose rate surprised hawkishly; every row is a mechanism, not a guarantee — Exercise 13.4 asks for the full walk-through on a concrete scenario.

**FX.** Nothing in this row is new: Sections 13.1 and 13.4 above, and Chapter 5 in full, already derived and repeatedly applied the mechanism — a widening rate advantage appreciates the currency in proportion to the horizon the market expects it to persist. What this chapter adds is the vocabulary connecting it to the rest of the curve: the "differential" Chapter 5 prices is a two-year point on exactly the curve Section 13.2 built, and Section 13.3's tie showed that moving it is, in a strict and computable sense, the same shock that moves a matched-duration bond.

**Equities.** Section 13.3 closed with the "equity duration" $1/(r-g)$: a higher discount rate $r$ mechanically lowers the present value of any positive future cash-flow stream, and does so more violently the further out those cash flows sit, which is why a rate shock hits an unprofitable, high-growth name harder than a mature dividend payer of identical market capitalisation. A second, distinct channel compounds the first: $r$ is not only the risk-free curve but the risk-free curve *plus* a required equity risk premium, and a hawkish surprise can raise the premium investors demand — tighter financial conditions read, correctly, as a signal that the probability of a future growth shock has risen — independently of, and on top of, the pure discounting arithmetic. Bernanke and Kuttner (2005) measured the combined effect directly, extracting the policy surprise from interest-rate futures exactly as Section 13.1 described, and found an unanticipated 25 basis-point cut in the federal funds rate associated with an increase of roughly one per cent in broad US equity indices — and, by the same logic run in reverse, a comparably sized unanticipated *hike* with a fall of similar magnitude. Exercise 13.4 uses that figure, scaled linearly to a stated surprise size, as a first approximation; treat the linear scaling as illustrative, since Bernanke and Kuttner's own regression is not obliged to pass through the origin or stay linear at every surprise size, only as a reasonable basis for ordering the pillars' magnitudes against one another.

**Gold.** Gold pays no coupon, no dividend, and no yield of any kind; holding it costs, at minimum, whatever a truly riskless asset would otherwise have paid, which is why the relevant rate for this row is specifically the *real* rate, not the nominal one Sections 13.1–13.4 have priced

throughout. Section 13.6 below develops the nominal/real distinction properly; the mechanism, stated now, is simply opportunity cost — a higher real yield on a safe, coupon-bearing asset raises the return an investor forgoes by holding a zero-yield one instead, so gold typically falls when real yields rise and rallies when they fall. Chapter 12 treats gold's own price formation and its typical correlation partners in full; this chapter's job is only the rate-transmission mechanism, and the honest caveat that goes with it.

**Trap — A nominal hike is not automatically a real-rate hike**

Every mechanism in this section so far is driven by a *nominal* rate surprise, and gold is the one row in Table 13.1 that does not respond to a nominal move mechanically at all — it responds to the *real* move, and the two are the same only if expected inflation is unchanged. A central bank that hikes purely to validate inflation the market had already priced in — delivering, say, 50bp against 50bp of already-embedded expected inflation — can leave the real rate, and therefore gold's opportunity cost, almost untouched, even though every nominal-rate covariate in the feed just moved sharply. Section 13.6's Fisher relation is the whole of the correction: check what happened to breakeven inflation, not only to the nominal print, before reading a rate surprise into a gold view.

**Carry and funding, everywhere.** Chapter 4's nightly rollover already showed the smallest instance of this channel: every CFD position, in every one of this book's four pillars, is financed overnight at a rate that moves when policy rates move, so a hiking cycle raises the running cost of holding *any* levered position, long or short, in *any* instrument — not only in the currency pair whose central bank just moved. Widen the lens from one book's financing line to the whole leveraged financial system and the same fact becomes the global financial cycle Rey (2015) describes: a move in one dominant central bank's policy rate — historically and structurally, most often the Federal Reserve's — propagates through the funding conditions facing every leveraged investor who borrows in or against that currency, *regardless of their own country's exchange-rate regime or nominally independent monetary policy*. Rey's "dilemma, not trilemma" framing is precisely that a floating exchange rate does not insulate a country from this channel the way the textbook trilemma promises it should; funding conditions transmit through balance sheets, not through the currency regime.

The mechanism that turns this into more than a financing-cost line item is constraints. As long as leverage and margin headroom are ample, a higher funding cost is exactly that — a cost, priced and absorbed. Once constraints start to bind — margin calls, risk limits, a funding line no longer offered at any price — the same rate rise forces *deleveraging*: positions are sold not because the trade thesis changed but because the cost or the capacity to hold them did. This is the market-structure justification for treating regimes as genuinely discrete states [Course Book §13.1] rather than as a continuously varying backdrop, and for building a router [Course Book §13.8] whose job is to detect which state currently obtains; Chapter 14 derives the mechanism in full, including the specific funding-currency arithmetic of a carry unwind, and Chapter 15 studies six historical episodes in which exactly this channel, not any single pillar's own fundamentals, did most of the damage.

### Balance-sheet policy: the second lever

Everything above treats monetary policy as one-dimensional — a short-rate corridor, moved at scheduled meetings. Since the 2008 financial crisis, and again from 2020, the major central banks have used a second, genuinely distinct lever: the size of their own balance sheet. *Quantitative easing* (QE) is large-scale purchase of assets — overwhelmingly government bonds, sometimes other securities — funded by creating central-bank reserves, expanding the balance sheet; *quantitative tightening* (QT) is the reverse, letting holdings run off at maturity or selling them outright, shrinking it. The channels through which balance-sheet size is thought to matter are structurally

distinct from the short-rate channel this chapter has otherwise treated: a *portfolio-balance* channel, in which removing duration and safe collateral from the market compresses term premia and pushes investors further out the risk spectrum than the short rate alone would explain; and a *signalling* channel, in which the scale and pace of purchases (or runoff) communicates a commitment to the policy stance for longer than the current short rate, on its own, would convey.

The consequence for a covariate feed is the honest point. Policy stance is genuinely two-dimensional — rate level, and balance-sheet size or flow — and a model, or a trader, reading only the short rate or Section 13.1's two-year differential is missing a source of variation that has, at times, dominated: a tightening cycle that combines rate hikes with active balance-sheet runoff is a materially different, and typically more restrictive, policy stance than the same rate path delivered with a static or expanding balance sheet, and the 2022-onward tightening cycle, which ran hikes and QT concurrently across several major central banks, is the obvious recent illustration (structural description; the specific pace and scale of any QT programme drift and should be verified against each central bank's current published schedule rather than assumed).

## 13.6 Inflation and real rates

Every rate this chapter has quoted so far is *nominal*: a return measured in units of currency, which is not the same thing as a return measured in units of what that currency buys. This section makes the distinction precise, because Section 13.5's gold row — and, more generally, any covariate meant to capture the "cost of money" rather than the mere level of a policy rate — needs the *real* rate specifically, not the nominal one.

### The price index, lightly

A consumer price index is built from a basket of goods and services, each assigned an expenditure weight meant to reflect its share of a representative household's spending, priced repeatedly against the same basket, and aggregated into a single index level; the year-on-year percentage change in that index level is the inflation rate as usually quoted. The weights are periodically re-based as spending patterns shift, and the basket's composition, its weighting methodology, and the frequency of re-basing are all matters of each statistical agency's own published methodology — structural machinery this book states and does not re-derive, and whose current specification should be checked against the relevant agency (the US Bureau of Labor Statistics, the Australian Bureau of Statistics) rather than assumed fixed.

**Result — The Fisher relation**

A nominal rate $i_{\text{nom}}$, a real rate $i_{\text{real}}$, and an expected inflation rate $\pi^e$ over the same period are connected by an argument with exactly the shape of Chapter 5's proof of covered interest parity: two routes to the same terminal purchasing power. *Route 1*: invest one unit of purchasing power directly at the real rate, $1 + i_{\text{real}}$ units of purchasing power at maturity. *Route 2*: invest one unit of currency at the nominal rate, $1 + i_{\text{nom}}$ units of currency at maturity, then deflate by expected inflation into purchasing-power terms, $(1 + i_{\text{nom}})/(1 + \pi^e)$. Both routes start from the same one unit of purchasing power and must agree:

$$\boxed{1 + i_{\text{real}} = \frac{1 + i_{\text{nom}}}{1 + \pi^e}} \qquad \Longrightarrow \qquad i_{\text{real}} \approx i_{\text{nom}} - \pi^e,$$

the familiar linear form dropping the (usually small) cross term, exactly as Chapter 5's linearised forward points dropped a denominator.

At $i_{\text{nom}} = 4.5000\%$ and $\pi^e = 2.5000\%$, the linear approximation gives $i_{\text{real}} \approx 2.0000\%$; the exact relation gives $(1.0450/1.0250) - 1 = 1.9512\%$, a gap of 4.88 basis points — the same order of

magnitude as the gap Chapter 5 found between its linearised and exact forward points, and for the identical reason: dropping a cross term that is small only when both rates are small.

### TIPS and breakevens: the market's own inflation forecast

Treasury Inflation-Protected Securities (TIPS) — US government bonds whose principal is indexed to CPI — trade directly on a *real* yield: their coupon and redemption are already inflation-adjusted, so the yield the market prices into a TIPS is, by construction, an estimate of $i_{\text{real}}$ without needing the Fisher relation to back it out. The *breakeven inflation rate* for a given maturity is the ordinary nominal Treasury yield at that maturity minus the TIPS yield at the same maturity — the market's implied compensation for expected inflation over the period, and, to a first approximation via Result 13.6, the market's own inflation forecast. At a nominal 10-year yield of 4.5000% and a 10-year TIPS yield of 2.1000%, the 10-year breakeven is 2.4000%. The same honest qualification Section 13.1 attached to the 2-year differential applies here without modification: a breakeven is contaminated by whatever liquidity premium and inflation-risk premium separate TIPS from nominal Treasuries, so treat it as a market *price* for inflation compensation, not a survey forecaster's clean expectation.

**Your project — The real rate as a v4 covariate**

The platform's own asset universe has, so far, treated "the differential" as Chapter 5 built it: a nominal two-year rate difference, entered as an FX covariate. Widening the platform to XAU/XAG under v4 exposes a gap that canon does not paper over: gold's best-founded transmission channel runs through the *real* yield specifically, not the nominal differential the FX pairs already consume, and no nominal covariate — however well it serves the currency pairs — substitutes for it. The fix is exactly the pattern [Course Book §22.4]'s master table already generalises across asset classes: add a real-yield (or breakeven-derived) series as its own point-in-time covariate [Course Book §18.2], typed and stored under the same discipline as every other macro input, rather than assuming a nominal-rate feature already captures gold's channel by proxy. It does not, and Section 13.5's trapbox is the reason why.

## 13.7 Exercises

**Exercise 13.1 — A humped curve: forward rates and shape, segment by segment**

A hypothetical government curve, not market data: $z_{0.25} = 4.55\%$, $z_1 = 4.20\%$, $z_2 = 4.00\%$, $z_{10} = 4.35\%$.

(a) Compute $f_{1,1}$, the 1-year rate one year forward. Where does it sit relative to $z_1$ and $z_2$, and why — run the same class-average argument Worked Example 13.1 used, in reverse.

(b) Compute the 2s10s spread, $z_{10} - z_2$, and classify it normal or inverted.

(c) Compute the three adjacent segment spreads — $z_1 - z_{0.25}$, $z_2 - z_1$, $z_{10} - z_2$ — classify each normal or inverted, and describe the curve's overall shape in one sentence.

▷ working space

**Solution — Exercise 13.1**

**(a)**

$$f_{1,1} = \frac{(1+z_2)^2}{1+z_1} - 1 = \frac{1.040000^2}{1.042000} - 1 = \frac{1.081600}{1.042000} - 1 = 0.038004 = 3.8004\%.$$

$f_{1,1} = 3.8004\%$ sits *below* both $z_2 = 4.0000\%$ and $z_1 = 4.2000\%$. The class-average argument runs in reverse of the normal-curve case: since $z_2 < z_1$, the two-year zero's average return is already *below* the one-year's, so the second year's implied rate must fall below the two-year average itself to pull that average down from $z_1$'s level — an inverted segment forces its forward rate below both spot rates, the exact mirror image of Worked Example 13.1's normal-segment result.

**(b)**

$$z_{10} - z_2 = 4.3500\% - 4.0000\% = 0.3500\% = 35.00\text{bp} \quad \text{(normal / upward)}.$$

**(c)**

| Segment | Spread | Classification |
|---|---|---|
| $z_1 - z_{0.25}$ | $-35.00$bp | Inverted |
| $z_2 - z_1$ | $-20.00$bp | Inverted |
| $z_{10} - z_2$ | $+35.00$bp | Normal |

The curve is inverted from three months out to two years, then normal (upward) from two years out to ten — a front-inverted, *humped* curve, exactly Table 13.2's fourth row: the near end priced for imminent easing or at least no further hiking, the long end priced for eventual normalisation above wherever the front end troughs. (All figures verified in `verify_mk13.py`, §5.)

**Exercise 13.2 — Duration P&L on a rally, and the same shock priced twice**

(a) A hypothetical \$5,000,000 position in a bond with $D = 4.50$ and $C = 32.0$. Yields *fall* 25bp. Find the duration-only and convexity-adjusted price change and dollar P&L, and say, without recomputing, whether convexity should add to or subtract from the duration-only gain — then confirm.

(b) A hypothetical GBPUSD position, spot $S_0 = 1.2650$. An illustrative pass-through takes $\Delta(\text{2y GBP-USD differential}) = +5.0$bp (GBP side firms), sustained over $H = 3$ years. Compute the UIP-benchmark spot change in pips. Separately, a 3-year zero-coupon bond ($D = T = 3$) is exposed to the identical 5.0bp shock: compute its duration-only price change and confirm Section 13.3's tie holds here too, magnitude for magnitude.

▷ working space

**Solution — Exercise 13.2**

**(a)** Convexity is positive for any option-free bond (Result 13.3), so it *adds* to whichever direction duration alone predicts — here, a rally, so it should add to the gain, not subtract from it.

$$\left.\frac{dP}{P}\right|_{\text{dur.}} = -D\,dy = -4.50 \times (-0.0025) = 0.011250 = 1.1250\%,$$

$$\tfrac{1}{2}C\,dy^2 = 0.5 \times 32.0 \times 0.0025^2 = 0.000100 = 0.0100\%.$$

$$\frac{dP}{P} \approx 1.1250\% + 0.0100\% = 1.1350\%.$$

On \$5,000,000: duration-only \$56,250.00; with convexity \$56,750.00 — an extra \$500.00 of gain, confirming the prediction.

**(b)** FX side:

$$\text{spot change} = \Delta \times H = 5.0 \times 3.0 = 15.0\text{bp} = 0.1500\%,$$

$$\Delta S = 1.2650 \times 0.001500 = 0.001898, \qquad S_1 = 1.266898 \;\; (+18.98 \text{ pips}).$$

Bond side, $D = T = 3$, $dy = +5.0$bp:

$$\frac{dP}{P} = -D\,dy = -3.0 \times 0.0005 = -0.001500 = -0.1500\%.$$

$|+0.1500\%| = |-0.1500\%|$ exactly: the same 5.0bp shock, applied over the same 3-year horizon on both sides, produces price moves of identical magnitude in the currency pair and in the matched-maturity bond — precisely Section 13.3's tie, now with the maturity and the shock size both changed and the equality still holding, because the equality was never about these particular numbers: it is about $T = H$ and $dy = \Delta$ being the same multiplication regardless of their value. (All figures verified in `verify_mk13.py`, §6.)

**Exercise 13.3 — The known-future covariate table, for one week's shape**

A given release schedule for the trading week of Monday 9 November 2026 to Friday 13 November 2026 (a week entirely after United States daylight saving ends on 1 November 2026, so New York is on Eastern Standard Time, UTC−5, throughout):

| Release | Native time | Native day |
|---|---|---|
| RBA decision | 14:30 AEST | Tuesday 10 November |
| FOMC decision | 14:00 ET | Wednesday 11 November |
| US CPI | 8:30 ET | Thursday 12 November |
| US NFP | 8:30 ET | Friday 13 November |

(a) Convert the three US releases to the reader's own AEST clock (ten hours ahead of UTC, no daylight saving), and list all four events in AEST chronological order. Flag any release whose AEST calendar date differs from its native US calendar date.

(b) Build the "hours to next scheduled release" covariate, as of 8:00 AEST on each of the five trading days.

(c) Suppose instead the FOMC decision above had been scheduled for Friday 13 November at 14:00 ET rather than Wednesday. Convert it to AEST, compare it against Chapter 4's Friday 22:00 UTC trading-week close, and say whether the release falls inside or outside that week's continuous trading session.

▷ working space

**Solution — Exercise 13.3**

**(a)** New York is on EST (UTC−5) all week, so converting to AEST (UTC+10) adds a flat 15 hours:

| Release | Native (ET) | AEST | Date shift |
|---|---|---|---|
| FOMC | Wed 14:00 | Thu 05:00 | Rolls forward one day |
| US CPI | Thu 8:30 | Thu 23:30 | Same AEST day |
| US NFP | Fri 8:30 | Fri 23:30 | Same AEST day |

The FOMC decision is the one to flag: read naively off a US calendar as "Wednesday", it lands on *Thursday* at 05:00 on the reader's own AEST clock — exactly the trap Chapter 4 warned against when it insisted a Wednesday-New-York event is not a Wednesday event on the reader's own calendar, now for a scheduled decision rather than a rollover. In AEST chronological order: RBA (Tue 14:30), FOMC (Thu 05:00), US CPI (Thu 23:30), US NFP (Fri 23:30).

**(b)**

| Day (ref. 8:00 AEST) | Next release | Hours to next release |
|---|---|---|
| Monday | RBA (Tue 14:30) | 30.5h |
| Tuesday | RBA (Tue 14:30) | 6.5h |
| Wednesday | FOMC (Thu 05:00) | 21.0h |
| Thursday | US CPI (Thu 23:30) | 15.5h |
| Friday | US NFP (Fri 23:30) | 15.5h |

Thursday is the row to check by hand: at 8:00 AEST Thursday, the FOMC decision (Thursday 05:00 AEST) has *already* occurred, so the next scheduled release is US CPI at 23:30 the same AEST day, not the FOMC — a covariate table built carelessly, by matching calendar dates rather than comparing full timestamps, would wrongly point Thursday's row at an FOMC decision that is already three hours in the past.

**(c)** A Friday 14:00 ET FOMC converts, by the identical +15-hour offset, to Saturday 05:00 AEST. Chapter 4 fixes the trading week's close at Friday 22:00 UTC, which on the reader's own AEST clock is Saturday 8:00. The hypothetical release, at Saturday 05:00 AEST, falls 3.0 hours *before* that close — inside the continuous trading week, not after it, however strange "Saturday" looks on a calendar the reader associates with the market being shut. A covariate keyed only to the reader's own local calendar date, rather than to the actual UTC trading-week boundary Chapter 4 defines, would misclassify this release as happening after the week had ended. (All figures verified in `verify_mk13.py`, §7.)

**Exercise 13.4 — The channel map, worked: a hawkish surprise hits**

Ahead of a scheduled FOMC meeting, meeting-dated pricing has $m = 10$bp of a possible $\delta = 25$bp hike priced in. The Fed delivers the full 25bp *and* revises its guidance hawkishly. Walk all four pillars of Table 13.1 through the surprise, in order, stating the sign and the mechanism for each, and compute figures where Sections 13.1–13.6 give you the tools to. AUDUSD trades at $S_0 = 0.6520$; take an illustrative net pass-through (target *and* path factors combined) of $\Delta(\text{2y AUD-USD differential}) = -22.0$bp, over Chapter 5's $H = 2$-year benchmark.

▷ working space

**Solution — Exercise 13.4**

**The surprise itself.** $p = m/\delta = 10/25 = 0.40$, so $\mathbb{E}[\text{move}] = 0.40 \times 25 = 10.0$bp; delivered 25.0bp; target-factor surprise $= 25.0 - 10.0 = 15.0$bp (Section 13.1's $p = m/\delta$ machinery, applied exactly as there). The hawkish guidance revision is, in Chapter 5's terms, an additional *path*-factor surprise on top of this target-factor one — not separately quantified here, but real, and part of why the illustrative differential pass-through below is larger than a bare target-factor surprise alone would produce.

**FX — sign + for the hiking currency, USD.** The 2-year AUD-USD differential ($i_{\text{AUD}} - i_{\text{USD}}$) narrows as the USD side firms:

$$\text{spot change} = \Delta \times H = -22.0 \times 2.0 = -44.0\text{bp} = -0.4400\%,$$

$$\Delta S = 0.6520 \times (-0.004400) = -0.002869, \qquad S_1 = 0.649131 \;\; (-28.69 \text{ pips}).$$

AUDUSD falls: the USD, the currency whose rate surprised hawkishly, strengthens against AUD, exactly Section 13.5's FX row.

**Equities — sign −.** Scaling Bernanke and Kuttner's (2005) approximate magnitude linearly to the 15.0bp *target-factor* surprise (their finding is specifically about the surprise in the rate itself, not the guidance):

$$\text{impact} \approx -\frac{15.0}{25} \times 1.00\% = -0.6000\%,$$

before accounting for the additional, unquantified path-factor drag the hawkish guidance adds on top — so $-0.6000\%$ is a floor on the likely equity move, not a ceiling.

**Gold — sign −, conditional.** Falls on two reinforcing channels *if* the real-rate condition holds: the opportunity-cost channel (Section 13.6), which fires only to the extent the nominal move is not offset by a matching rise in expected inflation — check breakevens, not just the nominal print, per Section 13.5's trapbox — and the stronger-USD channel (Section 13.5), since gold is dollar-quoted and a broadly stronger dollar mechanically lowers its price in every other currency simultaneously. Both channels point the same way here, which is precisely why gold's reaction to a clean hawkish surprise is usually the least ambiguous of the four rows — and precisely why Chapter 12 flags the cases (a liquidation-driven risk-off, chiefly) where the sign flips regardless.

**Carry and funding — sign − for levered risk positions, conditional on constraints binding.** USD funding costs rise for every book financed in or against dollars, a cost before it is anything else (Chapter 4's rollover arithmetic, at systemic scale). Whether it becomes more than a cost — whether it forces deleveraging rather than merely being paid — depends on whether the surprise is large enough, or arrives at a moment fragile enough, to bind margin and risk-limit constraints somewhere in the system [Course Book §13.1]; that is a regime question, not a mechanical certainty, and it is exactly the question Chapter 14's router [Course Book §13.8] exists to condition on rather than assume.

**The pattern.** Every pillar's sign follows from the same 15.0bp (target-factor) surprise; none of the four magnitudes equals another, none is derived from the others, and the FX and equity figures above are the only two this chapter can actually compute without borrowing machinery Chapters 12 and 14 own. That asymmetry — confident sign and a computable magnitude for two pillars, sign-with-caveats for the other two — is itself the honest answer to "what happens when a hawkish surprise hits": not a single number, but a signed, conditional map. (Surprise and FX figures verified in `verify_mk13.py`, §9; equity figure in the same block.)

**Key idea — What to remember**

A central bank sets one administered overnight rate; everything else in this chapter is that one number propagating outward — through a corridor to the money market, through no-arbitrage to a full curve of forward rates, through duration to every bond's price, and, on a calendar published months in advance, through scheduled surprises to the differential Chapter 5 already showed moves spot. The curve's shape is a fact about prices today; what it forecasts about tomorrow's policy is a hypothesis, tested like any other, not read off a spread and trusted. Duration and the FX differential turned out, exactly, to be the same arithmetic on two different instruments, and the same rate shock reaches equities through a discount-rate channel this chapter can size, gold through a real-rate channel that depends on what inflation expectations did at the same time, and every levered position in every one of this book's four pillars through a funding cost that, when constraints bind, stops being a cost and starts being a forced sale. That last channel is where this chapter ends and the next begins: Chapter 14 takes the funding-cost mechanism sketched in Section 13.5 and builds it into the full account of why cross-asset correlation is not a fixed number, and of what a regime actually is.

# 14

# Intermarket Dynamics and Regimes

*Everything in the previous thirteen chapters described one market at a time. This one describes the wiring between them, and immediately makes an uncomfortable admission: the wiring is real, it is economically explicable, and it is not stable. A cross-asset relationship is not a constant of nature like a contract specification; it is the current output of a mechanism — funding, invoicing, discounting, hedging, margining — and when the mechanism changes state, the sign of the relationship can change with it. That is not a defect in the data. It is the single most important structural fact a systematic trader needs, because it is the market-level reason regimes exist at all. The chapter therefore does three things in order: it maps the typical relationships and labels every one of them unstable; it derives, mechanically rather than narratively, why stress compresses those relationships towards a single common factor; and it closes by arguing that binding constraints — margin rules, leverage limits, risk budgets — create genuinely state-dependent market behaviour, which is the market-structure justification for this platform's entire regime apparatus* [Course Book §13.1]. *Nothing here is offered as a signal; every claimed lead and lag is routed to the deflation machinery* [Course Book §15.4] *before it may be believed.*

## 14.1 The correlation map

Ask a trading floor what drives the Australian dollar and you will be told, without hesitation, that it is a growth-sensitive currency: it rises when global equities rise, when commodity demand is strong, and when risk appetite is healthy. That answer is not wrong. It is also not a law, not a constant, and not something you may use without measuring it, and the distance between "not wrong" and "usable" is what this section is about.

Begin with what a cross-asset relationship actually is. It is a claim about the joint distribution of two return series — most often about the sign of their covariance — and every such claim is made *over some sample*. Change the sample and you change the claim. There is no such thing as "the correlation between gold and equities"; there is only the correlation *estimated over a particular window, at a particular sampling frequency, in a particular alignment*, and each of those three choices moves the answer. This is not pedantry. Two of the three choices are pure measurement artefacts that a careless analyst will mistake for economics.

**Frequency.** Correlations measured on daily returns are systematically lower than the same relationship measured on weekly or monthly returns, because at short horizons the idiosyncratic, microstructure-driven component of each series is proportionally larger and the common component proportionally smaller. Aggregating to a longer horizon averages the idiosyncratic noise away faster than it averages the common component away, so the measured correlation rises with the sampling interval even when nothing about the underlying economics has changed. A relationship quoted without its frequency is not quoted at all.

**Alignment.** This one matters acutely for a trader in Australia. The ASX closes hours before the New York session opens, so a "same-day" return on an Australian index and a "same-day" return on a US index are computed over largely disjoint intervals of wall-clock time: today's ASX session already contains yesterday's US news and contains none of today's. The measured

contemporaneous correlation between them is therefore mechanically depressed, and part of the true dependence is displaced into the *lagged* correlation between the US index today and the ASX index tomorrow. That lagged relationship is a clock artefact, not a forecast, and Section 14.6 returns to the general problem of mistaking one for the other. The honest remedies are to align on overlapping intervals, to work at a frequency long enough that the non-synchrony is a small fraction of the interval, or to model the lead explicitly and treat it as the calendar fact it is (Hasbrouck, 2007, develops the general machinery for attributing price discovery across venues that do not trade simultaneously).

**Window.** The third choice is the subject of the rest of this chapter, and it is not an artefact: relationships genuinely change.

### The map

With those caveats attached in advance, Table 14.1 states the relationships a practitioner is expected to know. Read the **Sign** column as "the sign this relationship usually shows" and the **Mechanism** column as the reason anyone believes it; read the **Stability** column as the operative content. Every row is typical and unstable. The instability is not a caveat appended to the table — it *is* the table's subject.

Two rows deserve emphasis because they carry the chapter's argument in miniature.

The **bond–equity** row is the most instructive relationship in modern markets precisely because its sign is not merely noisy but has changed *regime* and stayed changed for decades at a time. The mechanism is transparent once stated. An equity price is a claim on future cash flows discounted at a rate; a nominal bond price is a claim on fixed nominal cash flows discounted at a related rate. If the dominant shock is to *growth expectations*, equities fall while bonds rally (the discount rate falls and safety is bid), giving a negative correlation and making bonds a hedge. If the dominant shock is to *inflation*, both fall together, because the same higher discount rate hits both claims, and bonds are not a hedge at all — they are a second position in the same trade. Nothing about either asset changed. What changed was which kind of shock the macroeconomy was producing, and every multi-asset risk model calibrated on one of those epochs was wrong about the other. This is the clearest available demonstration that a correlation is a statement about the prevailing shock mix, not about the assets.

The **gold–equity** row demonstrates the same point on a shorter clock. Gold is simultaneously a zero-coupon real asset (which argues for a negative relationship with equities via the real-yield channel) and a liquid, widely held, marginable position (which argues for a positive relationship whenever holders are being forced to raise cash). Both mechanisms are always present; the question is only which one is binding this week. Worked Example 14.2 runs that switch explicitly.

### Rolling correlation, defined

If a relationship is a function of the prevailing state, the natural estimator is one that looks only at the recent past.

**Definition — Rolling (windowed) correlation**

Let $x_t$ and $y_t$ be two return series observed at $t = 1, \dots, n$, and fix a window length $W \geq 3$. The *rolling correlation at* $t$ uses only observations $t - W + 1, \dots, t$:

$$\hat{\rho}_t(W) \;=\; \frac{\sum_{s=t-W+1}^{t} (x_s - \bar{x}_t)(y_s - \bar{y}_t)}{\sqrt{\sum_{s=t-W+1}^{t} (x_s - \bar{x}_t)^2}\,\sqrt{\sum_{s=t-W+1}^{t} (y_s - \bar{y}_t)^2}}, \qquad \bar{x}_t = \frac{1}{W}\sum_{s=t-W+1}^{t} x_s,$$

and $\bar{y}_t$ analogously. The estimator is defined for $t \geq W$. Three properties matter operationally. It is *backward-only*: $\hat{\rho}_t$ uses no observation dated after $t$, which is the condition for it to be admissible as a covariate at $t$ [Course Book §18.2]. Its denominators cancel any common scaling of

| Pair | Sign | Mechanism | Stability |
|---|---|---|---|
| Broad USD vs growth-sensitive FX (AUD, NZD, EM) | − | USD is the funding and invoicing currency; its strength tightens global credit and reduces demand for higher-beta claims | Typical, unstable. Weakest when the shock originates in the US itself |
| Global equities vs JPY, CHF | − | Both are funding currencies; risk-off forces unwinds that buy them back (Section 14.3) | Typical, unstable. Depends on whether the pair is currently being used to fund |
| Global equities vs broad USD | − | Safe-asset and dollar-funding demand in stress | Typical, *least reliable of the havens*: inverts when the Federal Reserve is the easing party (mk07 flagged the same instability) |
| Gold vs real yields | − | Gold pays no coupon; a higher real yield raises the opportunity cost of holding it | The best-founded row in the table, and still unstable: it fails outright during forced liquidation (Section 14.3) |
| Gold vs broad USD | − | Gold is quoted in dollars; a stronger dollar mechanically lowers the dollar price of a globally demanded asset | Typical, unstable; overwhelmed whenever both are bid as havens simultaneously |
| Gold vs equities | ? | Two opposing channels: haven demand (negative) and liquidation demand for cash (positive) | *No reliable unconditional sign at all*; conditional on phase (Worked Example 14.2) |
| Crude oil vs petro-currencies (CAD-, NOK-style links) | + | Terms of trade: a higher oil price raises the export revenue and national income of a net exporter | Typical, unstable; breaks when the oil move is itself a global demand shock hitting both legs |
| Copper and industrial metals vs growth-sensitive FX | + | Shared exposure to global industrial demand | Typical, unstable; distorted by supply shocks and by one large consumer's inventory policy |
| Equity implied volatility vs equity level | − | Demand for downside protection rises as prices fall; leverage rises as equity value falls (Cboe, 2019, documents the index methodology) | The most reliable row in the table — and still a statistical regularity, not an identity |
| Equity implied volatility vs carry and high-beta FX | − | Carry returns load negatively on global FX volatility innovations (Menkhoff, Sarno, Schmeling and Schrimpf, 2012) | Documented in the cross-section; magnitude and timing unstable |
| Government bonds vs equities | ? | Discount-rate channel (positive co-movement) versus flight-to-quality (negative co-movement); which dominates depends on whether the dominant shock is to inflation or to growth | *The sign has itself switched across decades*; treat as a regime variable, never as a constant |
| Crypto (BTC, ETH) vs equities | +? | Traded largely as a high-beta risk asset by leveraged, retail-heavy participants | Drifting: the beta has moved substantially over the asset's short history, with episodes of outright decoupling (indicative; re-estimate) |

**Table 14.1.** Typical cross-asset relationships. Every row is a historical regularity with an economic mechanism, not a constant; every sign has inverted in some documented episode. Sampling frequency and alignment as discussed in the text. Indicative, mid-2020s vintage; re-estimate on your own sample before use.

$x$ or $y$, so $\hat{\rho}_t \in [-1, 1]$ regardless of units. And it is a *step function of the window*: an observation entering or leaving the window at $t$ can move $\hat{\rho}_t$ discontinuously, which is why a short window produces a jagged path and a long window a smooth but stale one. There is no correct $W$; there is only an explicit trade-off between responsiveness and estimation error, and $W$ must be chosen before looking at the outcome, never tuned to make a story work.

The estimation-error half of that trade-off is quantifiable, and it is larger than intuition suggests. Under bivariate normality with true correlation $\rho$, Fisher's transform $z = \operatorname{artanh} \hat{\rho}$ is approximately normal with standard error $1/\sqrt{W-3}$. At $W = 6$ that standard error is $1/\sqrt{3} = 0.5774$ on the $z$ scale — enormous. A six-observation window can produce a sample correlation near $\pm 1$ from data whose true correlation is modest, purely by chance. Any claim that a relationship "flipped" must therefore clear a bar: is the change larger than the noise in the estimator? The worked example below both exhibits a flip and tests it.

**Worked Example 14.1 — A rolling correlation that changes sign**

Hypothetical scenario, constructed to be computed exactly; not market data. Twelve consecutive weekly percentage returns for a global equity index and for gold, spanning a calm stretch followed by a stress stretch:

| Week | Equity % | Gold % | Week | Equity % | Gold % |
|---|---|---|---|---|---|
| 1 | +1.20 | −0.80 | 7 | −2.60 | −1.50 |
| 2 | −0.90 | +1.20 | 8 | −3.40 | −2.60 |
| 3 | +0.60 | +0.30 | 9 | +1.90 | +0.50 |
| 4 | +1.70 | −1.10 | 10 | −4.50 | −2.90 |
| 5 | −1.30 | +0.90 | 11 | −1.20 | +0.40 |
| 6 | +0.40 | −0.50 | 12 | +2.80 | +1.60 |

**Two non-overlapping windows.** Applying Definition 14.1 with $W = 6$ to weeks 1–6 and then to weeks 7–12 (`verify_mk14.py` §1):

$$\hat{\rho}_{1:6} = -0.9235, \qquad \hat{\rho}_{7:12} = +0.9392.$$

The first window says gold is a hedge: when equities rise, gold falls, and vice versa. The second says gold is a risk asset: it falls with equities and rallies with them. The mechanism changed — Section 14.3 identifies which one — and the estimator changed with it.

**The pooled estimate describes neither state.** Over all twelve weeks,

$$\hat{\rho}_{1:12} = +0.6543.$$

This number is not a compromise between the two regimes in any useful sense. It is not the correlation in the calm state ($-0.9235$), it is not the correlation in the stress state ($+0.9392$), and it does not have the right sign for half the sample. A risk model fed $+0.6543$ would have overstated risk in the calm window — it thinks the two legs reinforce each other when in fact they offset — and understated it in the stress window. The pooled estimate is not merely imprecise; it is wrong in both states in opposite directions. Note also that the pooled figure sits closer to the stress window's value than to the calm window's, because the stress window's returns are larger and the Pearson estimator is dominated by large observations: pooling across regimes silently weights by volatility.

**The full rolling path.** Evaluating $\hat{\rho}_t(6)$ at every admissible $t$ shows how the transition is actually seen by a backward-only estimator:

| $t$ (week) | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|
| Window | 1–6 | 2–7 | 3–8 | 4–9 | 5–10 | 6–11 | 7–12 |
| $\hat{\rho}_t(6)$ | $-0.9235$ | $-0.0575$ | $+0.4940$ | $+0.5579$ | $+0.8244$ | $+0.8866$ | $+0.9392$ |

The sign changes between $t = 7$ and $t = 8$ — one and two weeks *after* the regime actually changed at week 7, which is the unavoidable lag of any windowed estimator. At $t = 7$ the estimator reads $-0.0575$: essentially zero, and therefore maximally uninformative, exactly at the moment the information would have been most valuable. A rolling correlation is a lagging description of the recent past, never a forecast, and its least reliable readings are the ones taken during the transition it exists to detect.

**Is the flip real, or is it noise?** Fisher-transform the two window estimates: $z_{1:6} = -1.6125$ and $z_{7:12} = +1.7310$, each with standard error $1/\sqrt{6-3} = 0.5774$. The individual 95% intervals are wide — $[-0.9918, -0.4470]$ and $[+0.5366, +0.9935]$ on the correlation scale — confirming that neither point estimate deserves its apparent precision. But the *difference* is

$$z_{7:12} - z_{1:6} = 3.3435, \qquad \text{s.e.} = \sqrt{2} \times 0.5774 = 0.8165, \qquad \frac{3.3435}{0.8165} = 4.0949,$$

i.e. 4.09 standard errors, two-sided $p = 0.000042$. The flip survives its own noise: these two windows are not two draws from one correlation. That test is the minimum standard for asserting that a relationship has changed, and it is a standard most published claims of "correlation breakdown" never attempt.

**Trap — A long-sample correlation matrix is the wrong matrix in every regime**

Estimating a covariance matrix on ten years of data does not average away the regime problem; it manufactures a matrix that describes no state the market was ever in. In the worked example the pooled $+0.6543$ has the wrong sign for half the sample. On a real book the error is worse than a sign: risk is understated in the state where it matters, because the calm periods that dominate the sample by count are the ones contributing the low correlations, while the stress periods that dominate the losses are a small fraction of the observations. Lengthening the sample to reduce estimation error therefore *increases* model error — the two are traded against each other, not jointly improved. The only defensible responses are to estimate conditionally (a rolling or state-dependent matrix), to stress the matrix deliberately (Section 14.4), or both. "More data" is not one of them.

**Your project — Correlation as a point-in-time covariate**

If a rolling cross-asset correlation is to enter the platform as a covariate — and Section 14.7 argues it should — it must be built to the same standard as every other input. Three requirements follow directly from Definition 14.1. First, backward-only: $\hat{\rho}_t$ at bar $t$ may use returns dated $t$ and earlier and nothing else, which rules out the centred windows most charting packages default to [Course Book §18.2]. Second, alignment: for the two index CFDs and the FX pairs the sessions do not coincide, so the correlation must be computed on a stated alignment convention that is frozen in code, not chosen per experiment. Third, the window length $W$ is a hyperparameter like any other, and searching over $W$ to find a value that improves a backtest is a selection decision that must be declared to the deflation machinery [Course Book §15.4]; choosing $W$ from the platform's existing timeframe ladder for a stated reason, before evaluation, is the cheaper and more honest route. A correlation covariate computed as a single fixed number over the whole history is not a covariate at all — it is a

constant, and by the trap above it is the wrong constant.

## 14.2 The dollar at the centre

Table 14.1 has the US dollar on one side of five rows, which is not an accident of selection. The dollar is not one currency among many that happens to be large; it occupies three distinct structural positions in the world financial system simultaneously, and each of them generates a channel through which a dollar move propagates into assets that have no dollar leg at all. These three roles are institutional facts, stable on the timescale of decades, and they should be stated as such — in sharp contrast to everything else in this chapter.

**Invoicing.** A large share of world trade is invoiced in dollars even when neither the exporter nor the importer is American. The consequence is that the dollar price of a traded good is the sticky price, and the local-currency price is the one that moves when the exchange rate moves. An exporter in a country whose currency has depreciated against the dollar therefore does not immediately gain the competitiveness a textbook pass-through would predict, because its invoice is denominated in a currency that did not move; what changes instead is the importing country's local-currency cost of the same goods. Dollar strength is thus, mechanically, a tightening of the terms on which most of the world buys imports, independent of any monetary policy decision anywhere.

**Funding.** A very large stock of credit outside the United States — borrowed by non-financial firms, banks and sovereigns whose revenues are in other currencies — is denominated in dollars. This is the most important of the three roles for our purposes. A borrower with dollar liabilities and local-currency assets is *short the dollar* in economic substance, whatever its stated hedging policy, and when the dollar appreciates that borrower's leverage rises without it having transacted. Its lender's measured risk rises at the same moment, so credit is withdrawn at exactly the point at which the borrower needs more of it. Dollar appreciation is therefore not a neutral relative-price change: it is a tightening of global financial conditions transmitted through balance sheets. This mechanism, not sentiment, is the reason the first row of Table 14.1 has the sign it does.

**Reserve and safe asset.** Official reserves are held predominantly in dollar assets, and the US Treasury market is the deepest pool of nominally risk-free collateral in existence. Because collateral quality is what determines who can borrow in a stress, demand for the dollar in a crisis is partly demand for the *collateral*, not for the currency as an investment. This is why the dollar can rally in a crisis whose origin is American: the demand is for the safe asset and for the funding currency, and both are supplied by the country in trouble.

### The dollar smile

Practitioners summarise the resulting behaviour in a framework rather than a model. The *dollar smile* is the observation that the dollar tends to be strong at both ends of the global growth distribution and weak in the middle, so that plotting dollar strength against global growth traces a $\smile$ shape. It is a mnemonic for three mechanisms operating in different states, and it should be presented as exactly that — a practitioner framework with an economic story for each leg, not a law with a testable functional form.

*Left leg: global crisis or sharp global slowdown.* The reserve and funding roles dominate. Demand for dollar collateral rises, dollar-denominated borrowers scramble to secure the currency their liabilities are written in, and risk positions funded in dollars are unwound, all of which buy dollars. The dollar strengthens even though — and sometimes because — the news is bad.

*Middle: synchronised global growth.* No one is scrambling for collateral, and capital flows towards the higher expected returns available outside the United States, which requires selling dollars to buy those claims. The dollar weakens. This is the state in which the growth-sensitive currencies in Table 14.1 do best, and it is the state in which a carry book is comfortable.

*Right leg: US outperformance.* If the United States grows materially faster than the rest of the world, its policy rate and its expected returns both rise relative to elsewhere, and capital flows towards it. The dollar strengthens — for reasons diametrically opposed to the left leg's.

The framework's usefulness is precisely that it forces the question "which leg are we on?" rather than "is the dollar a risk asset or a haven?", a question with no answer. Its limitation is equally important: the two ends of the smile are produced by different mechanisms and are therefore associated with different behaviour in *other* assets. On the left leg the dollar rallies while equities fall and havens are bid; on the right leg the dollar rallies while equities may be rising and US yields are climbing. Any covariate that encodes "the dollar is strong" without encoding which mechanism is operating has thrown away the information that distinguishes two states with opposite implications for every other position in the book. Chapter 7's warning that the dollar is the least reliable leg of a risk-off basket is the same statement, arrived at from the other direction.

> **Key idea — The dollar's sign is not a fact about the dollar**
>
> The dollar's three structural roles are stable; its correlation with any risk asset is not, because that correlation depends on which of the three roles is currently doing the work. A model that carries a broad-dollar covariate and no covariate identifying *why* the dollar is moving is being asked to learn one coefficient for two states in which the true coefficient has opposite signs — and, exactly as in Chapter 5's carry-versus-volatility case, it will learn a compromise that is correct in neither. This is the first of several places in this chapter where the honest conclusion is not "use a better estimator" but "condition on a state".

### The global financial cycle

The funding role has a macro-level counterpart that gives the dollar's centrality its sharpest formulation. Rey (2015) documents that gross capital flows, credit growth, leverage and risky-asset prices across a wide set of countries co-move with a single global factor, and that this factor is closely associated with US monetary policy and with measures of global risk appetite. The implication she draws is uncomfortable for the standard open-economy trilemma: a country with a freely floating exchange rate does not thereby obtain monetary independence, because the global financial cycle transmits through credit conditions regardless of the exchange-rate regime. The choice, in her formulation, is a *dilemma* — independent monetary policy requires managing the capital account, not merely floating the currency.

For a systematic trader the operational content is direct and does not depend on taking a side in the macro debate. It is that a common global factor exists, that it is substantially driven by conditions determined outside any single currency pair, and that it moves many of the platform's instruments at once. Two consequences follow. First, apparent diversification across a 21-instrument book is much less than the instrument count suggests, because a large share of the variation in the risk-sensitive instruments is one factor wearing twelve costumes — Section 14.4 makes this quantitative. Second, a shock that originates nowhere near AUD/USD — a change in the expected path of US policy, a widening of dollar funding spreads — is a legitimate driver of AUD/USD, and a model with no covariate carrying that information is missing a first-order input, not a refinement. Rey's factor is the reason Chapter 7's risk-on/risk-off composite is a coherent object rather than a bundle of correlated headlines; the next section supplies the mechanism by which it acts on prices.

## 14.3 Risk-on/risk-off, mechanically

Chapter 7 introduced risk-on/risk-off as a co-movement pattern and built a gauge from it. That chapter deliberately stopped at description. This section supplies what description omits: the balance-sheet mechanism that makes the pattern happen, derived step by step, because a trader

who knows only the pattern cannot tell when it will fail, and a trader who knows the mechanism can.

### The funding-currency mechanism

**Definition — Funding currency**

A *funding currency* is one that leveraged investors borrow, in aggregate and persistently, in order to finance positions denominated in other currencies. The qualifying condition is a low nominal interest rate relative to the alternatives, sustained long enough that borrowing in it becomes a standing practice rather than an opportunistic trade; deep and liquid short-term funding markets are a practical prerequisite. The Japanese yen and the Swiss franc have occupied this role for extended periods, and the US dollar does so as well by virtue of the funding role in Section 14.2. The status is not a property of the country: it is a property of the current interest-rate configuration and of market practice, and it can be lost when rates normalise. Whether a given currency is currently a funding currency is an empirical question about the present, to be re-answered rather than assumed.

The central and initially counter-intuitive fact about funding currencies is that they *appreciate* when markets fall. A currency belonging to an economy with no particular good news, often with weak growth and near-zero rates, rallies hard in precisely the weeks when global risk assets are being sold. Explaining this by "safe-haven demand" is not wrong but is not an explanation; it renames the observation. The mechanism is a forced flow, and it can be derived from the position's construction.

**Result — Why funding currencies rally in a risk-off event**

Let $L$ denote a low-yielding funding currency and let the "risk position" be any leveraged exposure financed by borrowing in $L$ — a carry trade in a higher-yielding currency $H$, an equity or credit position, or a CFD book whose financing leg is denominated in $L$. The chain runs:

**(1) Borrowing in $L$ is selling $L$.** To fund an exposure in another currency the investor borrows $L$ and converts. The conversion is a spot sale of $L$ in the market. The investor's balance sheet now holds an asset in another currency and a liability in $L$: it is *short* $L$, economically, regardless of the instrument used to express it.

**(2) The aggregate short accumulates slowly.** Because the trade is voluntary and its attraction is a slow-moving interest differential, the aggregate short in $L$ builds over months and quarters, in small increments, spread across many independent investors. No single participant observes the aggregate.

**(3) A shock raises measured risk.** Realised and implied volatility rise, for whatever reason and from whatever origin. Nothing about $L$'s economy need have changed.

**(4) Required equity per unit of exposure rises.** This is the crucial step, and it has three independent sources that all fire at once: brokers and clearing houses raise margin and haircut requirements when volatility rises; internal risk limits expressed as value-at-risk bind harder when the estimated variance rises [Course Book §14.5]; and volatility-targeting mandates mechanically reduce position size when the volatility estimate rises. All three are *rules*, not opinions, and all three demand the same action.

**(5) Reducing the exposure requires buying $L$.** To cut the position the investor must repay the $L$ liability, which means buying $L$ spot (or letting a hedge roll off, which the counterparty offsets by buying $L$). *Deleveraging a short is a purchase.* There is no way to reduce the position that does not generate demand for the funding currency.

**(6) The purchases are simultaneous and same-signed.** The investors do not coordinate; they do not need to. They share broadly similar risk models, face similar margin schedules from a small number of prime brokers, and observe the same volatility. Step (4) therefore fires for everyone at approximately the same time, and step (5) converts that into one-directional demand for $L$.

**(7) $L$ appreciates, which is a further loss for everyone still in the trade.** The appreciation marks every remaining short in $L$ to a worse price. That mark-to-market loss reduces equity, which returns the investor to step (4) with less capital and a larger measured risk. The loop closes.

Two structural features follow immediately and are the whole reason this matters. *Asymmetry*: entry is voluntary, incremental and spread over quarters; exit is compelled, simultaneous and compressed into days, because margin calls have deadlines and risk limits are breached discretely. The return distribution of the trade therefore has a long, thin right side and a short, fat left one — negative skew is a mechanical consequence of the position's funding structure, not a stylised fact requiring separate explanation (Brunnermeier, Nagel and Pedersen, 2008). *Endogeneity*: the size of the unwind is a function of how crowded the trade had become, which is a function of how long and how quietly it had worked. The crash's magnitude is generated by the trade's own prior success, not by the size of the triggering news.

The mechanism also tells you when the pattern will *fail*, which is what a mere description cannot do. If a currency has ceased to be a funding currency — because its rates rose and the aggregate short was already closed — then step (5) has nothing to unwind, and the currency will not rally in the next risk-off event. The relationship in Table 14.1's second row is conditional on a positioning state that itself changes. This is why Chapter 7's positioning data and this chapter's correlation map are not two separate topics: the positioning *is* the correlation's precondition.

A small piece of arithmetic makes step (5)'s flow concrete without repeating Chapter 5's derivation. In that chapter's hypothetical yen-funded position, the investor borrowed 9,750,000 JPY to fund 100,000 AUD at a spot rate of 97.50. After a year at a 0.50% JPY rate, closing the position requires repaying 9,798,750 JPY: the investor must buy back 48,750 JPY more than it originally sold, a 0.5000% excess (`verify_mk14.py` §4). The accrued funding interest is itself an additional short position in the funding currency, so the demand generated by an unwind strictly exceeds the supply generated when the trade was put on — before considering that the unwind happens all at once and the accumulation did not.

### Haven demand, and why the havens are not alike

Funding-unwind flow is one source of risk-off demand. It is not the only one, and the four assets routinely called "havens" reach that status by different routes, which is why they behave differently under stress.

The **yen and Swiss franc** are bought primarily through the mechanism above, reinforced by the fact that both belong to large net creditor positions: in a stress, domestic investors repatriate foreign assets, which is again a purchase of the home currency. The **US dollar** is bought through the collateral and funding channels of Section 14.2; its haven bid is a demand for the world's settlement asset, which is why it can strengthen in a crisis of US origin. **Gold** is different in kind from all three: it is bought because it is nobody's liability. Its haven property is the absence of counterparty risk, which is the one form of risk that rises fastest in a financial crisis. But the same feature that makes it attractive — it pays no coupon — makes it costly to hold when real yields are high, and, crucially, it is a liquid, marginable asset that leveraged investors own.

That last clause is the whole of gold's conditionality, and it produces a two-phase pattern in crashes that is repeated often enough to be worth working through explicitly.

**Worked Example 14.2 — Gold's two phases in a crash: the liquidation week and the haven week**

Hypothetical scenario, computed exactly; not market data. Ten consecutive daily percentage returns for a global equity index and for gold, spanning a two-week crash. Sessions 1–5 are the initial liquidation; sessions 6–10 follow, with equities still under pressure.

| Session | Equity % | Gold % | Session | Equity % | Gold % |
|---|---|---|---|---|---|
| 1 | −2.10 | −0.90 | 6 | −1.50 | +1.80 |
| 2 | −3.80 | −1.70 | 7 | +0.60 | +0.90 |
| 3 | −1.40 | −0.40 | 8 | −2.20 | +2.40 |
| 4 | −5.20 | −2.80 | 9 | +0.40 | +1.10 |
| 5 | −2.60 | −1.10 | 10 | −0.90 | +2.00 |

**Compounded moves and within-phase correlations** (`verify_mk14.py` §2):

| | Equity | Gold | Correlation |
|---|---|---|---|
| Phase 1 (sessions 1–5) | −14.2563% | −6.7285% | +0.9941 |
| Phase 2 (sessions 6–10) | −3.5770% | +8.4652% | −0.9568 |
| Whole episode | −17.3234% | +1.1671% | +0.7152 |

In phase 1 gold falls with equities, and the daily correlation is +0.9941 — gold behaves as a risk asset. In phase 2 gold rallies while equities continue to grind lower, and the correlation is −0.9568 — gold behaves as a haven. Over the whole episode gold is up 1.1671% against an equity index down 17.3234%, and the pooled correlation of +0.7152 describes neither phase, for exactly the reason Worked Example 14.1 gave.

**What it costs a book that treated gold as a hedge.** Consider a fixed-unit CFD book held without rebalancing: long 200,000 of index exposure and long 100,000 of gold exposure, struck at session 0's prices. For fixed-unit positions the P&L over any interval is the original notional times the price change measured against the strike, so the phase figures add exactly to the episode total. Measured against the session-0 strike, phase 2's moves are −3.0671% for the index and +7.8956% for gold, giving:

| | Index leg | Gold leg | Book |
|---|---|---|---|
| Phase 1 | −28,512.70 | −6,728.48 | −35,241.17 |
| Phase 2 | −6,134.18 | +7,895.58 | +1,761.40 |
| Episode | −34,646.88 | +1,167.10 | −33,479.78 |

The gold leg lost money in phase 1 and made money in phase 2, ending the episode up 1,167.10. Read carelessly, that says the hedge worked. Read carefully, it says something much more uncomfortable. Compare against an unhedged book of the same total size, 300,000 all in the index: over phase 1 that book loses 42,769.05 against the hedged book's 35,241.17, so the gold allocation *did* help — by 7,527.87 — but only because gold fell by less than equities, not because it rose. **In the phase that mattered, gold was a diversifier in degree and not a hedge in sign.**

**Why that distinction is not academic.** The book's worst mark-to-market is −35,241.17, at the end of phase 1 — 1.0526 times the episode's final loss of 33,479.78. A margin engine does not observe the episode; it observes the path. If the account's equity buffer is smaller than the phase-1 trough, the position is closed there, and the holder never receives phase 2's +1,761.40 at all. The haven property arrived, but it arrived after the margin call. This is the general shape of the problem: the correlation that a hedge relies on breaks in the first phase of a crisis and repairs in the second, and it is the first phase that determines whether you are still solvent to enjoy the second.

The mechanism behind the two phases is now easy to name. In phase 1 the binding constraint is cash: leveraged holders face margin calls across their whole book and sell what they *can* sell, which means the liquid, profitable, marginable positions — and gold is all three. The seller's balance sheet, not gold's fundamentals, sets the price. Once the forced sales are exhausted, the constraint stops binding and gold's ordinary determinants — real yields, counterparty fear, the dollar — reassert themselves, which is phase 2. BIS (2020) documents the same pattern at market scale in March 2020, when even conventional safe assets were sold for several sessions as participants of every type raised dollar cash indiscriminately; Brunnermeier (2009) describes the same dynamic across the 2007–2008 episode. Neither reference is needed to see the logic, but both establish that the pattern is a matter of record rather than of theory.

#### The equity–volatility–FX loop

One more circuit deserves explicit statement because it closes the link between three of this book's four pillars. It runs: equities fall → implied volatility rises, both because demand for downside protection increases and because falling equity value raises the leverage of the underlying firms (Cboe, 2019, documents the index construction that makes this measurable) → every value-at-risk and volatility-targeting rule in the system reduces permitted position size, since the variance input has risen → risk positions are cut, which includes buying back funding currencies by Result 14.3 and selling growth currencies → the FX moves mark further losses on multi-asset books → equities fall further. Option dealers' hedging can reinforce the circuit: a dealer short downside protection is short gamma, and hedging that exposure requires selling the underlying as it falls and buying it as it rises, which amplifies moves in both directions.

Every arrow in that loop is a rule or a hedging identity, not a sentiment. That is the point. The loop is not a description of how traders feel; it is a description of what mechanically must happen when a volatility estimate rises inside a system whose participants are constrained by that estimate.

#### RORO, in one paragraph

Chapter 7 built the platform's risk-on/risk-off gauge: orient a basket of cross-asset barometer series so a rise means more risk-off, standardise each causally against its own backward-only window, and combine into one scalar reading per bar. This chapter has now supplied what that chapter's construction assumed — the reason the basket's legs move together is not that they share a narrative but that they share a funding condition, and Rey's (2015) global factor plus Result 14.3's flow chain is the mechanism. The gauge is a *measurement* of the state of that mechanism, which is exactly why it is routed to the Stage-2 regime router [Course Book §13.8] as a conditioning variable and never used as a trading signal in its own right. A measurement of which state you are in is not a forecast of what happens next, and Chapter 7's warning against collapsing the one into the other stands unchanged.

## 14.4 Correlation in stress

"In a crisis, all correlations go to one" is the best-known sentence in risk management and one of the least well explained. Taken literally it is false — correlations do not reach one, and some, as Worked Example 14.2 showed, invert rather than converge. Taken as shorthand for "the common component of returns grows until it dominates the idiosyncratic component" it is both true and derivable, and there are exactly three mechanisms that produce it. They are not alternative explanations; they operate simultaneously and reinforce one another.

**Mechanism 1: a common discount factor.** Every risky asset is a claim on uncertain future cash flows, and its price is those cash flows discounted at a rate incorporating a risk premium. Decompose any asset's return into a component driven by news about *its own* cash flows and a component driven by news about the *common* discount rate. In ordinary times cash-flow news is

mostly idiosyncratic — a mining company's ore grade, a country's harvest — and dominates the variance of individual assets, so measured correlations are moderate. A shock to the common discount factor, whether to the risk-free rate or to the price of risk, moves every claim in the same direction at once, regardless of the individual stories. The correlation does not rise because the assets became more alike; it rises because the shared term's variance exploded while the idiosyncratic terms' variance did not. Correlation, being a ratio, records the change in the mix.

**Mechanism 2: the seller's balance sheet becomes the common factor.** Leveraged intermediaries hold portfolios spanning many assets and asset classes. When such an intermediary must raise cash, it sells across its holdings. The assets it sells then move together for a reason that has nothing to do with any of them: their common owner was constrained. Under this mechanism the correlation matrix ceases to describe economic relationships between assets and starts to describe overlaps in the ownership structure. Lowenstein (2000)'s account of the 1998 LTCM episode is the canonical illustration — positions with no economic connection moved together because a small number of large, similarly positioned funds were unwinding the same book at the same time, and other participants, correctly anticipating the unwind, traded ahead of it.

**Mechanism 3: constraints fire simultaneously.** Margin requirements, haircuts, value-at-risk limits and volatility targets are all functions of measured volatility, and measured volatility is the one thing that certainly rises in a crisis. Every constrained participant therefore receives the same instruction, in the same direction, at the same time. This is the mechanism that converts a shock into a spiral, and Brunnermeier and Pedersen (2009) give its two canonical forms precise statements.

**Definition — The loss spiral**

A *loss spiral* operates through the net worth of levered holders at a *fixed* leverage ratio. Let a holder own assets $A$ financed by debt $D$ with equity $E = A - D$, so leverage is $A/E$. A price fall reduces $A$ and reduces $E$ one-for-one, so leverage *rises* mechanically without any transaction. To restore its target leverage the holder must sell assets — and because equity fell by the full amount of the loss while assets fell by the same absolute amount, restoring the ratio requires selling a *multiple* of the loss. Those sales push the price down further, reducing $E$ again, requiring further sales. The spiral's amplification factor is the leverage ratio itself: at 10:1 leverage, a 1 unit price loss requires roughly 10 units of asset sales to restore the ratio.

**Definition — The margin spiral**

A *margin spiral* operates through the *haircut* rather than through net worth. Margin and haircut requirements are set from estimated volatility, so when volatility rises the required equity per unit of position rises even if no price has moved against the holder and no position has changed size. The holder must therefore either post more equity — which in a crisis is when equity is hardest to raise — or reduce the position. Reductions push prices down and raise realised volatility, which raises the required haircut again, which forces further reductions. Brunnermeier and Pedersen (2009) show that this loop links *funding liquidity* (a trader's ability to obtain financing) and *market liquidity* (an asset's ability to be traded without moving its price) into a mutually reinforcing pair: constrained funding makes market-makers withdraw, which widens spreads and raises measured volatility, which tightens funding further. Their central result is that this feedback makes market liquidity *fragile* — it can collapse abruptly from a small shock once the constraint binds, so liquidity conditions are naturally two-state rather than continuous.

The two spirals are distinct and it is worth being pedantic about the difference. The loss spiral requires a price fall; the margin spiral does not — a rise in volatility with prices unchanged is sufficient to force deleveraging. The loss spiral is proportional to leverage; the margin spiral is

proportional to the sensitivity of haircuts to volatility. In practice they compound: the margin spiral's forced sales cause the price falls that drive the loss spiral, and the loss spiral's sales raise the volatility that tightens the margin spiral. The last sentence of Definition 14.4 is the one that matters most for this book: fragility means the system has two states with a fast transition between them, which is a regime structure derived from institutional plumbing rather than assumed for statistical convenience.

### What it costs: diversification measured in a stress

The consequence for a portfolio is arithmetic, and the arithmetic is worth doing exactly, because the intuition "diversification helps less in a crisis" badly understates what happens.

**Result — Portfolio variance under a common correlation**

Let a portfolio hold weights $w_1, \dots, w_n$ (summing to one) in assets with volatilities $\sigma_1, \dots, \sigma_n$ and pairwise correlations $\rho_{ij}$. Its variance is

$$\sigma_p^2 \;=\; \sum_{i=1}^{n} w_i^2 \sigma_i^2 \;+\; 2 \sum_{i<j} w_i w_j \sigma_i \sigma_j \rho_{ij}.$$

If every pairwise correlation takes the common value $\rho$, this separates into a part that does not depend on $\rho$ and a part that is linear in it:

$$\sigma_p^2(\rho) \;=\; \underbrace{\sum_{i} (w_i \sigma_i)^2}_{S} \;+\; \rho \, \underbrace{2 \sum_{i<j} (w_i \sigma_i)(w_j \sigma_j)}_{C}, \tag{14.1}$$

so $\sigma_p(\rho) = \sqrt{S + \rho C}$ with $C \geq 0$ whenever the weights are positive. Two reference points bound the range: $\sigma_p(0) = \sqrt{S}$ and $\sigma_p(1) = \sum_i w_i \sigma_i$, the undiversified weighted sum of standalone volatilities. The *diversification benefit* at correlation $\rho$ is naturally measured as

$$B(\rho) \;=\; 1 - \frac{\sigma_p(\rho)}{\sum_i w_i \sigma_i},$$

the fraction of the undiversified risk that combining the positions removes. $B$ is zero at $\rho = 1$ by construction. In the equal-weight, equal-volatility case $w_i = 1/n$ and $\sigma_i = \sigma$, (14.1) collapses to

$$\frac{\sigma_p(\rho)}{\sigma} = \sqrt{\tfrac{1}{n} + \left(1 - \tfrac{1}{n}\right)\rho} \;\xrightarrow[n \to \infty]{}\; \sqrt{\rho}, \tag{14.2}$$

which shows both that diversification cannot reduce risk below $\sigma\sqrt{\rho}$ no matter how many instruments are added, and that the *ratio* of portfolio volatility between two correlation states tends to $\sqrt{\rho_{\text{high}}/\rho_{\text{low}}}$ as the book grows.

**Worked Example 14.3 — A three-position book when correlation moves from 0.2 to 0.8**

Hypothetical scenario, computed exactly; not market data. A book holds three positions with the weights and annualised volatilities below — indicative magnitudes of the mid-2020s vintage for a major FX pair, gold and an equity index CFD; verify current levels before using any of them.

| Leg | Weight $w_i$ | Volatility $\sigma_i$ | $w_i\sigma_i$ |
|---|---|---|---|
| AUD/USD | 0.40 | 9.00% | 0.036000 |
| XAU/USD | 0.25 | 14.00% | 0.035000 |
| Index CFD | 0.35 | 16.00% | 0.056000 |
| $\sum_i w_i\sigma_i$ (the $\rho = 1$ case) | | | 0.127000 |

From (14.1), $S = \sum_i (w_i\sigma_i)^2 = 0.0056570$ and $C = 2\sum_{i<j}(w_i\sigma_i)(w_j\sigma_j) = 0.0104720$ (`verify_mk14.py` §3). Then

| | $\sigma_p^2$ | $\sigma_p$ | Diversification benefit $B$ |
|---|---|---|---|
| $\rho = 0.2$ (calm) | 0.0077514 | 8.8042% | 30.6756% |
| $\rho = 0.8$ (stress) | 0.0140346 | 11.8468% | 6.7183% |

**Three readings of the same shift.** The reported volatility rises by 34.56% ($11.8468/8.8042 = 1.3456$). The *variance* — the quantity the risk model actually carries and the one that scales position limits — rises by a factor of 1.8106, very nearly doubling. And the diversification benefit falls from 30.6756% to 6.7183%: the shift destroys $1 - 0.0672/0.3068 = 78.0987\%$ of the entire benefit of having diversified at all. The third figure is the honest one to quote, because it measures what was lost rather than what remains.

**In cash.** On a 250,000 book, a one-day 99% value-at-risk under a normal approximation with 252 trading days per year uses $z_{0.99} = 2.3263$ and $\sigma_{\text{daily}} = \sigma_p/\sqrt{252}$:

$$\text{VaR}_{0.2} = 2.3263 \times 0.5546\% \times 250{,}000 = 3{,}225.56,$$
$$\text{VaR}_{0.8} = 2.3263 \times 0.7463\% \times 250{,}000 = 4{,}340.25,$$

an increase of 1,114.69, or 34.56%, from a change in nothing but the correlation input.

**The book's own response is part of the mechanism.** Under a volatility target the book must scale positions by $\sigma^*/\hat{\sigma}_p$, so the same correlation shift requires multiplying every position by $8.8042/11.8468 = 0.7432$ — a 25.6827% cut in size, executed in the market conditions of Section 14.4's spirals. The risk model is not a passive observer of the stress; its instruction to sell is one of the flows that constitutes it.

**Why "correlations go to one" hurts a large book more.** With only three legs the diversification available was modest to begin with, so its destruction moves volatility by a third. Apply (14.2) to an equal-weight book of the platform's 21 instruments instead: the portfolio volatility is $0.4880\sigma$ at $\rho = 0.2$ and $0.8997\sigma$ at $\rho = 0.8$, a ratio of 1.8439 — the risk very nearly doubles, and in the limit of many instruments the ratio tends to $\sqrt{0.8/0.2} = 2.0000$ exactly. **The more instruments a book holds, the more of its risk budget depends on the correlation assumption, and the more violently that budget moves when the assumption fails.** Breadth is not protection against this; breadth is exposure to it.

**Trap — A hedge ratio estimated in calm is a leverage multiplier in stress**

Worked Example 14.1's data gives gold a regression beta on equities of $-0.7425$ in the calm window and $+0.5888$ in the stress window. Suppose you hedge one unit of equity exposure by buying $h = 1/0.7425 = 1.3467$ units of gold, which sets the calm-window net equity beta to exactly zero. In the stress window the same holding gives a net beta of $1 + 0.5888 \times 1.3467 = 1.7929$ (`verify_mk14.py` §1): the "fully hedged" book carries 79.29% *more* equity exposure than the unhedged one it was built to protect. The failure is not a matter of degree. A hedge whose ratio was fitted in one correlation state does not merely

> stop working when the state changes — when the sign flips, it becomes a second position in the trade it was supposed to offset, and it is levered by the ratio you chose for confidence. Any hedge ratio, spread weight or pair hedge estimated on a sample must be stress-tested at the opposite sign, not merely at a larger magnitude.

The general statement is the one that gives risk management its most uncomfortable line. Diversification is a claim about a correlation matrix; the correlation matrix is a function of the state; and the state that destroys the correlation assumption is precisely the state in which the diversification was supposed to pay. Diversification does not fail randomly. It fails conditionally, and it fails on exactly the days you built it for. Menkhoff, Sarno, Schmeling and Schrimpf (2012) give the currency-specific version — carry returns load negatively on global FX volatility innovations, so the strategy loses in the states where a diversified investor most wants gains — and Lustig, Roussanov and Verdelhan (2011) supply the cross-sectional counterpart: sorting currencies on interest rates produces portfolios loading on a common factor, so adding pairs reduces idiosyncratic noise without touching the factor exposure that generates both the return and the crash.

## 14.5 Carry unwinds as regime events

The carry trade is the cleanest instance of everything the previous two sections derived, which is why it recurs in this library. Chapter 5 established its economics — what it earns, why the premium exists, and what a crash costs in cash. This section makes a different claim: that the carry unwind is not a large draw from the trade's return distribution but a *change of state*, and that treating it as the former is the modelling error the platform's regime machinery exists to prevent.

### The recurring shape

The pattern is stable enough across episodes to be described as a shape. Over months or quarters, an interest differential attracts capital; positions accumulate quietly, each individually sensible; the trade's own success suppresses realised volatility, because a steady one-directional flow into the high-yielding currency dampens its fluctuations; the suppressed volatility loosens every volatility-scaled constraint in Result 14.3's step (4), permitting larger positions still. Then a shock — often unrelated to either currency — raises volatility, step (4) fires in reverse for everyone at once, and months of accumulation unwind in days.

Chapter 5's arithmetic gives the magnitude in a hypothetical but exactly computed case: a yen-funded AUD position earning a 3.6000% gross differential, hit by a 9.7436% spot fall, loses 7.2494% net of the interest earned — which is 2.0137 years of gross carry erased by a single move, or 2.7066 years measured on the spot leg alone (`verify_mk14.py` §4). The asymmetry between accumulation and unwind is not a rhetorical flourish about stairs and lifts; it is the arithmetic consequence of a slow accrual meeting a fast price move, and it is why the distribution is skewed rather than merely fat-tailed.

### Positioning as the accelerant

What converts a shock into an unwind of that size is crowding, and crowding is measurable — imperfectly, and with all the caveats Chapter 7 attached to positioning data. The causal role of positioning is specific and worth stating precisely. Positioning does not predict the timing of the unwind; it determines its *magnitude conditional on a trigger arriving*. A crowded trade and an uncrowded trade facing the same shock produce very different price paths, because step (6) of Result 14.3 — simultaneous, same-signed exit — has a size proportional to the aggregate position. This is exactly the distinction Chapter 7 drew between reading a positioning extreme as a directional signal (which it is not) and reading it as a statement about the conditional tail (which

it is).

Three results from the literature converge on the same operational conclusion. Brunnermeier, Nagel and Pedersen (2008) tie carry's negative skew directly to funding conditions and crowding, making the crash endogenous to the trade's popularity rather than exogenous news. Menkhoff, Sarno, Schmeling and Schrimpf (2012) show carry returns load negatively on global FX volatility innovations, so the payoff's sign is conditional on a volatility state. Lustig, Roussanov and Verdelhan (2011) show interest-sorted currency portfolios load on a common factor, so spreading the trade across many pairs diversifies the noise and not the risk — Worked Example 14.3's $n$-asset arithmetic is the same point in a different notation. The conclusion all three support is that a carry book's risk is a single factor exposure conditioned on a volatility state, and that the appropriate model object is a state-dependent conditional distribution, not an unconditional average.

**Trap — An unwind is a change of state, not a tail draw**

It is tempting to model a carry crash as the left tail of a fixed, fat-tailed distribution: same process, unlucky realisation. That model is wrong in a way that matters operationally. Under it, tomorrow's distribution after a 5% adverse move is the same as it was before, so the correct response is to hold or add. Under the mechanism of Result 14.3, the 5% move has *changed the state*: margin requirements are higher, remaining holders have less equity, measured volatility is higher, and the conditional distribution for tomorrow is now wider, more negatively skewed and serially dependent on today. The two models prescribe opposite actions after the first bad day, and only one of them is consistent with the institutional plumbing. Fat tails in an unconditional fit and regime dependence in the conditional law are not competing descriptions of the same phenomenon — an unconditional fat tail is often what regime dependence *looks like* when you refuse to condition.

### Instances, at public-record level

Named episodes are useful as existence proofs and dangerous as evidence, so they are stated here at the level of the public record and no further. The 2007–2008 crisis provides the fullest documented case of funding-driven cross-asset co-movement, including forced deleveraging by levered intermediaries and the associated unwinding of funded positions (Brunnermeier, 2009). March 2020 supplies a compressed, well-documented instance in which the initial phase saw indiscriminate selling of risk and safe assets alike as participants raised dollar cash, before the usual haven relationships reasserted themselves (BIS, 2020) — the market-scale version of Worked Example 14.2's two phases. In 2024 a widely reported episode of sharp yen-funded unwinding compressed a long accumulation into a very short window; it is described here generically, without magnitudes, because the specific figures are indicative and outside this book's citable record — verify against primary sources before relying on any number attached to it. The Swiss National Bank's discontinuation of the franc's minimum exchange rate on 15 January 2015 (Swiss National Bank, 2015) is a different species of the same genus: a heavily crowded, one-directional position whose unwinding was mechanically forced, in that case by the removal of an official floor rather than by a volatility shock, and whose price adjustment happened faster than positions could be exited. Chapter 15 treats several of these as case studies; the reason to name them here is only to establish that the mechanism has repeatedly operated, not to invite fitting a model to four data points.

## 14.6 Leads, lags, and honest testing

Every practitioner vocabulary contains a set of claimed leading indicators: copper leads global industrial activity; the slope of the yield curve leads the business cycle (Campbell and Shiller, 1991, analyse the predictive content of yield spreads for subsequent rate movements); rate differentials

lead exchange rates (Chapter 5); volatility leads everything. These claims share a structure. Each has a plausible mechanism, each has supporting evidence in some sample, and each is routinely used as though the ordering were both causal and stable. This section states the rule this book applies to all of them and then does the arithmetic that justifies the rule's severity.

### Why predictive orderings drift

Four distinct forces erode a lead-lag relationship, and it is worth separating them because they call for different responses.

*Arbitrage.* If a relationship is real, simple and profitable after costs, capital flows to it and the lead shortens until the profit is gone. The mechanism that destroys the relationship is the same mechanism that made it worth finding (Fama, 1970, states the general principle; Chapter 1 of this book gives the cost-band version). A lead that persists for years in public data usually persists because it is not tradable after costs — which is a statement about your execution, not about the relationship's truth.

*Changing infrastructure.* Which venue reacts first is a fact about market structure, not about economics. As instruments become electronically tradable, as one venue's liquidity migrates to another, and as participants co-locate, the ordering of price discovery changes (Hasbrouck, 2007, develops the information-share machinery that measures which venue leads). A lead-lag estimated on a decade-old sample may be measuring an infrastructure that no longer exists.

*Changing mechanism.* A relationship's sign is conditional on which shock dominates — Table 14.1's bond–equity row is the extreme case. When the shock mix changes, a predictive ordering estimated under the old mix carries the old mix's sign.

*Measurement artefact.* Section 14.1's alignment problem manufactures apparent leads out of nothing: if the ASX closes before New York opens, US returns will "lead" Australian returns at a one-day lag with a large and highly significant coefficient, and no part of that is information. Any claimed lead must first be checked against the trading calendar.

### Granger-style testing, and what it does not establish

The standard tool for the claim "$x$ leads $y$" is a Granger-style test: regress $y_t$ on its own lags and on lags of $x$, and ask whether the $x$ lags add explanatory power. The test answers a narrow question well — does past $x$ help predict $y$ beyond past $y$? — and three things must be said about what it does not answer. It does not establish causation: a third variable $z$ that moves both, but is incorporated into $x$'s price faster than into $y$'s, produces exactly this pattern with no causal arrow from $x$ to $y$ at all, and in cross-asset work such a $z$ (a global funding shock) is not hypothetical but the default explanation. It does not establish economic significance: statistical predictive content of the size typically found is far below the level that survives Chapter 1's no-trade band once spreads and financing are charged. And it does not establish stability: the test is estimated on a sample, and everything in this chapter argues that the relationship it measures is a property of the sample's regime. A Granger-style rejection is evidence that predictive ordering existed in that sample, nothing more — and it is one test among however many were run.

### How many were run: the arithmetic of a lead-lag scan

That last clause is the decisive one, and it can be made quantitative. Consider a modest, entirely reasonable-sounding search on this platform's own universe: 21 instruments, all ordered pairs ($21 \times 20 = 420$), each tested at 5 candidate lags, giving $m = 2{,}100$ candidate lead-lag relationships. Suppose each is evaluated on $n = 60$ weekly observations. Under the null hypothesis that *none* of them has any predictive content, Fisher's transform of each sample correlation is approximately normal with standard error $1/\sqrt{n-3} = 0.1325$ (`verify_mk14.py` §5).

The relevant question is not how large a single correlation would need to be to look significant, but how large the *best of* 2,100 will be when every one of them is pure noise. Integrating the

expected maximum of $m$ independent $|N(0,\, 0.1325^2)|$ draws gives

$$\mathbb{E}\left[\max_{k\le m} |z_k|\right] = 0.4812 \qquad \Longrightarrow \qquad r = \tanh(0.4812) = 0.4472,$$

confirmed by Monte Carlo at 0.4476 over 20,000 trials. **A scan of this size over pure noise is expected to deliver a best lead-lag correlation of about** 0.45**.** It will have a beautiful scatter plot, a plausible economic story invented after the fact, and no content whatsoever.

For contrast, take a single claimed relationship with $r = 0.30$ on the same 60 observations. Its Fisher $z$ is 0.3095, i.e. 2.3368 standard errors, with a two-sided $p$-value of 0.0194 — comfortably "significant" by the usual convention, and comfortably *below* what the noise scan produces. Correcting for the 2,100 tests by a Bonferroni bound at the 5% level requires $|z|/\text{s.e.} > 4.2258$, i.e. $|r| > 0.5078$. A relationship must be roughly 70% stronger than the nominally significant one merely to clear the multiplicity it was found under.

**Trap — A scanned lead-lag is a selection statistic, not a measurement**

The number reported after a search is not an estimate of the relationship; it is the maximum of a set of estimates, and the maximum of noise is not zero. Two rules follow, and they are not negotiable within this programme. First, the count $m$ must include *every* specification examined — all pairs, all lags, all window lengths, all transformations, and all the variants abandoned along the way — not just the ones written up. Second, the resulting statistic must be deflated for that count before it is believed, which is precisely what [Course Book §15.4]'s Deflated Sharpe Ratio machinery does for strategy selection and what the arithmetic above does, cruder but in the same spirit, for correlation selection. An undeflated lead-lag correlation of 0.45 found by scanning 2,100 candidates is not weak evidence of a relationship. It is the expected outcome under no relationship at all, and reporting it as a finding is a category error.

The rule this book applies, then, is simple to state and expensive to follow. *Test, do not trust.* Every intermarket relationship in Table 14.1 — including the ones with the best mechanisms — enters the platform as a hypothesis with a declared search count, is evaluated under the purged, embargoed cross-validation the programme mandates, and must clear the deflation gate [Course Book §15.4] before it may condition a live decision. A good mechanism raises the prior; it does not substitute for the test. The reason this chapter spends so long on mechanisms is not that mechanisms exempt a relationship from testing, but that they tell you *which* relationships are worth spending a test on — and, more importantly, tell you in advance the conditions under which a tested relationship will stop holding.

## 14.7 Why regimes exist

Everything above can now be collected into a single argument, and the argument is the chapter's reason for existing. A regime model is usually introduced statistically: returns exhibit volatility clustering, fat tails and periods of persistent co-movement, so one fits a mixture or a hidden Markov model because it describes the data better than a single-state alternative. That justification is sound but shallow. It says regimes are a convenient description. The claim of this section is stronger: *regimes are a consequence of market structure*, and specifically of the fact that the constraints under which the marginal price-setter operates are inequality constraints.

### The argument from binding constraints

An inequality constraint has two states: slack and binding. A leverage limit, a margin requirement, a value-at-risk budget, a house risk limit, a mandate's volatility target — each is an inequality, and each therefore partitions the state space into a region where it does not affect behaviour and a region where it determines behaviour. That partition is the regime.

When constraints are slack, price formation works the way an economics textbook describes it. Participants trade on information and preferences; a price fall makes an asset cheaper relative to its expected cash flows, which attracts value-motivated buyers; the feedback is *negative* and therefore stabilising; idiosyncratic news dominates each asset's variance, so measured correlations are moderate; liquidity providers earn a spread for supplying immediacy and continue to supply it.

When constraints bind, the marginal participant changes identity. The trader who sets the price is no longer the one with the best view but the one who *must* transact: the holder facing a margin call, the fund hitting a leverage limit, the mandate whose volatility target has been breached. Four consequences follow directly, and each is a stylised fact that regime models are usually asked to fit.

*Price-insensitive supply.* A forced seller sells at the available price, not at a price it considers fair. Its supply curve is inelastic with respect to value, so the usual stabilising response is absent exactly when it is needed.

*Same-signed simultaneity.* Constraints are functions of common inputs — measured volatility, mark-to-market prices, haircut schedules published by a handful of clearing houses and prime brokers. Different participants therefore receive the same instruction at the same time, so individual actions aggregate rather than offset.

*Positive feedback.* Forced selling lowers prices and raises volatility, which tightens the constraints further (Definitions 14.4 and 14.4). The system's gain switches sign relative to the unconstrained state: amplification replaces stabilisation.

*A common factor dominating idiosyncratic variation.* Because the shared driver is the constraint rather than any asset's fundamentals, the common component of returns grows relative to the idiosyncratic component, which is exactly Section 14.4's "correlations go to one".

The four together say that the constrained state has higher volatility, higher cross-asset correlation, fatter tails, and *persistence* — persistence because deleveraging takes time and because the very act of deleveraging keeps volatility elevated, which keeps constraints binding. A state with a distinctive joint distribution and a tendency to persist, entered and exited through a fast transition, is precisely what a regime is. Brunnermeier and Pedersen (2009)'s fragility result is the formal version: the liquidity system has a stable region and an unstable one, with an abrupt boundary between them, and small shocks near the boundary produce large discontinuous changes in behaviour.

**Key idea — The economic content of a regime variable**

Regimes are not a statistical convenience imposed on continuous data; they are the shadow of inequality constraints in the balance sheets of the participants who set prices. This is why the discrete-state modelling of [Course Book §13.1] is well founded here rather than merely convenient: the underlying object genuinely has states, because a constraint is either binding or it is not, and the transition between those conditions is fast relative to the horizons this platform trades. It also explains why regime detection lags. A constraint's binding is observed through its consequences — volatility, correlation, spread, flow — and every one of those is measured on a backward window, so the detector necessarily reports the new state after it has begun, exactly as Worked Example 14.1's rolling correlation did. A regime router is therefore a device for being approximately right about the present, not for being right about the future, and it should be evaluated as such [Course Book §13.8].

### From mechanism to observable

If regimes exist because constraints bind, then the observables worth feeding a regime router are those that measure how close constraints are to binding — not those that describe how the market feels. The distinction is sharp and it determines the covariate list. Table 14.2 maps this chapter's

mechanisms one-for-one onto quantities the platform already builds or is building.

| Mechanism (this chapter) | Observable | What the router does with it |
|---|---|---|
| Constraint tightness via measured volatility (Def. 14.4) | VolRegime state from realised and, where available, implied volatility | Selects the regime; widens the envelope's tails in the high-volatility state |
| Global risk appetite / funding conditions (Rey, 2015; Result 14.3) | RORO composite (Chapter 7): oriented, causally standardised cross-asset basket | Conditions which state the router assigns; never traded directly |
| Crowding, which sets the magnitude of an unwind conditional on a trigger | COT net positioning $z$-scores, point-in-time by release date (Chapter 7) | Conditions the tail asymmetry rather than the conditional mean |
| Common-factor dominance in stress (Section 14.4) | Rolling cross-asset correlation, backward-only, fixed $W$ (Def. 14.1) | Feeds portfolio-level sizing; flags when instrument count stops delivering diversification |
| Discount-factor shocks (Mechanism 1) | Rate and macro covariates from FRED-style series, point-in-time via a vintage-aware store | Conditions the level and the state; the known-future calendar is typed separately |
| Market-liquidity fragility (Brunnermeier and Pedersen, 2009) | The measured bid/ask cost surface replacing mid-only candles | Raises modelled cost in the stressed state, where fills are worst |

**Table 14.2.** The regime router's inputs, each traced to the market mechanism that makes it informative. Every observable is a measurement of how close a constraint is to binding, not a measurement of sentiment.

**Your project — Why the Stage-2 router earned its place**

The router [Course Book §13.8] is the platform component with the weakest statistical justification and the strongest *market* justification, and this chapter is where the second is supplied. Four consequences for the build follow from the argument above rather than from any backtest.

**(1) The router conditions; it does not decide.** Everything in this chapter says the state changes the conditional distribution — its width, its skew, its correlation structure. Nothing in this chapter says the state predicts direction. A router that emits a trade has converted a measurement into a forecast without evidence, which is the category error Chapter 7 warned against in the same words.

**(2) The covariates must measure constraints, not moods.** Table 14.2 is the list, and the criterion for adding to it is mechanical: can you name the constraint this series measures the tightness of? A series that fails that question is a candidate for the deflation machinery [Course Book §15.4] like any other, with no head start.

**(3) Correlation belongs in the covariate set, and in sizing.** Worked Example 14.3 showed that on a 21-instrument book the same correlation shift that moves a three-leg portfolio's volatility by 34.56% moves an equal-weight book's by 84.39%. A sizing rule that treats the instrument count as diversification while holding the correlation input fixed is not conservative; it is wrong in the direction that hurts. The rolling correlation must be point-in-time, backward-only and fixed-window, per the constraints already stated.

**(4) The evaluation must be per-regime.** Because the conditional law differs across states, a pooled performance statistic is the pooled correlation of Worked Example 14.1 in another guise: a number describing no state the system was ever in. Per-regime evaluation is not a refinement of the test battery; it is the only way the battery can answer the question the router poses.

The final observation is the one that should be uncomfortable, and it belongs here rather than in the summary. The platform's own risk machinery is part of the mechanism this chapter

describes. A volatility target that cuts positions by 25.68% when correlations rise is issuing the same instruction, at the same moment, as every other volatility-targeted book in the market — which is step (4) of Result 14.3 operating on your account. Being a small participant means your own selling does not move the price; it does not mean you are outside the loop. It means you are one of the many simultaneous, same-signed sellers whose aggregate does. The correct response is not to abandon the rule — an unconstrained book in a spiral is worse — but to model the cost of executing it in the state where it fires, which is exactly why the platform's cost surface must be state-dependent rather than a single average spread.

## 14.8 Exercises

Five exercises: two computational — the rolling-correlation flip and the portfolio-volatility shift, which are the chapter's two load-bearing calculations — and three that require the mechanisms to be assembled in prose, because a mechanism that cannot be reconstructed from memory has not been learned.

**Exercise 14.1 — A rolling correlation that inverts**

The table below gives ten consecutive weekly percentage returns for AUD/USD and for a global equity index (hypothetical, constructed; not market data).

| Week | AUD/USD % | Index % | Week | AUD/USD % | Index % |
|---|---|---|---|---|---|
| 1 | +0.80 | +1.50 | 6 | −1.40 | +1.60 |
| 2 | −0.50 | −0.40 | 7 | +0.90 | −1.30 |
| 3 | +1.10 | +2.00 | 8 | −0.60 | +1.20 |
| 4 | +0.30 | +1.10 | 9 | +1.80 | −1.90 |
| 5 | −0.90 | −1.70 | 10 | −1.10 | +0.30 |

(a) Using Definition 14.1 with $W = 5$, compute the correlation over weeks 1–5 and over weeks 6–10, and the pooled correlation over all ten weeks. State what each of the three numbers claims about the relationship.

(b) Compute the full rolling path $\hat{\rho}_t(5)$ for every admissible $t$ and identify between which two values the sign changes. How many weeks after the regime change at week 6 does the estimator's sign actually change, and why?

(c) Test whether the flip survives its own estimation noise, using the Fisher transform with standard error $1/\sqrt{W-3}$ and the difference-of-$z$ construction from Worked Example 14.1. Also report the 95% confidence interval for the weeks 1–5 estimate, and comment on the contrast between the two results.

(d) A trader hedges a long AUD/USD position by shorting the index, sizing the hedge on the weeks 1–5 regression beta of AUD/USD on the index. Compute that beta, compute the weeks 6–10 beta, and state in one sentence what happens to the hedged book in the second window.

▷ working space

**Solution — Exercise 14.1**

**(a)** (`verify_mk14.py` §6.) Weeks 1–5: $\hat{\rho} = +0.9812$. Weeks 6–10: $\hat{\rho} = -0.9371$. Pooled over weeks 1–10: $\hat{\rho} = -0.1761$.

The first window makes the standard claim: AUD/USD is a growth-sensitive currency moving with global equities. The second makes the opposite claim: AUD/USD moves *against* equities. The pooled number claims there is essentially no relationship at all — which is the worst of the three answers, because it is not a compromise between two states but the arithmetic residue of averaging a strong positive and a strong negative relationship. A model fed $-0.1761$ would treat the two series as nearly independent in both windows, when in fact they were nearly collinear in each, with opposite signs.

**(b)** Rolling path with $W = 5$:

| $t$ (week) | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|
| Window | 1–5 | 2–6 | 3–7 | 4–8 | 5–9 | 6–10 |
| $\hat{\rho}_t(5)$ | $+0.9812$ | $+0.4506$ | $+0.0858$ | $-0.3407$ | $-0.6778$ | $-0.9371$ |

The sign changes between $t = 7$ and $t = 8$, i.e. two weeks after the regime changed at week 6. The lag is structural, not a defect of these particular numbers: at $t = 6$ the window still contains four observations from the old regime and one from the new, so the old regime outvotes the new one; the estimator cannot cross zero until the new regime supplies enough of the window's observations and enough of its variance. A window of length $W$ carries the old state for up to $W - 1$ periods after the change, and the readings taken during the crossing — $+0.0858$ at $t = 7$ — are the least informative of the whole path, which is a general property of windowed estimators and not peculiar to correlation.

**(c)** With $W = 5$ the Fisher standard error is $1/\sqrt{5-3} = 1/\sqrt{2} = 0.7071$. The two window estimates transform to $z_{1:5} = +2.3285$ and $z_{6:10} = -1.7133$. Their difference is 4.0418 with standard error $\sqrt{2} \times 0.7071 = 1.0000$, so the flip is 4.0418 standard errors, two-sided $p = 0.000053$: the two windows are not two draws from a single correlation.

The 95% interval for the weeks 1–5 estimate alone is $\tanh(2.3285 \pm 1.9600 \times 0.7071) = [+0.7364,\ +0.9988]$.

The contrast is the instructive part. The individual estimate is very imprecise — the data are consistent with anything from a moderate $+0.74$ to a near-perfect $+0.999$ — yet the *difference* between the two windows is decisively established. Estimation noise and structural change are separate questions, and a five-observation window can answer the second convincingly while answering the first badly. It follows that reporting a rolling correlation as a point estimate, without an interval, overstates what has been measured, while refusing to conclude that a flip occurred because the individual windows are noisy understates it.

**(d)** Regression beta of AUD/USD on the index: weeks 1–5 gives $\beta = +0.5464$, weeks 6–10 gives $\beta = -0.8402$, pooled gives $-0.1302$. A trader who hedges a long AUD/USD position by shorting 0.5464 units of index per unit of AUD/USD sets the first window's net index exposure to zero. In the second window the true beta is $-0.8402$, so the short index leg no longer offsets anything — it *adds* exposure of the same sign as the AUD/USD position, and the "hedged" book has larger net exposure than the unhedged one, exactly as in the trap following Worked Example 14.3. The hedge did not degrade; it reversed.

**Exercise 14.2 — Portfolio volatility under two correlation states**

A book holds three positions with the weights and annualised volatilities below (indicative magnitudes; verify current levels before use):

| Leg | Weight $w_i$ | Volatility $\sigma_i$ |
|---|---|---|
| FX major | 0.50 | 10.00% |
| Gold | 0.30 | 20.00% |
| Crypto | 0.20 | 35.00% |

(a) Using Result 14.4, compute $S$, $C$, and $\sum_i w_i\sigma_i$, then the portfolio volatility at a common correlation of $\rho = 0.15$ and at $\rho = 0.75$.

(b) Report the ratio of the two volatilities and of the two variances, and compute the diversification benefit $B(\rho)$ in each state and the fraction of it destroyed by the shift.

(c) The book runs a volatility target. By what percentage must every position be cut when the correlation state changes from 0.15 to 0.75, holding the target and the individual volatilities fixed?

(d) On a 400,000 book, compute the one-day 99% value-at-risk in each state under a normal approximation with 252 trading days per year and $z_{0.99} = 2.3263$, and the increase.

(e) In two or three sentences, explain why part (c)'s answer is a market mechanism and not merely a portfolio-management convention.

▷ working space

**Solution — Exercise 14.2**

**(a)** (`verify_mk14.py` §7.) The risk contributions are $w_i\sigma_i = 0.050000,\ 0.060000,\ 0.070000$, so

$$S = \sum_i (w_i\sigma_i)^2 = 0.0110000, \qquad C = 2\sum_{i<j}(w_i\sigma_i)(w_j\sigma_j) = 0.0214000, \qquad \sum_i w_i\sigma_i = 0.180000.$$

Then $\sigma_p^2(0.15) = 0.0110000 + 0.15 \times 0.0214000 = 0.0142100$, giving $\sigma_p = 11.9206\%$; and $\sigma_p^2(0.75) = 0.0110000 + 0.75 \times 0.0214000 = 0.0270500$, giving $\sigma_p = 16.4469\%$.
**(b)** Volatility ratio $16.4469/11.9206 = 1.3797$; variance ratio $0.0270500/0.0142100 = 1.9036$. The diversification benefits are

$$B(0.15) = 1 - \frac{0.119206}{0.180000} = 33.77\%, \qquad B(0.75) = 1 - \frac{0.164469}{0.180000} = 8.63\%,$$

so the shift destroys $1 - 0.0863/0.3377 = 74.45\%$ of the diversification benefit. Note the pattern from Worked Example 14.3 repeating with different inputs: volatility rises by about a third, variance by about a factor of two, and roughly three-quarters of the diversification disappears. The third figure is the one that describes what was lost.
**(c)** Volatility targeting scales positions by $\sigma^*/\hat{\sigma}_p$, so the required multiplier is $11.9206/16.4469 = 0.7248$: every position must be cut by 27.52%.
**(d)** Daily volatilities are $11.9206\%/\sqrt{252} = 0.7509\%$ and $16.4469\%/\sqrt{252} = 1.0361\%$. Hence

$$\text{VaR}_{0.15} = 2.3263 \times 0.7509\% \times 400{,}000 = 6{,}987.65, \qquad \text{VaR}_{0.75} = 2.3263 \times 1.0361\% \times 400{,}000 = 9{,}640.91,$$

an increase of 2,653.26, or 37.97% — from a change in the correlation input alone, with no change in any position, any price or any individual volatility.
**(e)** The 27.52% cut is not a private housekeeping decision. Every volatility-targeted book in the market receives the same instruction from the same input at the same time, because measured volatility and measured correlation are public and shared. The aggregate of those simultaneous, same-signed reductions is itself the selling that Section 14.4's spirals describe and that pushes prices and volatility further in the direction that triggered the cut. The risk rule is therefore an input to the market state it is responding to — amplification by construction, not by intention — which is precisely the constraint-driven feedback that Section 14.7 identifies as the reason regimes exist.

**Exercise 14.3 — Ordering the risk-off cascade**

The eight events below are listed in scrambled order. They describe a single risk-off episode in a market where leveraged investors have accumulated positions funded in a low-yielding currency $L$.

- **A.** Leveraged investors buy $L$ to repay their funding liabilities.
- **B.** An exogenous shock raises realised and implied volatility.
- **C.** Brokers raise margin requirements and internal value-at-risk limits bind.
- **D.** Over several quarters, investors borrow $L$ and convert it to fund higher-yielding positions, building a large aggregate short in $L$.
- **E.** Remaining holders of the trade take further mark-to-market losses, and their measured risk rises again.
- **F.** $L$ appreciates sharply against the higher-yielding currencies.
- **G.** Cross-asset correlations rise as forced sellers liquidate across their whole book.
- **H.** Required equity per unit of exposure rises, so positions must be reduced.

(a) Place the eight events in causal order, and identify which pair of events forms the loop that makes the episode self-reinforcing.

(b) For each of the three constraints named in Result 14.3's step (4), state whether it would still fire if the shock raised volatility but left all prices unchanged.

(c) State the one condition under which this cascade would *not* occur despite a large volatility shock, and name the Chapter 7 data source that speaks to it.

▷ working space

**Solution — Exercise 14.3**

**(a)** The causal order is

$$\mathrm{D} \to \mathrm{B} \to \mathrm{C} \to \mathrm{H} \to \mathrm{A} \to \mathrm{F} \to \mathrm{E} \to \mathrm{G},$$

matching Result 14.3's steps (1)–(2), (3), (4), (4), (5), (6)–(7), (7), with G — the cross-asset correlation rise — as the consequence that appears once forced sellers are liquidating across their whole book rather than only in the funded trade (Section 14.4, mechanism 2).
The self-reinforcing loop is E → C/H: the mark-to-market losses and higher measured risk in E return the system to the constraint step (C and H), which produces more selling (A), which produces more appreciation (F), which produces more losses (E). Accept F → E → H → A → F as the same cycle traced with a finer step. The essential recognition is that the loop closes through the *constraint*, not through sentiment: no participant's opinion changes anywhere in the cycle.
**(b)** All three would still fire with prices unchanged, and this is the substantive content of the question. Broker margin and haircut schedules are functions of estimated volatility, so a volatility rise raises required equity per unit of position directly (Definition 14.4). A value-at-risk limit is a function of the estimated variance, so a variance rise breaches it with no price move at all. A volatility target scales position size by $\sigma^*/\hat{\sigma}$, which shrinks as $\hat{\sigma}$ rises regardless of P&L. The contrast is with the loss spiral (Definition 14.4), which requires an actual price fall to erode net worth. That is exactly why the two spirals are defined separately: one needs a loss, the other needs only a change in the estimate of risk.
**(c)** The cascade requires an aggregate short in $L$ to exist. If the trade has already been unwound — because $L$'s policy rate rose, because a previous episode forced the exit, or because the differential compressed — then step (5) has nothing to buy back, and $L$ need not rally at all despite a large volatility shock. The relationship in Table 14.1's second row is conditional on a positioning state. Chapter 7's Commitments of Traders data, read as net positioning and its $z$-score, is the public proxy that speaks to this condition — with all of that chapter's caveats attached: it measures a futures population that is at best correlated with the OTC and CFD positioning that actually matters, and it is dated by reference Tuesday and released the following Friday, so it must be used point-in-time.

**Exercise 14.4 — Classifying four macro states on the dollar smile**

For each of the four hypothetical macro states below, state which leg of the dollar smile (Section 14.2) it corresponds to, which of the dollar's three structural roles is doing the work, the implied direction for the broad dollar, and the implied direction for a growth-sensitive currency such as AUD. Then state one observation that would falsify your classification.

**A.** Global growth is broadly synchronised and above trend; credit spreads are tight; volatility is low; the United States is growing at about the same pace as its peers.

**B.** A large financial institution outside the United States fails; dollar funding spreads widen sharply; equity volatility doubles within a week.

**C.** US growth and inflation surprise strongly to the upside while European and Asian growth disappoints; US policy rates are expected to rise relative to other centres; equity markets are firm.

**D.** A shock originating in the United States — a domestic banking problem — causes US equities to fall sharply, and the Federal Reserve is expected to cut rates faster and further than any other central bank.

▷ working space

**Solution — Exercise 14.4**

**A — the middle of the smile.** No role is under stress; capital seeks higher expected returns abroad, which requires selling dollars. Broad dollar: weaker. AUD: stronger. This is the state in which a carry book is comfortable and Table 14.1's first row holds most reliably. *Falsifier*: the dollar strengthening anyway while credit spreads stay tight and volatility stays low — which would indicate a relative-return story (the right leg forming) that the "synchronised" description has missed.

**B — the left leg.** The funding and reserve roles dominate: dollar liabilities must be serviced, dollar collateral is demanded, and positions funded in dollars are unwound. Broad dollar: stronger. AUD: weaker. This is the canonical risk-off configuration, and by Result 14.3 the yen and franc should strengthen too if they are currently being used to fund. *Falsifier*: the dollar weakening while funding spreads widen, which would say the stress is being met by an expected policy response large enough to dominate the funding demand — i.e. the episode is closer to state D than to state B.

**C — the right leg.** The relative-return channel dominates; neither the funding nor the reserve role is stressed. Broad dollar: stronger. AUD: weaker, but for reasons opposite to B — here the dollar rises because the news is *good* for the United States. *Falsifier*: equity volatility rising sharply alongside the dollar, which would say the move is a funding event rather than a growth-differential event, and reclassify the state as B. Note the practical importance: B and C both produce "strong dollar, weak AUD", but they imply opposite signs for equities, so a covariate that records only the dollar's direction cannot separate them.

**D — ambiguous by construction, and that is the point.** The reserve and funding roles push the dollar up; the relative-return and policy-easing channels push it down. Which dominates is an empirical question about the specific episode and cannot be resolved from the description. Broad dollar: indeterminate. AUD: weaker against most crosses on the risk-off channel, but its direction against the dollar specifically is exactly the ambiguity. *Falsifier*: none is needed, because no directional claim has been made — and refusing to make one is the correct answer. This is the state Chapter 7 had in mind when it called the dollar the least reliable leg of a risk-off basket, and it is a concrete instance of why a regime router must carry more than one observable: the distinction between B and D is invisible in the dollar's price and visible in the pairing of funding spreads with expected policy paths.

**Exercise 14.5 — Why the regime router earned its place**

Write an argument, of roughly a page, justifying the inclusion of a regime-conditioning layer in a systematic trading system, using *only* market mechanisms established in this chapter. You may not appeal to backtest performance, to the fact that regime models fit financial data well, or to any statistical property of returns that is not derived from a mechanism. Your argument must identify what the router may and may not be used for, and must state the strongest objection to it that you can construct, together with your answer to that objection.

▷ working space

**Solution — Exercise 14.5**

A model answer follows; the marking criterion is whether every claim traces to a mechanism rather than to a fit.

**The argument.** Market participants who set prices operate under constraints — margin requirements, haircuts, leverage limits, value-at-risk budgets, volatility targets — and these are inequality constraints, each with a slack region and a binding region. In the slack region the marginal price-setter is the participant with the best view, a price fall attracts value buyers, feedback is negative and stabilising, and idiosyncratic news dominates each asset's variance, so cross-asset correlations are moderate. In the binding region the marginal price-setter is the participant who *must* transact; its supply is price-insensitive; the constraints that bind are functions of shared inputs, so many participants act simultaneously and in the same direction; the resulting price falls and volatility rises tighten the constraints further, making the feedback positive; and the common driver of returns becomes the constraint rather than any asset's fundamentals, so correlations rise towards a single factor. The two regions therefore have different conditional distributions — different volatility, different correlation, different skew — and the binding region persists, because deleveraging takes time and itself sustains the elevated volatility that keeps constraints binding. A system whose conditional distribution has two persistent states with a fast transition between them is a regime process by definition. A model that assumes one state is therefore not conservative; it is misspecified in a known direction, fitting a single law to a mixture whose components differ in width, in tail asymmetry, and in the correlation structure that governs portfolio risk.

Three chapter-specific consequences make the layer necessary rather than merely defensible. First, the funding-currency mechanism shows that a relationship's *sign* can be conditional: a funding currency rallies in stress only while an aggregate short exists to be bought back, so a model carrying an unconditional coefficient learns an average of two opposite-signed states. Second, the portfolio arithmetic shows that the same nominal book is a materially different risk object in the two states — on a 21-instrument book, a correlation shift from 0.2 to 0.8 raises portfolio volatility by 84.39% with no position changed — so a sizing rule that ignores the state is systematically wrong at the moment it matters most. Third, the two-phase behaviour of havens shows that even a well-founded relationship such as gold's real-yield link is suspended while liquidation is the binding constraint, so a hedge sized outside the state can invert inside it.

**What the router may be used for.** It may condition: select which conditional distribution applies, widen the envelope's tails in the constrained state, adjust the cost model to the state in which fills are worst, size positions using a state-appropriate correlation, and evaluate performance per state. **What it may not be used for.** It may not emit a direction. Nothing in the mechanism above says a constrained state predicts the sign of the next return; it says the distribution around whatever happens is wider, more skewed and more correlated. Converting a state measurement into a trade is an unsupported inference.

**The strongest objection.** Every observable that reveals the state — realised volatility, rolling correlation, spread, positioning — is measured on a backward window, so the router detects the new state only after it has begun. In the worked example the rolling correlation's sign changed two weeks after the regime did, and read approximately zero at the transition itself, which is exactly when a state estimate would have been most valuable. If the router is always late, the objection runs, it is describing history and can add nothing.

**The answer.** The objection is correct that the router cannot call the transition and wrong that this makes it useless, for two reasons that follow from the same mechanisms. The constrained state *persists*: deleveraging takes days to weeks, and while it proceeds the elevated volatility keeps the constraints binding, so a state identified late is still the state that will obtain over the horizon on which positions are held. And the router's outputs are used to condition

distributional width, sizing and cost — quantities for which being approximately right about the present is sufficient — rather than to time entries, for which it would need to be right about the future. A device that says "we are now in the constrained state, so widen the envelope, cut size and charge stressed costs" is useful the whole time the state lasts, and its lag costs it only the transition itself. It is worth conceding explicitly that this makes the router weakest in precisely the sharp, single-day dislocations where lag is fatal; the honest defence is that its value is concentrated in episodes that persist, and that the appropriate protection against instantaneous dislocation is position sizing and the cost model, not detection.

**Key idea — What to remember**

Cross-asset relationships are real, mechanically explicable, and unstable, and all three of those facts are load-bearing: the mechanisms tell you which relationships are worth testing and under what conditions they will stop holding, while the instability means no sign in this chapter's map may be used without being measured on a backward-only window and re-measured continuously. The central mechanism is that funding, margin and risk-limit constraints are inequalities, so the same market has two modes — one in which value-motivated trading stabilises prices and idiosyncratic news dominates variance, and one in which constrained participants transact simultaneously in the same direction, correlations converge on a common factor, and the feedback amplifies rather than damps. That is why diversification fails on exactly the days it was built for, why a hedge ratio fitted in calm can become a leverage multiplier in stress, and why a carry unwind is a change of state rather than an unlucky draw from a fixed distribution. It is also the market-structure justification for this platform's entire regime apparatus [Course Book §13.1]: the router exists because the market genuinely has states, it may condition but never decide, its observables must each measure how close a constraint is to binding, and every claimed intermarket lead must clear the deflation gate [Course Book §15.4] before it may condition anything at all. Chapter 15 takes six episodes in which these mechanisms ran to completion and examines what broke in each, which is the same argument told through the public record rather than through derivation.

# 15

# Crises: Six Case Studies in Market Failure

*Chapter 14 named the mechanisms — funding constraints, margin and loss spirals, correlation collapsing to one — and argued that they make market behaviour state-dependent. This chapter runs those mechanisms at full amplitude and watches what they break. Six episodes are examined at the level of the public record: what the machinery was, what it did when stressed, and what a systematic CFD trader should change because of it. The selection is not a history; it is a taxonomy of failure modes, and the chapter's centre of gravity is the one episode that killed retail traders and their brokers outright — the removal of the Swiss franc floor on 15 January 2015 — because that is the failure mode the reader's own book is most exposed to. Each case is treated the same way: mechanics, then what broke, then the lesson, and the lesson is always arithmetic rather than sentiment. The chapter ends with the honest ledger — which failure modes the platform's machinery genuinely defends against, which it merely detects after the fact, and which nothing whatsoever defends against except holding a smaller position than the margin rules permit.*

## 15.1 October 1987: the feedback machine

The crash of 19 October 1987 is the cleanest case in this chapter because the amplifying mechanism was not a hidden exposure or an opaque funding chain: it was a published, widely adopted, and internally coherent trading rule. Portfolio insurance promised institutional equity holders a floor on their portfolio value without the cost of buying listed puts, and it delivered that promise by *manufacturing* the put — replicating its payoff dynamically in index futures. The replication is correct. The problem is what the replication instructs you to do, and what happens when everyone is instructed to do it at the same instant.

### The mechanics: a hedge that sells into weakness

A protective put is a long index position plus a long put struck at the desired floor $K$. Under the Black–Scholes assumptions (Hull, 2021; the Greeks are derived in the Course Book, [Course Book §20.3]), the value of that combined position decomposes exactly as

$$S + p(S,K,T) \;=\; S\,N(d_1) \;+\; Ke^{-rT}N(-d_2), \qquad d_1 = \frac{\ln(S/K) + \left(r + \tfrac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}, \quad d_2 = d_1 - \sigma\sqrt{T},$$

where $S$ is the index level, $T$ the horizon to the floor date, $\sigma$ the volatility used by the replicator, $r$ the financing rate, and $N(\cdot)$ the standard normal distribution function. Read the right-hand side as a portfolio instruction rather than a valuation formula: to hold the insured position synthetically, hold $N(d_1)$ units of index per unit insured and place the remainder, $Ke^{-rT}N(-d_2)$, in cash. Since $N(d_1)$ is increasing in $S$, the instruction is unambiguous — *as the market falls, sell*. And since $N(\cdot)$ is steepest near the money, the amount you must sell per unit of decline *grows* as the market approaches the floor. That second derivative is the gamma of the replicated option, and it is the entire story of October 1987.

**Worked Example 15.1 — The replication schedule: how much a synthetic put sells**

A fund insures a portfolio of 1,000,000,000 currency units against a floor set 5 per cent below the current index, with three months to the floor date. Take the index at $S_0 = 300.00$, floor $K = 285.00$, $T = 0.25$ years, replication volatility $\sigma = 0.20$, and financing $r = 7\%$. To isolate the price effect, hold $T$, $\sigma$ and $r$ fixed across a single session. Compute the index exposure the replication demands at $S = 300$, 285 and 270, and the resulting sell orders.

*Units insured.* $n = 1{,}000{,}000{,}000/300 = 3{,}333{,}333.33$ index units.
*Replication delta at each level.* With $\sigma\sqrt{T} = 0.20 \times 0.5 = 0.10$ and $(r + \frac{1}{2}\sigma^2)T = (0.07 + 0.02) \times 0.25 = 0.0225$:

$$\begin{aligned} S = 300: &\quad d_1 = \tfrac{\ln(300/285)+0.0225}{0.10} = \tfrac{0.0512933+0.0225}{0.10} = 0.737933, &\quad N(d_1) = 0.769722, \\ S = 285: &\quad d_1 = \tfrac{0+0.0225}{0.10} = 0.225000, &\quad N(d_1) = 0.589010, \\ S = 270: &\quad d_1 = \tfrac{-0.0540672+0.0225}{0.10} = -0.315672, &\quad N(d_1) = 0.376126. \end{aligned}$$

*Index units held.* $n\,N(d_1)$ gives 2,565,741.34, then 1,963,367.88, then 1,253,752.20 units.
*The sell orders.*

$$\begin{aligned} &\text{leg 1 } (300 \to 285,\ -5.00\%): &&\text{602,373.46 units, worth 171,676,435.97 at 285,} \\ &\text{leg 2 } (285 \to 270,\ -5.26\%): &&\text{709,615.68 units, worth 191,596,232.55 at 270.} \end{aligned}$$

*Reading.* The second leg requires $709{,}615.68/602{,}373.46 = 1.1780$ times the first leg's selling — 17.8 per cent more stock sold for a decline that is, in index points, identical. The insured book has gone from 77.0 per cent invested to 37.6 per cent invested on a 10 per cent fall. Ten funds running the same programme on the same notional would have placed 1.72 billion of sell orders on the first leg and 1.92 billion on the second, at prices they themselves were setting.

The worked example understates the real dynamic in three ways, each of which pushed in the same direction in 1987. First, $T$ shrinks as the floor date approaches, and $N(d_1)$ steepens as $T \to 0$ — late in the insurance period the same decline demands more selling. Second, the replication volatility $\sigma$ is an input, and a falling, volatile market forces it upward; raising $\sigma$ when $S < K$ raises $N(d_1)$, but the practical response of insurers whose realised path had already breached assumptions was to re-hedge more aggressively, not less. Third, and decisively, the worked example prices the sales at the market's undisturbed level. They are not executed there. Shleifer (1986) established the point this chapter needs: demand curves for equities slope down, so large uninformed sales move prices and the move is not a transient printing error to be arbitraged away instantly. The mechanical seller therefore moves the price that determines how much more he must sell.

### What broke

The Brady Commission (1988) inquiry into the events identified concentration of mechanical selling as a principal amplifier, and documented the market-structure failures that let the amplifier run without a governor. Three matter here.

*The futures market absorbed the selling and disconnected from cash.* Index futures were where portfolio insurers executed, because that is where a whole basket can be sold in one order. On the day, the futures traded at a substantial discount to the cash index. In normal conditions index arbitrage closes such a gap — buy the future, sell the basket (Chapter 9's fair-value arithmetic) — but the arbitrage requires selling the cash basket, and the cash market's order-routing and reporting systems were saturated, so arbitrageurs faced execution and reporting uncertainty that made the apparent discount uncapturable. When the arbitrage link fails, the two prices are no longer two quotes on one asset; they are two separate, thin markets.

*The designated liquidity providers ran out of capital.* Specialists on the cash market are obliged to make markets, but the obligation is bounded by their balance sheets. Absorbing one-directional flow of that size means taking a position, and taking a position while the price falls consumes exactly the capital that funds the next quote. This is the loss spiral of Chapter 14 acting on the market maker rather than the speculator, and it is the general result: *liquidity provision is a levered position in the asset*, so a fall in the asset reduces the supply of liquidity precisely when demand for it peaks.

*The price move had no proportionate news.* Public record puts the one-day fall in the S&P 500 at roughly 20 per cent, with no macroeconomic release of remotely commensurate content. That absence is the diagnostic: a price move without news is a move about the market's own plumbing, and the plumbing was mechanical selling meeting exhausted intermediation.

The structural legacy is visible today in the instruments the reader trades. Exchange-level circuit breakers and, later, single-stock limit mechanisms exist because of this episode: the venues concluded that the correct response to a self-feeding liquidity failure is to stop trading and restart the auction. Index option markets also changed permanently — the pronounced negative implied-volatility skew that Gatheral (2006) treats as a standing feature of index surfaces was not a standing feature before 1987. Markets priced a crash mechanism into the option surface once they had seen it work.

### The lesson for a systematic book

Three transfer directly.

*Crowding manufactures the tail.* A backtest is conditioned on a population of other market participants that appears nowhere in the data. Portfolio insurance was profitable and low-risk while few ran it, and catastrophic when many did, and no amount of historical resampling of pre-1987 prices would have revealed this, because the mechanism operated through a variable — how much other capital follows the same rule — that the price series does not contain. A strategy that is popular is a different strategy from the same strategy when it was obscure. The platform's honest defence is not detection but scepticism: the deflation machinery [Course Book §15.4] exists to stop a backtest earning belief it has not paid for, and crowding is one more reason the raw in-sample Sharpe overstates the future.

*Any rule that trades in the direction of the move is a short volatility position, whatever it is called.* Stop-losses, volatility targeting that cuts on realised-vol spikes [Course Book §14.4], trend entries, and margin-driven liquidation are all, mechanically, the same instruction as a replicated put: sell lower, buy higher. Individually each is defensible. What 1987 shows is that their aggregate is a demand curve with the wrong slope, and that in a crowded market the aggregate is what sets the price. The reader's own stop discipline is a small piece of that aggregate.

*The venue can stop.* A circuit breaker or limit halt suspends the reference price on which a CFD is quoted. What happens to an open CFD position when the underlying reference halts is a matter of the provider's terms, not of market logic: quoting may continue on a widened spread, may be suspended, or may resume at a gap. This is a structural feature to be confirmed in the reader's own contract — verify current terms — and simulated as a scenario, not discovered live.

## 15.2 LTCM 1998: leverage against convergence

Long-Term Capital Management is the case every quantitative trader is told about and few draw the right conclusion from. The wrong conclusion is that the models were bad. The models were, by the standards of the time, good; several of the trades converged as predicted, some within months of the fund's collapse. The right conclusion is structural: *a levered position has two success conditions, and being right is only one of them.* The other is surviving the path, and the path is governed by financing terms and by who else holds the trade — neither of which appears in a convergence model.

### The mechanics: why convergence trades must be levered

A convergence, or relative-value, trade takes two claims on nearly the same cash flows that trade at slightly different prices, buys the cheap one, sells the dear one, and hedges the common risk factor so that what remains is exposure to the *spread* alone. Off-the-run against on-the-run government bonds, swap spreads, cross-market sovereign spreads, and index volatility against constituent volatility are all instances. Lowenstein's (2000) account describes a book made of many such positions across markets and geographies.

The defining feature of these trades is not their sophistication but their thinness. Write $s$ for the spread in decimal terms and $D$ for its spread duration — the sensitivity, in units of position value, to one unit of spread change. Unlevered, a full convergence from $s$ to zero earns $Ds$ per unit of notional. At $D = 5$ years and $s = 10$ basis points, that is $5 \times 0.0010 = 0.50$ per cent of notional, gross, for a trade that may take a year and consumes a research team. No institutional investor funds that. Now apply leverage $L$ (assets per unit of equity):

$$\text{return on equity} \;=\; L\,D\,s.$$

To turn 0.50 per cent of notional into 12.5 per cent on equity requires $L = 25$. The leverage is not an appetite for risk; it is the arithmetic consequence of harvesting a thin, reliable spread. This is the first structural fact: *relative-value strategies are levered because their edge is small, and their edge is small because the position is hedged*. The hedging that makes the trade look safe is what forces the leverage that makes it lethal.

> **Result — Symmetry of the levered convergence trade**
>
> For a levered spread position with leverage $L$, spread duration $D$, and an adverse spread widening of $\Delta s$, the fraction of equity lost is
>
> $$\ell \;=\; L\,D\,\Delta s,$$
>
> the same expression that gives the profit on convergence, with the sign reversed. Widening and convergence are symmetric in profit but *not* in consequence: convergence pays into an account that survives, widening consumes the equity that finances the wait.

At $L = 25$ and $D = 5$: a 10 basis point widening costs 12.5 per cent of equity, 20 basis points costs 25 per cent, 40 basis points costs 50 per cent, and 80 basis points — less than one per cent of notional — eliminates the equity entirely. Note what has happened to the trade's prospects at that moment: the spread that was 10 basis points is now 90, so the expected profit on convergence has risen ninefold. The position is nine times more attractive and the trader is insolvent. That conjunction is the case in one sentence.

### The loss spiral, exactly

Losses do not merely reduce equity; they raise leverage, because the balance sheet does not shrink as fast as the capital supporting it. Let assets be $A_0 = L_0 E_0$, and suppose a mark-to-market loss of $\ell E_0$ — assets and equity fall by the same number of currency units, since the loss is a markdown of the assets themselves. Then

$$A_1 = (L_0 - \ell)E_0, \qquad E_1 = (1 - \ell)E_0, \qquad \boxed{L_1 = \frac{L_0 - \ell}{1 - \ell}}$$

With $L_0 = 25$ and a 12.5 per cent equity loss, $L_1 = 24.875/0.875 = 28.43{:}1$. The trader who does nothing is more levered after the loss than before it, so the *next* identical spread move costs more than the last. To restore the original leverage he must sell

$$A_1 - L_0 E_1 \;=\; (L_0 - \ell)E_0 - L_0(1 - \ell)E_0 \;=\; \ell\,(L_0 - 1)\,E_0,$$

which at $\ell = 0.125$ and $L_0 = 25$ is 3.00 units of assets per unit of *original* equity — 12.06 per cent of the surviving book — sold into the market that just moved against him. This is the loss spiral of Brunnermeier and Pedersen (2009) written out with numbers: the deleveraging required by a loss is proportional to the loss *multiplied by the leverage*, so a highly levered book must sell a large multiple of what it lost.

### What broke

Through 1998 a sovereign default and the flight to quality that followed widened, at the same time, spreads that the fund's risk model treated as weakly related. The correlation structure was estimated from a sample in which each spread had been driven mostly by its own local supply and demand. In the stress, all of them were driven by one thing: the price of taking risk with borrowed money. This is the general result of Chapter 14 — the funding factor is a common factor that does not appear as a column in a covariance matrix estimated on calm data — and it turns a portfolio of thirty diversified relative-value bets into one bet, thirty times over.

Three amplifiers converted that into failure. Repo financing that had been extended at very small haircuts on the strength of reputation was repriced, which is the margin spiral: a haircut rise mandates deleveraging regardless of the trader's own view (§15.3 works this arithmetic). The positions were not private: dealers who had financed the fund, or who had put on the same trades themselves, could infer what had to be sold, and a forced seller with a known position faces liquidity that is not merely absent but adversarial. And the fund's own size meant that its liquidation was a material fraction of the market in several of its spreads — it was, in the language of §15.1, a mechanical seller large enough to move the price that determined how much more it had to sell. The public record is that the fund's capital fell by roughly ninety per cent over a few months of 1998 and that a consortium of major dealers, convened under the auspices of the Federal Reserve Bank of New York, recapitalised it in September of that year; balance-sheet leverage before the episode is generally cited on the order of 25:1 (indicative). Many of the spreads subsequently converged.

### The lesson for a systematic book

*Leverage sets the horizon over which you are allowed to be right.* A forecast has a terminal distribution; a levered position has a *path*, and the path's maximum adverse excursion is a different random variable with a much fatter left tail than the terminal return. Expectancy arithmetic [Course Book §14.1] scores the terminal distribution; survival is scored by the path. This is precisely why the platform's evaluation pipeline reports the worst-decile path across combinatorial folds rather than the average path [Course Book §16.4], and why sizing runs off conditional value at risk rather than variance [Course Book §14.5]: both are attempts to make the path, not the mean, the object of the decision.

*A correlation assumption is a position.* Diversification across $n$ trades is worth $1/\sqrt{n}$ in volatility terms only if the correlations hold; when they go to one, the $n$-trade book has the risk of a single trade at $n$ times the size. A risk system that consumes an estimated correlation matrix is implicitly short an option on that matrix, and the option is struck exactly where the trader most needs it to be worthless. The practical consequence for a 21-instrument FX-led book is that the instrument count is not the diversification count: fourteen USD-legged pairs in a dollar-funding shock are close to one position, and the sizing machinery must be stressed under that assumption rather than the sample one.

*If your position is inferable, your liquidity is conditional.* A retail CFD trader is not large enough to be front-run individually, but he is a member of a population whose aggregate positioning is observable — that is exactly what Chapter 7's positioning data measures, and what a provider sees directly in its own book. Crowded retail positioning at obvious stop levels is a known feature of the FX landscape, and the honest conclusion is not paranoia about the broker but the same one as before: do not place size at a level everyone else has placed size at, and do not assume the fill.

## 15.3 2007–2008: the funding run

The crisis of 2007–2008 is treated here narrowly, as a funding event, because that is the part of it a currency trader needs. Brunnermeier's (2009) anatomy is the reference frame: a maturity- and liquidity-mismatched financing structure, a run on the short-term liabilities that supported it, and two amplification loops that turned a credit repricing into a market-wide liquidity failure. The credit losses that started the episode were, in aggregate, not large enough to explain what followed. The leverage they were sitting under was.

### The mechanics: what a modern run looks like

The classical bank run is depositors in a queue. The modern equivalent has no queue and no building. Two forms matter.

*Refusal to roll.* Long-maturity, hard-to-value assets were financed by short-maturity paper — asset-backed commercial paper issued by off-balance-sheet conduits, overnight and term repurchase agreements at dealers, and unsecured interbank borrowing. Each of these is a promise that must be renewed continuously. A lender who declines to renew is not making an accusation; declining costs him nothing and protects him against a small probability of a large loss. The structure therefore fails on the *suspicion* of a problem, not its confirmation.

*Haircut repricing.* A repurchase agreement lends against collateral with a haircut $h$: post collateral worth 1, receive $1 - h$ in cash. The borrower's own capital funds the haircut, so the maximum book a given equity can carry is

**Definition — Haircut and the leverage ceiling**

For a book financed entirely by repurchase agreements at haircut $h$ against equity $E$, the maximum asset position is

$$A_{\max} = \frac{E}{h}, \qquad \text{equivalently} \qquad L_{\max} = \frac{1}{h},$$

because each unit of assets requires $h$ units of own capital. A rise in $h$ is a mandatory reduction in $A$ that is entirely independent of the borrower's view, solvency, or forecast. Raising haircuts is therefore a system-wide margin call issued by the lending sector.

**Worked Example 15.2 — The haircut ceiling and the forced sale**

A repo-financed vehicle holds 200,000 million of high-grade securities against equity of 1,000 million, financed at a haircut of 0.50 per cent. Lenders reprice the haircut to 5.00 per cent. What must happen? (An illustrative scenario; haircut levels drift and must be verified against current market practice.)

*Before.* $L_{\max} = 1/0.005 = 200{:}1$, so $A_{\max} = 1{,}000/0.005 = 200{,}000$ million — the book is exactly at its ceiling.
*After.* $L_{\max} = 1/0.05 = 20{:}1$, so $A_{\max} = 1{,}000/0.05 = 20{,}000$ million.
*Consequence.* $200{,}000 - 20{,}000 = 180{,}000$ million, or 90 per cent of the book, must be sold. Equivalently, holding the original book at the new haircut would require $0.05 \times 200{,}000 = 10{,}000$ million of equity — ten times what the vehicle has.

*Reading.* No asset defaulted, no forecast changed, and the vehicle may be entirely solvent. A nine-and-a-half-fold move in a financing parameter, from a small number to a slightly less small number, mandated the liquidation of nine-tenths of the balance sheet. Multiply by every vehicle financed the same way and the sales are simultaneous and one-directional.

The two loops now compound. Forced sales depress prices; depressed and volatile prices raise

measured risk; higher measured risk raises haircuts and margins; higher haircuts force more sales. That is the *margin spiral*. Simultaneously, the price falls mark down the equity of every levered holder, mechanically raising leverage by the identity of §15.2 and requiring further sales to restore it. That is the *loss spiral*. Brunnermeier and Pedersen (2009), whose formal frame Chapter 14 sets out, show that the two reinforce: market liquidity and funding liquidity are mutually determined, so a shock to either propagates into the other, and the system has a stable regime in which liquidity is cheap and an unstable one in which it collapses. Exercise 15.2 works the cascade numerically.

A third mechanism, specific to a networked dealer system, sat underneath both. Bilateral derivative and financing exposures make every firm's solvency a function of others' solvency, and the exposures are not publicly observable. Under that opacity the individually rational response to uncertainty is to withdraw funding from everyone, which is why term unsecured interbank lending — normally priced within a few basis points of the expected policy path — repriced to a large and persistent spread over it. That spread is the cleanest public thermometer of the episode, and the structural reason it exists is worth remembering: an unsecured term loan to a bank is a credit-risky instrument, and in calm periods the market prices it as if it were not.

### Foreign exchange in 2008

For a currency book, 2008 produced two distinct and simultaneous flows, both funding-driven, and neither of them a forecast about relative growth.

*The carry unwind.* Chapter 5's carry trade — borrow a low-yield currency, hold a high-yield one — had been built up for years, funded predominantly in yen and Swiss francs. Deleveraging a carry position means buying back the funding currency. When many positions unwind together, that purchase moves the funding currency up, which *increases* the mark-to-market loss on every remaining carry position, which forces more unwinding. The positive feedback is in the exchange rate itself, and it is the reason carry returns exhibit the pronounced negative skew that Brunnermeier, Nagel and Pedersen (2008) document: the payoff is a slow accumulation of interest differential punctuated by crashes whose speed is set by deleveraging mechanics rather than by macroeconomics. Menkhoff, Sarno, Schmeling and Schrimpf (2012) supply the cross-sectional counterpart — carry returns load on global FX volatility, which is another way of saying they are short a funding-stress factor. Indicatively, the yen-funded high-yield crosses of the period, of which AUD/JPY is the canonical example, fell by roughly half from their 2007 peaks to their 2008 troughs; the reader should treat that magnitude as indicative of scale rather than as a datum.

*The dollar funding shortage.* Non-US banks held dollar assets financed by short-term dollar borrowing and by rolling FX swaps. When the dollar wholesale market closed, those institutions needed dollars they could no longer borrow, so they bid for them — and the dollar appreciated against nearly everything even though the crisis originated in the United States. The pressure showed up as a violation of the covered interest parity relation derived in Chapter 5: synthetic dollar borrowing through the FX swap market became persistently more expensive than direct dollar borrowing, which is the cross-currency basis. A no-arbitrage identity failed not because arbitrageurs stopped noticing but because executing it requires balance sheet, and balance sheet was the scarce good. Central bank swap lines were, in mechanical terms, an override of that scarcity. Rey's (2015) global financial cycle is the macro statement of the same fact: dollar funding conditions are a global state variable, not an American one.

### The lesson for a systematic book

*Funding liquidity is a hidden common factor.* It does not appear in the price series, it is not one of the platform's instruments, and it drives all of them at once when it moves. Its public proxies are indirect — the cross-currency basis, unsecured-to-secured spreads, the shape of the front end, dealer-sector balance-sheet indicators — and none is available to a retail platform at bar frequency. The practical response is not to pretend it is observed but to treat the *regime* it induces as the observable: volatility state, RORO composite, and funding-currency behaviour are symptoms

of the factor, which is why the Stage-2 router [Course Book §13.8] conditions on them, and why they must be constructed point-in-time [Course Book §18.2] or they will encode the crisis before the crisis.

*Funding currencies do exactly what Chapter 14 says.* A stress that forces deleveraging is a bid for the currency the leverage was borrowed in. The forecast content of that statement is limited — it says nothing about *when* — but the conditional content is strong: given a funding shock, the sign of the JPY and CHF response is not a coin flip. A distributional model whose regime conditioning is honest should be able to express that as a shift in the conditional density; a point forecast cannot express it at all.

*The tail is a different mechanism, not a bigger draw from the same one.* The carry crash is not the left tail of a stationary carry distribution. It is a distinct dynamic — forced buying of the funding leg — that switches on when constraints bind and is absent otherwise. Fitting a fat-tailed marginal to carry returns [Course Book §12.4] captures the *shape* of the consequence and none of the trigger. That is an argument for regime conditioning over distributional flexibility, and it is one the reader's own architecture already makes.

## 15.4 6 May 2010: the liquidity vacuum

The two previous cases took months. This one took minutes, and that is its contribution: it established how fast modern liquidity can leave, and it demonstrated that the depth displayed on a screen is not a stock of anything. The official account is CFTC–SEC (2010); the academic reconstruction from the audit trail is Kirilenko, Kyle, Samadi and Tuzun (2017).

### The mechanics: a controller in a feedback loop

The proximate trigger, on the official account, was a large sell programme in equity index futures executed by an algorithm instructed to participate at a fixed percentage of trailing market volume, with no price limit and no time limit. Consider what that instruction means in a market where volume is not exogenous. The algorithm sells; market makers buy and accumulate inventory; having exceeded their inventory tolerance, they sell to each other and to directional participants in order to flatten; that inter-dealer flattening *is volume*; the participation algorithm observes the higher volume and accelerates its selling. The controller's input variable is contaminated by its own output. This is a positive-feedback loop of exactly the 1987 form — sell more because the market moved — but constructed out of an execution parameter rather than an option replication, and running on a timescale of minutes rather than days.

Kirilenko et al. (2017) reconstruct what the intermediaries did. High-frequency market makers did not initiate the decline; they absorbed the early flow, as their business requires. But their inventory is mean-reverted over horizons of seconds to minutes, and mean-reverting a long inventory in a falling market means selling into it. The finding that matters is that over the critical interval the fast intermediaries were net sellers in the direction of the move, and their aggregate net position barely changed — they were passing inventory between themselves, not warehousing it. Menkveld's (2013) characterisation of the modern market maker generalises the point: these are inventory-constrained intermediaries operating on thin capital and fast cancellation, not committed liquidity providers with a stabilising obligation. Their business model is the round trip, not the position.

### What broke

*Displayed depth vanished, and it was always allowed to.* A resting limit order is, as Harris (2003) puts it, a free option written to the rest of the market: it grants everyone else the right, but not the obligation, to trade at the quoter's price, and the quoter is compensated only by the spread he earns when the counterparty is uninformed. The value of that option to its holders rises with uncertainty. So the moment uncertainty spikes, the rational quoter cancels — not out of panic but out of arithmetic. It follows that displayed depth is systematically *largest when it is least needed*

*and smallest when it is most needed*, and that no measure of average book depth carries information about depth in the event.

*Prices with no counterparty on the other side were transacted against.* With genuine quotes withdrawn, market orders reached nominal quotes posted only to satisfy an obligation to be present — so-called stub quotes, at prices bearing no relation to value. Trades printed at absurd levels in individual securities, and the public record is that a large number of them, those executed beyond a defined distance from pre-event prices, were subsequently cancelled by the venues.

*The cross-market link failed in the direction that transmits, not the direction that repairs.* The single "price" of an equity index exists because arbitrageurs continuously equalise the future, the exchange-traded fund, and the constituent basket. Those arbitrageurs withdrew for the same reason the quoters did, so the mechanism that normally *imports* stability from the deepest market instead exported the dislocation and then failed to close it.

The whole episode — decline and substantial recovery — occupied roughly half an hour of the trading day. The structural legacy is again visible in today's instruments: limit-up/limit-down price bands, revised single-stock halts, and rules against nominal stub quoting followed (structural; specifics drift and must be verified against current rules).

**Trap — A stop order is a market order wearing a costume**

A stop is not a price at which you will trade. It is an instruction that, once a trigger price prints, converts your resting order into a *market* order, to be filled against whatever happens to exist. In a calm book that distinction is worth a fraction of a pip and is correctly ignored. In a vacuum it is worth everything, because the thing being relied upon — a counterparty near the trigger price — is exactly what disappears in the event that fires the stop. Every position-sizing formula in the Course Book takes "risk per trade" as an input [Course Book §14.5]; that input is the *intended* loss, and the intended loss equals the realised loss only when the market chooses to cooperate. A stop-limit avoids the bad fill by substituting a worse failure — no fill at all, and an unbounded open position. There is no order type that solves this. The only quantity the trader actually controls is size.

### The lesson for a systematic book

*The cost surface is a distribution, and its right tail is where the strategy dies.* The platform's programme of replacing mid-only candles with a measured bid/ask cost surface is the correct response to this case, but only if the surface is stored and used as a conditional distribution rather than a mean. The arithmetic is unforgiving. Suppose a strategy's assumed cost is 1.5 pips per trade, its per-trade edge is 3.0 pips, and in a stressed window the effective cost is 40 pips. One stressed fill costs 38.5 pips of excess, which is 12.83 trades' worth of edge. If a fraction $\pi$ of fills land in the stressed regime, the mean cost becomes $(1-\pi)\times 1.5 + \pi \times 40$, so at $\pi = 1$ per cent the average cost rises 25.7 per cent and the net edge falls to 2.615 pips; at $\pi = 2$ per cent the net edge is 2.230 pips; and the share that erases the edge entirely is $\pi^* = 3.0/(40-1.5) = 7.79$ per cent. A strategy evaluated on average spreads can be profitable in the backtest and unprofitable in life on the strength of a stressed-fill share of under eight per cent — and stressed fills are not randomly distributed across the strategy's trades, they cluster exactly where its signals fire.

*Cost is regime-dependent, so the regime router has a cost job as well as a signal job.* Chapter 3's decomposition and the platform's history of high hit rates with negative expectancy [Course Book §14.1] both point the same way: for this book the cost surface is not a nuisance parameter, it is the dominant term. Conditioning it on the observable state — news windows, rollover, thin-hours sessions, volatility regime — is the same modelling exercise as conditioning the return distribution, and deserves the same rigour [Course Book §14.2].

*A CFD quote is derived, and the provider's response to a vacuum is to widen or to stop.* There is no consolidated book behind a retail FX CFD (Chapter 4): the provider quotes a two-sided price it is

willing to trade, informed by the underlying interbank market and its own inventory. When the underlying market thins, the provider's hedging cost rises and the quote widens or is suspended — and its terms will typically reserve the right to cancel or adjust trades executed at manifestly erroneous prices, which means a profit taken during a dislocation is not necessarily final. This is structural and must be read in the reader's own contract; the modelling consequence is that execution realism [Course Book §18.3] in event windows should assume the worse of a wide fill and no fill, never the mid.

## 15.5 15 January 2015: the peg that broke

This is the case that matters most to the reader, because it is the only one in the chapter that destroyed retail traders in the reader's own instrument class, on the reader's own kind of account, using the reader's own leverage. It is also the case where every defence a retail trader believes he has — the stop, the margin close-out, the broker's risk system, the regulator — failed simultaneously and for the same reason. It deserves to be worked in detail.

### The mechanics of a floor

In September 2011, with the franc appreciating sharply as European sovereign stress pushed capital into it, the Swiss National Bank announced a minimum exchange rate of CHF 1.20 per euro and stated that it would enforce it by purchasing foreign currency in unlimited quantities. The mechanics deserve care, because a floor of this kind is structurally different from the currency defences that fill the intervention literature (Sarno and Taylor, 2001).

A central bank defending its currency *against depreciation* must sell foreign reserves to buy its own money. Reserves are finite, so the defence has a computable exhaustion point, and speculators attacking it are effectively racing a countdown. The SNB's problem was the mirror image: it was preventing *appreciation*, which requires selling francs and buying euros, and it can create francs without limit. In the narrow operational sense the floor was therefore unbreakable by market forces — there is no run that exhausts a printing press. The binding constraint was never capability; it was willingness. Every franc created to hold the floor bought a euro asset, so the central bank's balance sheet grew, and grew with a currency mismatch: euro-denominated assets against franc-denominated liabilities, on which any subsequent franc appreciation produces a valuation loss. By late 2014 the SNB's balance sheet had grown to a very large fraction of Swiss annual output (indicative), and expectations of large-scale euro-area asset purchases implied that maintaining the floor would require substantially more of the same. That is the shape of every peg's ending: the policy remains possible long after it has stopped being attractive, so the exit is a *decision*, and decisions are not forecastable from the price series.

### What the floor did to the observable statistics

While the floor held, EUR/CHF traded in a narrow band just above 1.20. For a systematic trader this produced an instrument with an extraordinary profile: a positive interest differential (euro rates above Swiss rates, so a long EUR/CHF position was paid to wait), a realised volatility close to zero, and — most seductively — an apparent downside underwritten by a central bank. That combination is the exact signature of a written option, and that is what it was.

**Trap — An administered price has administered statistics**

A pinned exchange rate's realised volatility measures the *administration*, not the economics. Take an illustrative pinned series with a daily standard deviation of 0.05 per cent against a free-floating major at 0.57 per cent: annualised, that is 0.79 per cent against 9.05 per cent, a factor of 11.4. Now feed both into a volatility-targeting sizer [Course Book §14.4] at a 10 per cent risk target. It scales notional by target/$\hat{\sigma}$, so it allocates 12.60 times notional to the

pegged instrument and 1.11 times to the floating one — the peg buys the position 11.4 times more size, and does so *because* its risk has been suppressed. Through a 15 per cent gap the pegged-sized position loses 1.89 times equity; the floating-sized position loses 0.17 times equity. Nor does any tail-aware estimator rescue this: measured against the pegged sample's own daily standard deviation, a 15 per cent gap is a 300-sigma event, and against the floating sample's it is still 26-sigma. No fitted distribution — Gaussian, Student-$t$, or extreme-value — assigns usable probability to either. The suppressed variance was not destroyed. It was stored, and the storage was uninsured.

The general statement is worth isolating: *a peg is a free option written by the authority to the market, with an unrestricted cancellation clause*. Holders of a long EUR/CHF position were long a put struck at 1.20 for which they paid nothing, issued by a writer under no contractual obligation, and revocable without notice. Any position whose risk profile depends on an administered price — pegged and managed currencies, price-capped commodities, supported bond markets — carries the same structure. The risk is not in the sample; it is in the writer's incentives, which are political and are not in any data feed the platform consumes.

### The announcement and the gap

On 15 January 2015 the SNB announced that it was discontinuing the minimum exchange rate with immediate effect, simultaneously lowering the interest rate on sight deposit balances to $-0.75$ per cent (SNB, 2015). There was no notice and no phased withdrawal — correctly so, from the central bank's point of view, since a pre-announced exit invites the entire market to transact ahead of it.

What happened next is the part every systematic trader should internalise. EUR/CHF did not *fall* through a sequence of prices. It *gapped*: liquidity was withdrawn essentially instantly and simultaneously across venues, and for a period there were no executable two-sided prices at all between the floor and levels tens of per cent below it. The public record shows prints far below parity on some venues before a partial recovery, with the pair settling roughly 15 to 20 per cent below the floor by the close (indicative magnitudes). The essential fact is not the size of the move but its *topology*: the path from 1.20 to the lows was not traversed, it was skipped. A stop order resting at 1.19 was not filled at 1.19, was not filled at 1.15, and was not filled at 1.10, because none of those prices existed as a trade. It was filled wherever the market reopened.

**Worked Example 15.3 — A 30:1 EUR/CHF long through the gap**

A hypothetical retail account holds equity of CHF 10,000 and is long EUR/CHF at the provider's maximum permitted leverage of 30:1, entered at $S_0 = 1.2000$, with a stop-loss resting one per cent below the entry — placed there in the belief that the floor made 1.1880 unreachable. The market gaps and the position is closed at $S_{\text{fill}} = 1.0200$, 15 per cent below entry. Compute the outcome. (Hypothetical scenario; leverage caps, margin rates and close-out rules drift by jurisdiction and provider — verify current terms.)

*Position.* Notional $= 30 \times 10{,}000 =$ CHF 300,000, which at 1.2000 is EUR 250,000. The margin rate is $1/30 = 3.3333$ per cent, so the account is fully committed.
*Intended risk.* Stop at $1.2000 \times (1 - 0.01) = 1.1880$. Profit and loss in the quote currency is EUR notional $\times (S_{\text{exit}} - S_0)$, so

$$250{,}000 \times (1.1880 - 1.2000) = -\text{CHF } 3{,}000$$

— 30 per cent of equity, leaving CHF 7,000. Note this already: at 30:1, a stop placed a mere one per cent away risks nearly a third of the account.

*Realised outcome.* Fill at $1.2000 \times (1 - 0.15) = 1.0200$:

$$250{,}000 \times (1.0200 - 1.2000) = -\text{CHF } 45{,}000,$$

which is 4.50 times the account equity. The closing balance is $10{,}000 - 45{,}000 = -\text{CHF } 35{,}000$: a *negative balance* equal to 3.50 times the money originally deposited. The realised loss is $45{,}000/3{,}000 = 15.00$ times the intended loss.
*Why no intermediate defence fired.* The price at which this account's equity reaches exactly zero is $1.2000 \times (1 - 1/30) = 1.1600$, and a close-out trigger set at 50 per cent of required margin sits at 1.1800. Both are *inside* the gap. The provider's risk engine was not slow; there was nothing at 1.18 or 1.16 to trade against. The gap cleared the zero-equity price by a further 11.67 percentage points of the entry level.
*The account-currency multiplier.* Suppose instead the account is denominated in Australian dollars, with AUD/CHF at 0.7500 francs per dollar before the announcement and 0.6000 after — the franc appreciated against everything, not only the euro. Initial equity is AUD 13,333.33. The CHF 45,000 loss converts at the *post-event* rate: $45{,}000/0.6000 = \text{AUD } 75{,}000$, against AUD 60,000 at the pre-event rate — 25 per cent more. Closing balance: AUD $-61{,}666.67$. In account-currency terms the loss is 5.625 times equity rather than 4.500, and the negative balance 4.625 times rather than 3.500. The ratio is exactly $0.7500/0.6000 = 1.25$.
*Reading.* The account-currency conversion is not a harmless rescaling. When a position loses money in a currency that is appreciating for the same reason the position lost money, the conversion is *wrong-way* and amplifies the loss. Any risk system that computes exposures in quote currency and converts at a static rate understates the tail of every cross whose quote currency is a haven.

### The sizing bound that actually binds

The worked example makes the governing arithmetic obvious once written down.

**Result — The gap-survivable position limit**

A position of leverage $L$ (notional divided by account equity) suffers a fractional equity loss of $L\,g$ through a price gap of size $g$, regardless of where any stop was placed. Hence

- to avoid a negative balance at all requires $L < 1/g$;
- to lose at most a fraction $\theta$ of equity requires $L \leq \theta/g$.

At $g = 15$ per cent: avoiding insolvency requires $L < 6.67$:1, and capping the loss at a quarter of equity requires $L \leq 1.67$:1 — which is 5.56 per cent of the 30:1 the account in Worked Example 15.3 was permitted to run. The binding constraint on a levered CFD book is not the margin rule. It is the gap the trader is willing to survive, and it is between one and two orders of magnitude tighter than what the provider allows.

The bound has no free parameters other than $g$ and $\theta$, both of which are choices rather than estimates. That is its virtue: unlike a value-at-risk number it does not depend on a fitted distribution, and it therefore does not evaporate when the distribution does. The uncomfortable consequence is that a book sized to survive a 15 per cent gap on any single instrument is running leverage a retail platform would consider trivially small, and that conclusion does not become false because it is unwelcome.

### Guaranteed stops: insurance, honestly priced

The market's answer to gap risk is the guaranteed stop-loss order: for a premium, the provider contracts to close the position at the stop level regardless of gaps. Structurally it is insurance —

the provider writes the gap risk and charges for it — and it can be assessed as insurance.

**Worked Example 15.4 — What a guaranteed stop costs and what it prices**

A provider charges a premium of $c = 0.30$ per cent of notional when a guaranteed stop is triggered (indicative structure and level; premiums, minimum stop distances and charging conventions vary by provider and drift over time — verify current terms). Consider the protection it would have supplied in Worked Example 15.3: the stop was one per cent away, the fill was 15 per cent away, so the shortfall the guarantee absorbs is $g = 14$ per cent of notional.

*The break-even jump probability.* The provider breaks even if the premium equals the expected payout, $c = \lambda g$ where $\lambda$ is the per-trade probability of a gap of that severity:

$$\lambda^* = \frac{c}{g} = \frac{0.0030}{0.14} = 0.021429 = 2.1429\% \text{ per protected trade,}$$

one depeg-scale gap every 46.67 protected trades. At 250 protected trades a year, that premium prices 5.36 depeg-scale gaps *per year*.

*Against a plausible frequency.* If the true probability of such an event is one trade in 2,000, the actuarially fair premium is $0.0005 \times 0.14 = 0.70$ basis points against 30 basis points charged — an overpayment of 42.86 times. At one in 500 it is 2.80 against 30, or 10.71 times; even at an implausible one in 100 it is 14.00 against 30, or 2.14 times.

*The drag on the strategy.* Suppose the premium is charged on every protected order rather than only on triggers, at 250 trades a year: $250 \times 0.30\% = 75$ per cent of turnover notional annually, which at 5:1 leverage is 375 per cent of equity and at 30:1 is 2,250 per cent. Charged only on triggers, with a 40 per cent stop-hit rate, the drag is $250 \times 0.40 \times 0.30\% = 30$ per cent of turnover notional — still 150 per cent of equity a year at 5:1. Against a per-trade edge of 2 basis points of notional, which over 250 trades is 5.00 per cent of notional a year, the trigger-only premium is 6.00 times the entire gross edge.

Two conclusions follow, and they point in opposite directions, which is why the honest answer is uncomfortable.

The expected-value comparison says the guarantee is grossly overpriced, and for the provider's book that comparison is correct: it writes many such contracts and diversifies across them. But the client does not. He writes one contract and cannot diversify across states of the world, and the loss being insured is 4.50 times his equity — a quantity for which expected-value arithmetic is simply the wrong tool, because ruin truncates every future period's compounding [Course Book §14.3]. Insurance against ruin is rationally purchased at prices far above fair value.

Yet the drag arithmetic then closes the door anyway. A high-turnover systematic book cannot pay tens of basis points of notional per trade, because that number is an order of magnitude larger than the edge the whole programme is trying to capture [Course Book §1.6]. The guaranteed stop is affordable for a discretionary trader placing a handful of large, long-held positions and unaffordable for the platform's intended operating mode. Which leaves exactly one instrument for managing gap risk on a high-turnover book: *size*, as bounded by Result 15.5. That is not a satisfying answer. It is the correct one.

There is also a structural caveat that no premium comparison captures. A guaranteed stop is a *contractual promise from the provider*, and the event it insures against is precisely the event that impairs the provider. The guarantee's value is conditional on the counterparty surviving the same gap that triggers it. It is wrong-way exposure in the exact technical sense.

### What broke: brokers, and the meaning of a negative balance

Retail CFD and margin-FX providers are not merely conduits. They extend leverage against client collateral and rely on being able to close a losing client out before his equity is exhausted. That model rests on one assumption — *price continuity* — and 15 January 2015 removed it. Across the industry, clients' positions were closed at prices far beyond the point at which their collateral was consumed, producing client balances that were not merely zero but negative. Those negative balances are the provider's loss unless collected, and collecting a debt many multiples of a retail client's deposit is slow, jurisdictionally variable, and frequently futile. The public record is that at least one UK-based retail broker entered insolvency proceedings within days, that another required emergency financing to continue operating, and that clients in several jurisdictions were pursued for, or forgiven, balances they never agreed to owe. In the years that followed, retail leverage caps and negative-balance protection requirements were introduced in a number of jurisdictions — structural, but the specifics vary and the reader must verify what applies to his own account, entity and jurisdiction rather than assume.

Negative-balance protection is worth understanding precisely, because it is easy to misread as a solution. It converts the client's unbounded downside into a bounded one — he can lose his deposit, not more — by transferring the excess to the provider. It therefore does not remove the loss; it *concentrates* it on the counterparty that is holding the other side of every client's position and has just absorbed the same gap on its own hedge book. A CFD is a bilateral contract: an open profit is an unsecured claim on the provider, and client money held in segregated accounts is protected by a legal structure whose effectiveness in an insolvency is a question of law and of what actually got segregated, not a market fact. This is the chapter's plainest instruction: counterparty risk in retail CFD trading is real, it is correlated with exactly the events that make the account most valuable, and the only defences are keeping surplus capital off the platform and, where the size of the book justifies it, spreading it across more than one provider.

### The lesson for a systematic book

*Pegs transfer tail risk; they do not remove it.* Every statistic computed on an administered price describes the administration. A model fitted to that sample — volatility, quantiles, conformal intervals, regime states — has learned the policy, not the market, and will be confidently wrong the moment the policy changes. The general rule is that any instrument whose price is being managed should be excluded from leverage-scaled sizing rules altogether, or sized on the *unpegged* counterfactual, because there is no estimator that recovers the suppressed risk from the pegged data.

*Gap risk defeats per-trade stops, so risk must be expressed as size.* A stop is a statement about the intended loss; a gap is a statement about the realised one; and Result 15.5 converts the difference into a leverage ceiling. The platform's sizing machinery should carry that ceiling as a hard constraint alongside its conditional value-at-risk budget [Course Book §14.5] — not because CVaR is wrong, but because CVaR is estimated from a sample and the ceiling is not.

*The event is not in the price history, so it must be injected.* Nothing in EUR/CHF's 2011–2014 bars contains the information that the floor could be withdrawn. Scenario tooling that stresses a book with imposed gaps of specified size — five, ten, twenty per cent, on the instruments and correlations most exposed — is the only way this risk enters the platform's evaluation at all, and its outputs should be read as constraints rather than as forecasts [Course Book §23.5].

*Backtest bar data hides the topology.* A daily bar with an open of 1.2010 and a low of 1.0200 is indistinguishable, in a naive simulator, from a day that traded smoothly through every intervening price. Any fill logic that assumes intra-bar prices were available — limit fills inside the bar, stop fills at the trigger, scaling in on the way down — will manufacture profits that did not exist and hide losses that did [Course Book §18.3]. The correct treatment of a gap bar is that only the open and the close were tradeable, and even that is generous.

## 15.6 March 2020 and the dash for cash

March 2020 is the modern reference episode for what a liquidity event does to a diversified book, and it contains a bonus: a price that the industry's models were not written to represent. The BIS (2020) bulletins on the turmoil are the reference frame used here.

### The mechanics: two phases, and why the second one surprises people

The first phase was conventional and would have satisfied Chapter 14's correlation map: equities fell, credit spreads widened, the yen and the franc and gold firmed, and the dollar was mixed. Had it stopped there, diversification would have worked as advertised.

The second phase inverted it. Over roughly a fortnight, *everything* was sold, including government bonds, gold, and the highest-quality short-dated paper — assets whose entire investment case is that they rise when other things fall. The mechanism is not mysterious once stated, and it has three legs.

*Cash demand was the binding objective, not risk reduction.* Corporates drew credit lines preemptively; funds facing redemptions had to raise cash; every levered position facing a variation-margin call had to settle it in cash the same day. None of those actors was expressing a view. They needed currency, and they needed it now.

*What gets sold is what* can *be sold.* An agent raising cash under a deadline sells the position that will transact at a knowable price, not the position he likes least. That ordering is by liquidity, so the most liquid, highest-quality assets are hit *first and hardest* — exactly the inversion of the intuition that quality protects. Chapter 14's account of gold falling in the first week of a crash and rallying afterwards is this mechanism at the instrument level, and March 2020 ran it across the whole cross-section.

*The shock absorber was itself constrained.* Dealer balance sheets are what stand between a wave of selling and a price collapse, and intermediation capacity is finite and priced (Duffie, 2012). When the wave includes the government bond market — the asset dealers use to hedge everything else — the absorber saturates. Bid-ask spreads in instruments that normally trade in fractions of a basis point widened by an order of magnitude (indicative), and levered relative-value positions in that same market were being unwound under rising margin, adding supply to the asset everyone was supposed to be buying.

Two amplifiers deserve naming because they recur. Initial margin at clearing houses and under bilateral agreements is typically set from recent realised volatility, so requirements rise *after* a large move, mechanically demanding cash from exactly the participants who have just lost it — the margin spiral of §15.3, now hard-wired into the market's risk infrastructure rather than negotiated bilaterally. And the retail layer moved the same way: providers widened spreads to multiples of their normal levels and raised margin requirements mid-event. That last point converts directly into the reader's arithmetic. A provider raising the FX margin rate from 3.33 per cent (30:1) to 20 per cent (5:1) multiplies the equity required to hold an unchanged book by 6.00; an account at the old limit must liquidate irrespective of its view, at the widened spread, in the middle of the event. Even a move to a 5 per cent requirement forces a 33.33 per cent reduction in position size.

The reversal was as abrupt as the collapse and had the same character: it was a policy event, not a change of view. Large-scale asset purchases, dollar swap lines and a series of lending facilities restored the funding side, and prices retraced. For a forecaster the honest observation is that both the collapse and the reversal were driven by decisions — margin policy, central bank policy — that are not encoded in the price history.

### Negative front-month oil, 20 April 2020

Chapter 12 established the delivery mechanics; this is the case where they bit. A physically deliverable crude contract obliges the holder at expiry to take barrels at a specified location. In

April 2020 demand collapse had filled storage; the marginal barrel had nowhere to go. A long who could not take delivery had to exit before expiry, and the buyers who could absorb the contract were those with storage — of whom there were effectively none left. The publicly reported outcome is that the expiring front-month contract settled at roughly −$37.6 per barrel, having settled near $18.3 the previous session.

The economics are not exotic. Working's (1949) storage relation, $F = Se^{(r+u-y)T}$ with storage cost $u$ and convenience yield $y$, is written as though $u$ were a finite rate. When physical capacity binds, the cost of storing the marginal barrel is not a rate — it is unbounded, because the storage does not exist at any price. And the reason we casually believe prices cannot go negative is *free disposal*: an asset you can costlessly throw away is worth at least zero. Crude oil has no free disposal — you may not legally dump it — so the lower bound on its price is not zero but minus the cost of lawfully getting rid of it. The negative print was not a market failure in the sense of the other cases in this chapter. It was a correct price for an obligation nobody wanted, revealed when the assumption that quietly supports the zero bound stopped holding.

What broke was software and models, and that is the part a systematic trader must take personally.

**Your project — Domain errors are a live failure mode on this platform**

The platform trades WTI as one of its 21 instruments, and the Course Book's entire return convention is logarithmic [Course Book §1.1]. Consider the transition from the publicly reported +18.27 to −37.63 per barrel. The simple return is $(-37.63 - 18.27)/18.27 = -305.97$ per cent — computable, but outside the $[-100\%, \infty)$ range every downstream component assumes. The log return $\ln(P_t/P_{t-1})$ is simply *undefined*: the argument is negative. There are only three things a pipeline can do with that, and two of them are catastrophic. It can raise and halt — inconvenient, and by far the best outcome. It can propagate a not-a-number that silently poisons every rolling statistic downstream. Or it can drop the observation, which is worst of all, because the discarded row is the single most informative bar in the sample and its removal is invisible in every summary statistic. The same applies to a shifted transform: $\ln((P_t + \alpha)/(P_{t-1} + \alpha))$ with $\alpha = 50$ gives −1.708196, a number that is finite, defensible, and quietly changes the meaning of every coefficient fitted on it.

The remedies are unglamorous and belong in Stage 0 [Course Book §14.7]: assert the domain explicitly at ingest rather than relying on the transform to fail loudly; log and *retain* non-positive prints as flagged observations instead of dropping them; test the pipeline with a synthetic negative-price bar as part of the leak-and-integrity test suite [Course Book §18.2]; and record which instruments are physically deliverable, because that is the property that makes the domain assumption falsifiable. The wider point generalises past oil: options models built on a lognormal forward (Black, 1976) are undefined for negative underlyings, and the industry's response in 2020 — shifting to displaced-lognormal or normal dynamics — is the same class of fix. A model has a domain, the domain is an assumption, and assumptions do not appear in risk reports.

### The lesson for a systematic book

*Correlation-to-one is the tail's normal state, and its cause is the sale ordering.* Because a cash-raising seller sells what he can rather than what he wants to, the cross-section of stress selling is ordered by liquidity. Diversification computed on fundamental distinctions — asset class, geography, factor — therefore fails on precisely the axis it was built on, while the surviving structure is a single dimension: how quickly the thing converts into currency. For a 21-instrument book that is a concrete instruction. The sizing machinery should be stressed under an assumed stress correlation, not the estimated one, and the stressed matrix should group instruments by their funding and liquidity characteristics rather than by their labels.

*Regime shifts in the cost and margin surfaces are discontinuous and endogenous.* Spreads did not drift wider; they stepped. Margin requirements did not tighten gradually; they were reset. Both were set by counterparties responding to the same event the strategy was trading. A simulator with a constant cost assumption and a constant margin assumption is not a conservative approximation of this — it is a simulation of a different market, and the difference is largest exactly in the periods that determine the strategy's tail.

*Policy is a state variable the price history cannot teach.* Both the collapse and the reversal in March 2020, like the SNB's decision in 2015, were discrete choices by institutions outside the market. A model can condition on the *consequences* of such decisions once they appear — volatility state, funding proxies, spread regime — and cannot anticipate the decisions. That asymmetry is the honest limit of the platform's regime machinery, and it should be stated in the model card rather than discovered in production.

## 15.7 The failure-mode taxonomy

Six episodes, six mechanisms. They are not six instances of one thing — they differ in what initiates them, how fast they run, and which part of a trader's cost and risk structure they attack — and the practical value of the chapter lies in keeping them distinct, because the defences differ too. Tables 15.1 and 15.2 set out the taxonomy: the first classifies the mechanism and its timescale, the second records what each mode does to the four quantities a levered CFD book actually cares about.

| Failure mode | Mechanism | Characteristic speed | Canonical case |
|---|---|---|---|
| Feedback crowding | A rule shared by many participants instructs selling into weakness; the aggregate demand curve slopes the wrong way and steepens as the price falls | Hours to days | October 1987 |
| Leverage unwind | Thin-edge levered positions are marked down; leverage rises mechanically with losses; required deleveraging scales as loss × leverage | Weeks to months | LTCM, 1998 |
| Funding run | Short-term financing is not rolled, or is repriced through haircuts; balance-sheet contraction is mandatory and independent of the borrower's view | Days to months | 2007–2008 |
| Liquidity withdrawal | Quoters cancel because a resting quote is a free option whose value to takers rises with uncertainty; displayed depth was never a stock of anything | Seconds to minutes | 6 May 2010 |
| Peg break | An administered price is a free option written by an authority with an unrestricted cancellation clause; withdrawal removes the price path, not merely the level | Instantaneous | 15 January 2015 |
| Dash for cash | Cash demand under deadline orders selling by liquidity rather than preference; the dealer absorber saturates; margin requirements rise procyclically | Days to weeks | March 2020 |

**Table 15.1.** The six failure modes by mechanism and timescale. The speeds are the distinguishing feature as much as the mechanisms: a defence that requires a human decision is useless against the bottom two rows.

Two structural observations fall out of the tables. First, the speed column and the gap column are related: the faster the mode, the more of the price move arrives as discontinuity rather than as a traversable path, because there is less time for any quoter to re-enter between trades. Second, the margin column contains the only channel in the table that acts on a position *regardless of the trader's view or intent* — a haircut or margin increase is a liquidation order issued by a third party. Everything else in the table can, in principle, be waited out by an unlevered holder. That is the sense in which leverage is not a magnifier of risk but a change in its kind.

| Mode | Spreads and depth | Gaps | Margin and financing | Correlation |
|---|---|---|---|---|
| Feedback crowding | Widen materially; depth becomes one-sided | At opens and around halts | Raised after the event; venues impose breakers | Rises within the crowded asset class |
| Leverage unwind | Widen in the specific instruments the levered holders own | On liquidation prints and at each mark | Haircuts repriced by financiers; the dominant channel | Goes to one across a book via its common holder |
| Funding run | Widen broadly and *persistently*; the level does not revert for months | At each financing repricing | Haircuts step up repeatedly; the defining channel | Goes to one via the funding factor, across asset classes |
| Liquidity withdrawal | Explode transiently; book empties to nominal quotes | To stub levels, then recover | Unchanged — the event is faster than any margin reset | Rises transiently as cross-market arbitrage withdraws |
| Peg break | Quoting ceases; there is no spread because there is no market | *Is* the event; unbounded and untraversed | Reset afterwards: leverage caps cut, requirements raised | The haven complex moves as one; account-currency conversion is wrong-way |
| Dash for cash | Step up across everything, including the safest instruments | At session boundaries and policy announcements | Rise procyclically at clearing houses and at retail providers, mid-event | Goes to one across the whole cross-section, ordered by liquidity |

**Table 15.2.** What each failure mode does to the four quantities that determine a levered CFD book's survival. Read the "gaps" column first: it is the only one against which no order type helps.

**Your project — The platform's defences, mode by mode — and the residue**

Mapping each mode onto the machinery already specified in the Course Book makes clear what is covered, what is merely observed, and what is not addressed at all.

*Feedback crowding.* No observable warns of it, because the driving variable — how much other capital follows the same rule — is absent from the price series. The defence is epistemic: the deflated Sharpe and backtest-overfitting machinery [Course Book §15.4], [Course Book §15.5] stops a crowded backtest from earning belief, and the stressed cost surface prices the exit that crowding makes expensive. Honest status: *not detected, only discounted*.

*Leverage unwind.* Directly addressed. Conditional value-at-risk sizing [Course Book §14.5], fractional Kelly [Course Book §14.3] and volatility targeting [Course Book §14.4] all cap the loss-to-equity ratio that drives the spiral, and Result 15.5's ceiling binds above them. Honest status: *covered, conditional on the sizing inputs being stressed rather than sample-estimated*.

*Funding run.* The Stage-2 regime router [Course Book §13.8] conditions on RORO and volatility state, which are symptoms of the funding factor rather than the factor itself, and the covariates must be point-in-time [Course Book §18.2] or they encode the outcome. Change-point detection [Course Book §13.7] is by construction *lagging*: it needs observations after the change to declare one. Honest status: *detected late, by design*.

*Liquidity withdrawal.* This is what the measured bid/ask cost surface is for, and it only works if the surface is stored as a conditional distribution and the strategy is evaluated on a stressed decile rather than a mean. Execution realism [Course Book §18.3] must assume the worse of a wide fill and no fill in event windows, and the kill gates [Course Book App. E] should compare realised to modelled cost continuously, not at review time. Honest status: *covered in cost terms, provided the surface is never averaged*.

*Peg break.* Nothing statistical helps. The defences are three, all structural: the gap-survivable

leverage ceiling of Result 15.5 as a hard constraint; instrument screening that refuses leverage-scaled sizing on any administered price; and injected gap scenarios read as constraints rather than forecasts [Course Book §23.5]. Honest status: *unforecastable; survivable only by size*.

*Dash for cash.* Sizing must run on a stressed correlation matrix grouped by funding and liquidity characteristics rather than instrument labels, and evaluation must report the worst-decile path across combinatorial folds [Course Book §16.4], [Course Book §16.6] rather than the average. Honest status: *covered in the risk model, not in the forecast*.

*And what nothing defends against.* Three items belong on the ledger with no offsetting entry. **Broker failure**: a CFD is a bilateral contract, an open profit is an unsecured claim on the provider, and the provider is impaired by exactly the events that make the claim large. No model addresses this. The only responses are keeping surplus capital off the platform, diversifying counterparty where the book's size justifies the operational cost, and reading the segregation terms as a legal question rather than a market one. **A gap through everything**: no order type, no stop, no volatility estimate and no conformal interval survives a discontinuity, and the only control is position size, which is why Result 15.5 is stated as an absolute ceiling rather than as a model output. **Policy decisions**: the SNB's exit, the March 2020 facilities and every margin-rule change were choices made outside the market, unforecastable in timing, and observable only through their consequences. The machinery reduces the probability and the severity of ruin. It does not remove it, and a system that claims otherwise has hidden the assumption rather than retired it.

## 15.8 Exercises

**Exercise 15.1 — Gap arithmetic and the price of a guarantee**

A hypothetical account is denominated in Swiss francs with equity of CHF 24,000 and holds a long EUR/CHF position of EUR 400,000 entered at $S_0 = 1.2000$. A stop-loss rests 0.80 per cent below the entry. The pair gaps and the position is closed 18.0 per cent below the entry.

(a) Compute the position's franc notional and its leverage on equity; the intended loss at the stop, in francs and as a percentage of equity; the realised loss and the closing balance; and the ratio of realised to intended loss.

(b) What is the largest leverage this account could have run and still have lost no more than 40 per cent of its equity through an 18 per cent gap? Express the answer as a leverage ratio, a franc notional, and a euro position size, and state what fraction of the position actually held that represents.

(c) The provider offers a guaranteed stop-loss order at a premium of 0.25 per cent of notional, charged only when the guaranteed stop is triggered (indicative structure; verify current terms). Taking the shortfall the guarantee would have absorbed from part (a), compute the per-trade jump probability at which the provider breaks even. Then, for a strategy placing 180 such trades a year of which 40 per cent trigger their stop, compute the annual premium in francs and as a percentage of equity, compare it with the single gap loss from part (a), and derive the trade count $n^*$ at which the annual premium exactly equals one gap loss.

▷ working space

**Solution — Exercise 15.1**

**(a)** Notional $= 400{,}000 \times 1.2000 =$ **CHF 480,000**, so leverage $= 480{,}000/24{,}000 =$ **20.00 : 1**. Stop level $= 1.2000 \times (1 - 0.008) = 1.19040$. Profit and loss in the quote currency is the euro notional times the price change:

$$400{,}000 \times (1.19040 - 1.2000) = -\textbf{CHF 3,840.00} = -16.00\% \text{ of equity.}$$

Fill level $= 1.2000 \times (1 - 0.18) = 0.98400$:

$$400{,}000 \times (0.98400 - 1.2000) = -\textbf{CHF 86,400.00} = 3.60 \times \text{ equity,}$$

so the closing balance is $24{,}000 - 86{,}400 =$ **−CHF 62,400.00**, a negative balance of 2.600 times the original deposit. The realised loss is $86{,}400/3{,}840 =$ **22.50** times the intended loss. (Sanity check against Result 15.5: $L\,g = 20 \times 0.18 = 3.60$ times equity, as computed.)

**(b)** Result 15.5 gives $L \le \theta/g = 0.40/0.18 =$ **2.222222:1**. The corresponding notional is $2.222222 \times 24{,}000 =$ **CHF 53,333.33**, i.e. $53{,}333.33/1.2000 =$ **EUR 44,444.44** — which is $44{,}444.44/400{,}000 =$ **11.11**% of the position actually held. Note that this is still nine times more levered than the 1.67:1 that a 25 per cent tolerance through a 15 per cent gap would permit; "prudent" here is a choice of $\theta$ and $g$, not a fact.

**(c)** The guarantee absorbs everything beyond the stop: the shortfall is $18.0\% - 0.8\% = 17.2\%$ of notional. Break-even requires premium = expected payout, $c = \lambda g$, so

$$\lambda^* = \frac{0.0025}{0.172} = 0.014535 = \textbf{1.4535}\% \text{ per protected trade,}$$

one such gap every $1/0.014535 = 68.80$ protected trades.

Annual cost: $180 \times 0.40 = 72$ triggers, each charging $0.0025 \times 480{,}000 =$ CHF 1,200.00, so the annual premium is $72 \times 1{,}200 =$ **CHF 86,400.00**, which is **360.00**% of equity per year. That is *exactly* the single gap loss from part (a) — the ratio is 1.0000. The break-even trade count follows from setting $n\,h\,c = g$ with $h$ the trigger rate:

$$n^* = \frac{g}{h\,c} = \frac{0.18}{0.40 \times 0.0025} = \textbf{180.00} \text{ trades per year.}$$

*Reading.* At this size and turnover the guarantee is worth buying only if a depeg-scale gap is expected *more often than once a year*. Two conclusions follow, and neither is comfortable. Per-trade gap insurance is not a viable component of a high-turnover systematic book: its cost is of the same order as the catastrophe it insures, annually. And part (b) is therefore the operative answer — the affordable defence against an 18 per cent gap is to hold a position roughly one-ninth the size, permanently.

**Exercise 15.2 — The margin spiral, three rounds**

A fund holds 5,000 million of a single asset class against equity of 250 million, financed entirely by repurchase agreements at the haircut its leverage implies. Volatility rises and its financiers reprice the haircut in three steps: to 8 per cent, then 12 per cent, then 18 per cent. At each step the fund sells exactly enough to satisfy $A \le E/h$. Assume every 100 million sold marks the asset class down 0.20 per cent, that sales execute at the price prevailing at the start of the round, and that the impact marks the surviving book at the end of the round.

(a) State the fund's starting haircut and leverage.

(b) Work the three rounds: required assets, forced sale, price impact, mark-down, and

closing equity, assets and leverage for each.

(c) Report the cumulative sales as a percentage of the original book, the cumulative price decline, the total equity destruction, and the number of percentage points of equity destroyed per percentage point of price decline. Comment on what a value-at-risk model calibrated on this asset class would have made of the price move.

(d) Three funds hold identical books and deleverage simultaneously. What is the round-one sale and the round-one price impact?

▷ working space

**Solution — Exercise 15.2**

**(a)** $h_0 = E/A = 250/5{,}000 = \mathbf{5.00}\%$, so leverage is $1/0.05 = \mathbf{20.00{:}1}$.
**(b)** Round by round, using $A_{\text{req}} = E/h$, impact $=$ (sale$/100$) $\times$ 0.20%, mark-down $=$ impact $\times$ surviving assets:
*Round 1*, $h = 8\%$. $A_{\text{req}} = 250.0000/0.08 = 3{,}125.0000$; sale $= 5{,}000 - 3{,}125 = \mathbf{1{,}875.0000}$; impact $= 18.75 \times 0.20\% = \mathbf{3.7500}\%$; mark-down $= 3{,}125.0000 \times 0.037500 = 117.1875$; equity $250.0000 \to \mathbf{132.8125}$; assets $3{,}125.0000 \to 3{,}007.8125$; leverage $\mathbf{22.6471{:}1}$.
*Round 2*, $h = 12\%$. $A_{\text{req}} = 132.8125/0.12 = 1{,}106.7708$; sale $= 3{,}007.8125 - 1{,}106.7708 = \mathbf{1{,}901.0417}$; impact $= 19.010417 \times 0.20\% = \mathbf{3.8021}\%$; mark-down $= 1{,}106.7708 \times 0.038021 = 42.0803$; equity $132.8125 \to \mathbf{90.7322}$; assets $1{,}106.7708 \to 1{,}064.6905$; leverage $\mathbf{11.7344{:}1}$.
*Round 3*, $h = 18\%$. $A_{\text{req}} = 90.7322/0.18 = 504.0675$; sale $= 1{,}064.6905 - 504.0675 = \mathbf{560.6230}$; impact $= 5.606230 \times 0.20\% = \mathbf{1.1212}\%$; mark-down $= 504.0675 \times 0.011212 = 5.6518$; equity $90.7322 \to \mathbf{85.0803}$; assets $504.0675 \to 498.4157$; leverage $\mathbf{5.8582{:}1}$.
**(c)** Cumulative sales $= 1{,}875.0000 + 1{,}901.0417 + 560.6230 = 4{,}336.6646$ million, or $\mathbf{86.73}\%$ of the original book. Cumulative price decline $= 1 - (1 - 0.037500)(1 - 0.038021)(1 - 0.011212) = \mathbf{8.4477}\%$. Equity fell from 250.00 to 85.0803 million, $-\mathbf{65.97}\%$. Dividing, each percentage point of price decline destroyed $65.97/8.4477 = \mathbf{7.81}$ percentage points of equity. That last ratio is the exercise's point. An 8.45 per cent decline in a high-grade asset class is unremarkable — comfortably inside any reasonable one-year value-at-risk band, and the sort of move a risk report would describe as a normal drawdown. It destroyed two-thirds of the fund's capital and forced out seven-eighths of its book. The value-at-risk model was not wrong about the price distribution; it was answering a question about prices when the binding question was about *financing terms*, which do not appear in the price history at all. Note also that the fund's leverage fell from 20:1 to 5.86:1 — it succeeded completely at deleveraging, and that success is what the 65.97 per cent equity loss bought.
**(d)** With three identical funds, the round-one sale is $3 \times 1{,}875 = \mathbf{5{,}625}$ million and the round-one impact is $56.25 \times 0.20\% = \mathbf{11.25}\%$ — three times the price move, before the second round has begun. The spiral's severity scales with the amount of capital sharing the position and the financing arrangement, which is the formal content of "crowded trade". The individual fund's risk is not a property of the fund.

**Exercise 15.3 — Classifying six episodes**

Assign each of the following described episodes to one of the six failure modes of Table 15.1. For each, name the dominant mode, name any secondary mode present, and state which of the platform's defences (from the projbox of §15.7) would be the relevant one.

(i) A volatility-targeting overlay licensed to many institutional users rebalances to a fixed risk budget. A volatility spike triggers simultaneous de-risking across all users; the selling raises realised volatility further, triggering more de-risking.

(ii) A prime broker doubles the initial margin required on a fund's holdings after a ratings action on the collateral. The fund's positions and thesis are unchanged; it must nonetheless sell half the book within a week.

(iii) A listed security falls 40 per cent in ninety seconds on no news, prints twice at one cent, and closes the session one per cent below the previous close. The venue subsequently cancels the aberrant prints.

(iv) A central bank that has capped its currency's appreciation for two years announces without notice that the cap is removed. Quoting in the pair ceases for several minutes.

(v) Over eight sessions an investor's government bonds, investment-grade credit, gold and equities all fall together, and the bid-ask spread on the government bonds — normally

the tightest instrument he owns — widens by a larger multiple than the spread on the equities.

(vi) A multi-strategy fund running thirty hedged relative-value books at 20:1 loses 15 per cent of capital in three weeks as every one of its spreads widens together. Its financiers, who run similar books themselves, reduce their own positions in the same instruments.

▷ working space

**Solution — Exercise 15.3**

**(i) Feedback crowding**, with no meaningful secondary. The signature is a shared rule whose execution changes the input the rule reads — here realised volatility, exactly as the 1987 replication read price. Relevant defence: none detects it; the deflation machinery [Course Book §15.4] discounts the backtest, and the stressed cost surface prices the exit.

**(ii) Funding run.** The tell is that the mandated sale is independent of the borrower's view or solvency: a haircut change is a liquidation order issued by a third party (Definition 15.3). Secondary: *leverage unwind*, since the forced sale marks down the remaining book and raises leverage. Relevant defence: sizing that leaves headroom above the margin requirement, plus regime conditioning [Course Book §13.8] on funding-stress symptoms.

**(iii) Liquidity withdrawal.** Speed of seconds, a print at a nominal quote, near-complete recovery within the session, and ex-post trade cancellation are the four diagnostics together. Relevant defence: the stressed decile of the measured cost surface and execution realism [Course Book §18.3] — never assume a mid fill in an event window — plus the recognition that a resting stop is a market order.

**(iv) Peg break.** The cessation of quoting is the diagnostic: the event removes the price *path*, not merely the level. Secondary: *liquidity withdrawal*, which is the proximate mechanism by which the gap is realised. Relevant defence: only Result 15.5's size ceiling and instrument screening; no statistical machinery applies.

**(v) Dash for cash.** The diagnostic is the *ordering*: the most liquid, highest quality instrument suffers the largest relative spread widening, which is impossible under a fundamentals-driven repricing and automatic under liquidity-ordered selling. Relevant defence: stressed correlations grouped by liquidity rather than by asset label, and worst-decile path evaluation [Course Book §16.4].

**(vi) Leverage unwind**, with two secondaries that make it worse: *feedback crowding* (the financiers hold the same positions, so the correlation across the thirty books runs through their common holders rather than through fundamentals) and *funding run* (the financiers' own deleveraging will reprice the fund's financing). This is the LTCM configuration in miniature. Relevant defence: conditional value-at-risk sizing [Course Book §14.5] and fractional Kelly [Course Book §14.3] on a *stressed* correlation matrix, not the sample one.

**Exercise 15.4 — Ranking the failure modes for this platform**

Rank the six failure modes of Table 15.1 by the danger each presents specifically to the SST platform as it is currently configured — a leveraged CFD book of 21 OANDA instruments, fourteen of them FX pairs, forecast with a distributional envelope, with a documented history of high hit rates and negative expectancy, and with its bid/ask cost surface still being measured. Defend the ranking with mechanisms and arithmetic from this chapter, not with intuition, and state explicitly the criterion by which you are ranking. Identify at least one defensible alternative ordering and the assumption that would produce it.

▷ working space

**Solution — Exercise 15.4**

**Criterion.** Rank by expected damage to the programme's viability: the product of how likely the mode is to touch this book, how severe it is when it does, and — decisively — whether a defence exists that the platform can actually deploy. A mode that is frequent but fully defended ranks below a mode that is rare and undefendable.

**1. Peg break (gap through stops).** It is first because it is the only mode that can produce a loss *larger than the account*, and because no forecasting machinery addresses it. Worked Example 15.3: 30:1 through a 15 per cent gap is a loss of 4.50 times equity and a negative balance of 3.50 times the deposit; in an AUD-denominated account with a wrong-way conversion, 5.625 and 4.625 times. The book's composition makes this concrete rather than theoretical — fourteen FX pairs, with the franc and yen complex being precisely where administered and haven dynamics operate, and v1 FX being the primary version. The trapbox of §15.5 adds the aggravating factor: an administered price's suppressed volatility causes a volatility-targeting sizer to allocate roughly eleven times more notional to it. The defence is a leverage ceiling ($L < 1/g$) an order of magnitude below the provider's cap, and Worked Example 15.4 shows the insurance alternative costs several times the gross edge.

**2. Liquidity withdrawal and spread-regime shift.** Second on severity but first on *certainty*. The platform's own recorded history is high hit rate with negative expectancy [Course Book §14.1], which is the empirical signature of a book whose costs already exceed its edge; costs are therefore the dominant term, not a correction to it. Section 15.4 gives the arithmetic: against a 3 pip edge, a stressed-fill share of 7.79 per cent erases the edge entirely, and at 2 per cent the edge is down a quarter. This is not a tail event — OANDA spreads widen at news, at rollover and in thin hours, which are exactly the windows in which a short-horizon signal fires, so the stressed-fill share is *correlated with the strategy's own trade timing* rather than random. The mid-only candle history systematically understates it, which is why replacing it with a measured bid/ask surface is the highest-value piece of infrastructure in the programme.

**3. Dash for cash.** Third because it attacks the risk model rather than the account directly. Twenty-one instruments are nowhere near twenty-one independent bets — fourteen USD-legged FX pairs in a dollar-funding event are close to one — so a CVaR budget computed on sample correlations understates the tail by a factor that grows with the number of instruments. March 2020 adds the second blow: providers raise margin mid-event, and a book at the old limit must liquidate at the widened spread regardless of view (3.33 per cent to 20 per cent multiplies required equity sixfold). It ranks below the first two only because stressed-correlation sizing and worst-decile path evaluation [Course Book §16.4] genuinely address it.

**4. Funding run.** Fourth: it is the *state* in which modes 1–3 occur rather than a direct attack on a retail book, it runs over days to months, and its symptoms (RORO, volatility regime, funding-currency behaviour) are the covariates the Stage-2 router [Course Book §13.8] already consumes. Detection is lagging by construction [Course Book §13.7], which caps how much credit it can be given.

**5. Leverage unwind.** Fifth because the platform's leverage is its own choice and the Course Book's sizing machinery addresses it directly [Course Book §14.3], [Course Book §14.5]. The exposure is indirect: other participants' unwinds move the platform's instruments, which is a volatility and correlation event already covered under mode 3.

**6. Feedback crowding.** Last, not because it is benign but because the platform is too small to be part of the feedback and its strategies are not, at present, crowded. Its harm is epistemic — backtests of popular signal families overstate their future — and it is handled by the deflation and overfitting machinery [Course Book §15.4], [Course Book §15.5].

**A defensible alternative.** Swap ranks 1 and 2 if the platform commits *ex ante* to the gap-survivable ceiling of Result 15.5 — say 2:1 rather than 30:1. Under that assumption a 15 per cent gap costs 30 per cent of equity rather than 450 per cent: painful, survivable, and no

> longer existential. Spread-regime shifts then dominate, because they are certain, recurrent, and erode expectancy every week rather than once a decade. The assumption that flips the ranking is therefore a policy choice about leverage, which is precisely the argument for making that choice explicit and hard-coded rather than leaving it to the margin rule.

# 16

# Crypto Markets

*Chapter 15 spent six case studies establishing that markets fail in patterned, recognisable ways: strategy crowding creates its own selling pressure, leverage turns being right eventually into being dead first, funding runs starve the illiquid, and a counterparty's promise breaks exactly when it is needed most. Crypto is not a seventh case study. It is a market barely two decades old, still assembling institutional layers — deep prime brokerage, mature clearing, settled regulatory oversight — that took foreign exchange and equities the better part of a century to build, in which several of Chapter 15's failure modes are live structural possibilities rather than historical lessons. This chapter treats BTC and ETH exactly as the platform treats them: two CFD instruments among twenty-one, priced off a spot composite the platform does not control, running at a volatility several multiples of any FX pair's and a correlation to the rest of the book that drifts rather than holds. It traces what a blockchain actually settles, and why that makes "not your keys, not your coins" a statement about counterparty risk rather than a security slogan; draws out the parallel Chapter 4 half-completed for foreign exchange — a fragmented, tape-less venue landscape, radicalised into a market that never closes at all; works the funding-rate mechanism that anchors a perpetual future to spot in place of an expiry date; and closes by asking exactly one question of machinery this book has already built: does trading BTC and ETH require anything genuinely new, or does Chapter 14's sizing rule and Chapter 3's cost taxonomy already know what to do with an instrument three times as volatile as the one next to it in the book. Mostly, they already know.*

## 16.1 What Trades

Two of the platform's twenty-one instruments are crypto CFDs: BTC/USD and ETH/USD, referencing Bitcoin and Ethereum against the US dollar. Nothing else in the book's roster works quite the way the assets underneath these two instruments do, and the difference starts below the market entirely, at the ledger.

### What a blockchain settles

A blockchain is a shared, append-only ledger, replicated across a large number of independently operated nodes, that reaches agreement on the order and validity of transactions without a central operator adjudicating disputes. Bitcoin, the first working design, solved the specific problem of preventing the same unit of value from being spent twice absent any trusted intermediary, by having nodes compete to extend the ledger with new blocks of transactions and by defining the valid chain as whichever one embodies the most accumulated computational work — proof-of-work, in Nakamoto's (2008) original design. Ethereum ran the same consensus problem on a proof-of-work chain from its own 2015 launch and migrated the live network to proof-of-stake in 2022, replacing computational competition with economic stake as the resource nodes commit to earn the right to extend the chain; the specific consensus mechanism differs, but the settlement guarantee it exists to provide does not: once a transaction is buried under enough subsequent blocks, reversing it would require redoing an amount of work or stake the honest majority of the network is assumed never to concede control of.

What the ledger settles is narrow and precise: who controls which units, as recorded in the chain itself. Chapter 10 built the central counterparty as the institution that removes bilateral credit risk from a futures market, standing as universal buyer to every seller and seller to every buyer once a trade is matched. A public blockchain removes the intermediary at the settlement layer entirely rather than interposing a stronger one: there is no clearing house, because the ledger itself, once a transaction is sufficiently confirmed, is the authoritative record of ownership, with nothing left to novate. What this cannot do is remove counterparty risk one layer up, at whichever venue a trade actually happens on — and that is where custody becomes a precise, not a rhetorical, distinction.

**Definition — Custody: on-chain wallet versus exchange balance**

A *self-custodied* position exists as an entry on the blockchain itself, controlled by whoever holds the private key authorising a transfer of it; nobody else's ledger, database or promise stands between the holder and the asset, and nobody else's insolvency can make the position disappear. An *exchange balance* is not this. It is an entry in that exchange's own internal database — a liability the exchange owes the client, redeemable on request exactly as long as the exchange remains solvent and able to process the request — structurally identical in kind to a bank deposit or a broker's cash balance, whatever marketing language surrounds it.

**Trap — "Not your keys, not your coins" is a counterparty statement**

The slogan is not really about security hygiene, though it is often presented that way. It is a precise statement about whose liability a balance is. Every other instrument in this book is, by construction, a claim on an intermediary: an FX position is a claim on a broker (Chapter 4), a listed option a claim mediated by a clearing house (Chapter 11), a futures position free of bilateral credit risk only because a CCP stands behind it (Chapter 10). Crypto is the one asset class in the platform's roster for which a genuine alternative — direct, intermediary-free ownership — structurally exists. Choosing not to use it is not a mistake; it is choosing convenience and liquidity over that alternative, and §16.5 returns to exactly what is being chosen instead.

### The venue landscape

Chapter 4 opened by observing that foreign exchange has no building: no single order book, no consolidated tape, no regulator-mandated best bid and offer, only a tiered network of dealers each quoting the same handful of pairs slightly differently. Crypto reaches the identical diagnosis by a different structural route. Rather than a dealer network, the spot market is served by a large number of independently operated exchanges worldwide, most of them running conventional central-limit-order-book matching much like the equity venues of Chapter 8, alongside a newer, structurally distinct category — decentralised exchanges, which match trades against an on-chain liquidity pool via an automated pricing formula rather than a matching engine at all, with no operator taking custody of client funds at any point. Neither structure produces a consolidated tape. Each centralised exchange prints its own last trade; each pool prices off its own reserves; nothing forces any of them into agreement, and under stress they can and do diverge by economically meaningful amounts, for the same reason no analogous prime-brokerage credit network yet exists to arbitrage the gap the way it does across FX's dealer tiers. The lesson Chapter 4 drew for OANDA's own quote stream — one dealer's view, not the market's — applies here in a harder form: one *venue's* view, with materially less institutional machinery forcing convergence with anyone else's.

Most global trading volume quotes not against a fiat currency directly but against a stablecoin — a token engineered to hold a fixed peg, typically one-to-one to the US dollar, letting a venue

offer dollar-denominated pricing and settlement without a banking rail behind every single trade. USDT and USDC are, as of the mid-2020s, the two dominant examples by usage (indicative; the stablecoin landscape has shifted materially over the asset class's short history and should be verified current rather than assumed static). The platform's own two instruments sit outside this convention: OANDA is a regulated retail broker rather than a crypto exchange, and settles BTC/USD and ETH/USD in the account's own fiat currency against a constructed spot reference rather than any stablecoin pair — §16.3 returns to exactly what that reference is.

## 16.2 The 24/7 Clock

Chapter 4 built an entire session-bucketed cost surface on the fact that foreign exchange, continuous as it is within the week, still closes for roughly two days at the weekend and thins measurably around each day's rollover. Crypto removes even that. There is no scheduled close: not a daily auction boundary the way every cash equity and index venue in Chapter 8 has, not a contract's own trading hours the way the futures curve of Chapter 10 has, and not a weekend the way FX itself has. Some venue, somewhere, is open at every instant of every calendar day, public holidays included, and has been continuously since shortly after each network's launch.

This does not make liquidity uniform. Participation still tracks the working hours of the desks and retail populations that dominate flow, and weekend and deep-UTC-small-hours liquidity is widely and consistently documented as thinner than liquidity during the overlapping US and European trading day — a real, tradeable pattern, but one whose precise magnitude drifts with the market's own maturation and should be treated as indicative, not hard-coded as a constant.

BTC and ETH are, as a consequence, the only two instruments in the platform's twenty-one-instrument book with no scheduled closure of any kind. The gap to the platform's other data-wall arithmetic is not cosmetic. [Course Book §6.8]'s bar-count convention for the platform's intraday timeframes is built on FX's own trading week, roughly 118 of a possible 120 nominal hours once rollover and weekend-reopen thinning are discounted — a 70.24% share of a full 168-hour calendar week. Applying that same 118-hour convention to BTC or ETH would understate each instrument's own M1 bar supply by a clean, computable margin. At 60 minutes per hour across 156 weeks (three years), each hour-of-week is worth 9,360 M1 bars — the platform's own bar-count table implicitly uses exactly this scaling to reach its published 1,104,480 bars per instrument over three years from FX's 118-hour week. A full 168-hour week yields $168 \times 9{,}360 = 1{,}572{,}480$ M1 bars per instrument over the same three years: 468,000 more than the FX convention, a 42.37% undercount, before a single node outage or thin-liquidity minute is even subtracted — an idealised ceiling, not a claim about achieved uptime (verified in `verify_mk16.py`). The platform's data-wall arithmetic should size these two instruments' own bar budget on their own week, not borrow FX's.

Session bucketing needs the same rebuild, for a related but distinct reason: it is not merely that the hour count differs, but that the FX bucket names themselves — Sydney, Tokyo, London, the London–New York overlap, `rollover_hour` — describe boundaries a market with no session simply does not have. Nor do gaps behave the same way. The weekend gap Chapter 4 described, and the cash-open gap Chapter 9 decomposed exactly, are both, structurally, the market pricing in everything that happened while one specific venue was *closed*, released in a single jump at reopen. With no close, that mechanism has nothing to fire at: a large, fast crypto move is a genuinely different object, a rapid repricing that happens *within* continuous trading under the same liquidity-withdrawal dynamics Chapter 15 documented for other markets under stress, not a scheduled reopening after a period nobody could trade at all. Treating a fast BTC move as though it were a scheduled-closure gap would misattribute a liquidity event as a calendar event, and misdirect exactly the diagnostic Chapter 9's decomposition exists to run.

**Your project — The cost surface's crypto cells, and weekend volatility in evaluation**

Two rebuilds, not one relabelling. First, the session partition: the platform's eight FX buckets do not transplant onto a calendar BTC and ETH do not share, so their own cost surface needs hour-of-day and day-of-week fixed effects fitted directly on their own spread history, plus a `weekend_flag` standing in for `rollover_hour`'s role without pretending to be it — a period of typically thinner, not absent, liquidity. Second, evaluation: an FX-shaped harness can quietly assume weekend bars do not exist and lose nothing, because for FX they mostly do not; applied unchanged to BTC and ETH it would either drop two full days of genuinely tradeable, genuinely volatile data every week, or pool it with weekday bars as though the two regimes were exchangeable. Neither is correct. The calibration checks — coverage, PIT — that [Course Book §17.4]'s per-regime evaluation already runs by volatility and trend state need a weekday/weekend split added for exactly these two instruments, on the same principle Chapter 9 used to isolate the index CFD's cash-open bar: a bar drawn from a structurally different liquidity regime is not exchangeable with its neighbours, and calibration measured on the pool tells you nothing reliable about either regime alone.

## 16.3 Spot, Perpetuals, and the Funding Rate

Three ways of holding a crypto price exposure matter, and only one of them is what the platform actually trades. *Spot* is outright ownership (or, on an exchange, an exchange-balance claim on) the asset itself, settled immediately, with no expiry and no financing mechanic beyond whatever the holder does with the asset — the crypto analogue of a cash equity position. A *dated future*, structurally identical to Chapter 10's futures curve, is a contract on the asset's price at a fixed expiry $T$, converging to spot only as $T$ approaches, by the promise of eventual delivery or cash settlement. The *perpetual future* — commonly just "perp" — is neither: an exchange-traded derivative with no expiry and no delivery date at all, the dominant crypto derivative instrument globally by volume, and the instrument this section's mechanism exists to explain.

**Definition — Perpetual future**

A perpetual future tracks an underlying asset's price indefinitely, with no expiry to force convergence the way a dated future's approaching delivery date does. In its place, the contract uses a periodic *funding* payment, exchanged directly between the two sides of every open position, to keep the perpetual's own traded price economically tethered to a spot reference for as long as the contract exists — which, by design, is forever.

### The mechanism

**Result — The funding-rate mechanism**

Let $P_t^{\text{perp}}$ be the perpetual's own traded price and $P_t^{\text{idx}}$ a spot price index the venue constructs, typically itself a composite across several underlying spot markets. The funding rate for the next settlement, $F$, is set from the recent average premium or discount of the perpetual to that index, $\left(P^{\text{perp}} - P^{\text{idx}}\right)/P^{\text{idx}}$, usually adjusted by a small interest-rate offset reflecting the cost of carry between the two legs' collateral currencies. The exact averaging window, offset and any clamp on extreme values are venue-specific and drift; what is structural, and common to essentially every venue offering the instrument, is the sign logic and the settlement itself. At each fixed settlement instant — commonly, though not universally, every eight hours — every open position pays or receives

$$\text{funding payment} \;=\; N \times F, \tag{16.1}$$

where $N$ is the position's signed notional (positive long, negative short) at that instant. A long position with $F > 0$ (the perpetual trading rich to the index — long demand outweighs short interest) pays; a long position with $F < 0$ (the perpetual trading cheap) receives; a short position takes the opposite sign in both cases. Funding is a transfer between the two sides of the same contract, not a fee paid to the venue, which is what lets it run for years in either direction without itself draining the market the way a spread or a commission would.

**Worked Example 16.1 — A day's funding, and its annualised rate**

A trader holds a long BTC perpetual position with notional exposure $N = \$50{,}000.00$ on a venue settling funding every eight hours — three settlements a day, at 00:00, 08:00 and 16:00 UTC (a common, though not universal, convention; verify the specific venue's own schedule). Funding prints at $F = +0.0100\%$ at each of the day's three settlements, the perpetual trading persistently rich to its index.
By (16.1), each settlement's payment is

$$N \times F = \$50{,}000.00 \times 0.000100 = \$\mathbf{5.00},$$

paid by the long side each time, for a daily total of

$$\$5.00 \times 3 = \$\mathbf{15.00}.$$

Restated as an annualised rate, for comparison with the FX rollover swap of Chapter 3:

$$0.0100\% \times 3 \times 365 = \mathbf{10.9500}\% \text{ p.a.},$$

a genuinely carry-like cost, in the same units and the same order of magnitude as a stressed FX interest-rate differential, despite arising from a completely different mechanism — order-book positioning imbalance, not a central-bank policy rate. The short side of the same three settlements receives the identical \$15.00 a day; nothing is paid to or collected by the venue itself. Held at this constant rate for a full week the position would pay $\$15.00 \times 7 = \$105.00$; Exercise 16.1 computes the true weekly bill once the rate itself flips sign partway through the week, as it routinely does once positioning flips from crowded-long to crowded-short.

### The basis trade, and what it cannot hedge

Persistently positive funding — the common state in a strongly trending, leveraged-long market — is directly harvestable. A delta-neutral construction buys spot (or a spot-tracking instrument) for the long leg and sells the perpetual in equal notional for the short leg: the perpetual leg now receives funding every settlement, the mirror of Worked Example 16.1's long position, while the spot leg carries no funding obligation at all, since only the perpetual contract has one. Price risk on the underlying cancels between the two legs; ignoring execution costs, the position's expected return is the harvested funding stream itself — the same cash-and-carry logic Chapter 9 derived for an index future's cost of carry and Chapter 12 derived for a commodity's storage-and-convenience-yield curve, run here against a funding rate instead of an interest-rate or storage differential.

What the construction cannot cancel is counterparty risk. Both legs typically sit on the same venue, or are split across two venues, which multiplies rather than removes the exposure; the trade's realised profitability depends on whichever venue or venues hold the position remaining solvent and able to honour withdrawals for as long as it is held. Menkhoff et al. (2012) documented, for the FX carry trade, that the trade's own crowding is highest and its vulnerability greatest exactly when volatility is about to spike — carry earned slowly reverses violently, precisely

because everyone harvesting it is positioned the same way when the reversal arrives. The funding-harvest trade shares the pattern by construction: funding runs persistently positive exactly when the underlying has been trending hard and drawing in leveraged long demand, which is to say, exactly when a sharp reversal (and a venue under simultaneous liquidation stress) is least improbable. A funding-harvest position that looks like a riskless arbitrage on a spreadsheet is, uncancelled, a bet on exchange solvency the spreadsheet does not price at all — Chapter 15's counterparty lesson, applied to a venue class with a shorter history and less regulatory oversight than any exchange examined there directly.

#### What the platform's own CFD references

None of the mechanism above is what generates a charge on the platform's own BTC/USD or ETH/USD position directly. OANDA is a regulated retail broker, not a crypto exchange, and prices its two crypto CFDs off a constructed spot composite — an aggregated reference drawn across multiple underlying spot venues, in the same structural spirit as the index CFD's future-referenced construction of Chapter 9 — rather than trading an on-chain perpetual itself. Whatever overnight financing a held CFD position accrues is charged at OANDA's own published rate and administration, the same notional-times-rate-times-days-held construction as every other overnight financing line in this book (Chapter 3), not the exchange-native funding mechanism above.

Understanding that mechanism still matters to a platform trading the CFD proxy rather than the perpetual directly, for two reasons. It is the single most liquid, most continuously quoted gauge of aggregate crypto positioning anywhere in the market — a covariate available in real time, unlike the COT report's Tuesday-to-Friday release lag (Chapter 7). And it is the rate the composite's own underlying venues are themselves arbitraging against, so a large, persistent funding imbalance is exactly the condition under which the spot composite itself is most likely to move quickly. The exact current details of OANDA's own crypto financing — benchmark, administrative spread, charging cadence — are, as with every provider-specific figure in this book, to be verified against current product documentation, never assumed.

## 16.4 Volatility and Correlation Character

Crypto's realised volatility runs structurally higher than any FX pair in the platform's book. Across the mid-2020s, BTC and ETH are routinely cited at something in the order of three to six times a major FX pair's typical annualised volatility (indicative; the ratio is itself regime-dependent, and has at times compressed as crypto derivatives markets have deepened — verify current levels before treating any specific multiple as a constant).

What does not change is the *shape*. [Course Book §1.4]'s stylised facts — fat tails and volatility clustering — were established as properties of returns in general, not as properties peculiar to any one asset class, and BTC and ETH exhibit both exactly as documented there: large moves cluster in time, and the unconditional distribution is fatter-tailed than a Normal benchmark predicts. This is precisely why the platform's own per-instrument volatility forecasting needs no crypto-specific redesign, only crypto-specific parameters: [Course Book §22.4]'s own master table already lists "GARCH/HAR per pair; session-of-day heteroscedasticity" as the vol-dynamics row for the instrument class BTC and ETH sit inside, not as a special case bolted on afterwards.

Correlation is less stable than volatility's shape. Chapter 14 built the general case that cross-asset correlation is typical, not law, liable to drift and to break exactly when a portfolio most needs it to hold; crypto is less an exception to that thesis than an unusually visible illustration of it. BTC and ETH's beta to broad risk assets — US equities above all — has at various points across the 2020s run materially positive during macro-driven risk-on/risk-off episodes, a shared sensitivity to liquidity conditions and rate expectations, while at other times decoupling sharply on catalysts confined to the asset class itself — an exchange-specific stress event, a protocol upgrade,

a regulatory headline with no transmission channel into the equity market at all (indicative; the regime itself, not any single correlation estimate, is Chapter 14's point).

### Sizing: the rule that needs no special case

[Course Book §14.4]'s volatility-targeting rule sizes every instrument by the same construction,

$$w_t = \min\left(\frac{\sigma^*}{\hat{\sigma}_t},\ w_{\max}\right), \tag{16.2}$$

with $\sigma^*$ a chosen target volatility for the strategy's returns, $\hat{\sigma}_t$ an instrument-specific volatility forecast, and $w_{\max}$ a leverage cap. Nothing in (16.2) refers to an instrument's identity, only to its own forecast $\hat{\sigma}_t$ — which is exactly why a threefold gap between BTC's volatility and an FX pair's needs no branch in the sizing logic at all.

**Worked Example 16.2 — Same risk budget, three times the volatility**

A strategy targets $\sigma^* = 10\%$ annualised with a leverage cap $w_{\max} = 1.5$, exactly the Course Book's own convention. An FX pair's forecast volatility is $\hat{\sigma}_{\text{FX}} = 8\%$; BTC's is $\hat{\sigma}_{\text{BTC}} = 24\%$, three times the FX figure. By (16.2):

$$w_{\text{FX}} = \frac{0.10}{0.08} = \mathbf{1.250000}, \qquad w_{\text{BTC}} = \frac{0.10}{0.24} = \mathbf{0.416667},$$

neither cap binding. The ratio is exact:

$$\frac{w_{\text{BTC}}}{w_{\text{FX}}} = \frac{0.416667}{1.250000} = \mathbf{0.333333} = \frac{1}{3}.$$

At a base notional of \$100,000.00 per unit of weight, this is

$$\text{notional}_{\text{FX}} = 1.250000 \times \$100{,}000.00 = \$\mathbf{125{,}000.00},$$
$$\text{notional}_{\text{BTC}} = 0.416667 \times \$100{,}000.00 = \$\mathbf{41{,}666.67}.$$

At an illustrative BTC price of \$50,000.00, the BTC notional is \$41,666.67/\$50,000.00 = **0.833333** BTC. Three times the volatility, exactly one-third the units — the entire adjustment (16.2) makes, with $\hat{\sigma}_t$ fitted separately per instrument and nothing else changed.

Higher, differently-correlated volatility is therefore not, on its own, a reason to build a separate crypto sizing model. (16.2) already takes $\hat{\sigma}_t$ as given and divides by it; the estimation feeding that division still has to be fitted on BTC and ETH's own return history, at their own clustering dynamics, but the rule that consumes the estimate does not care which instrument produced it.

## 16.5 Market Quality, Honestly

Two honesty checks belong here, and neither is a reason to avoid the asset class; both are reasons to hold the CFD wrapper deliberately rather than by default.

**Market quality.** Many of the venues that set the global crypto price operate under materially lighter regulatory oversight than a national securities exchange or a CFTC-regulated futures market. Wash trading — a venue's own participants, or the venue itself, trading with themselves to inflate reported volume — and other manipulative practices are a persistent, publicly documented concern specifically on less-regulated venues. How large a share of any given venue's reported volume this affects is genuinely disputed and, from outside the venue, effectively unverifiable, precisely because the limited transparency that permits the practice also prevents a reliable external estimate of its scale; treat any specific percentage encountered elsewhere with real scepticism, since this book will not manufacture a precision the underlying data cannot support.

**Exchange failures.** Counterparty risk is not a theoretical add-on to holding a crypto exposure through an exchange balance rather than on-chain (§16.1). Exchange insolvencies have occurred; this is a matter of public record, and is the crypto-market instance of the same lesson Chapter 15 drew from six market failures in entirely different asset classes — a claim on an intermediary is only as good as that intermediary's solvency, and an exchange looking fine yesterday has never reliably predicted that it looks fine tomorrow.

**Regulatory drift.** How crypto assets, venues and derivatives are classified and regulated — as commodities, as securities, as a bespoke category, or as some jurisdiction-specific mixture of the three — continues to move, jurisdiction by jurisdiction, and has moved substantially in the years since Nakamoto's (2008) design was first published. Nothing here describes a settled regulatory perimeter in any jurisdiction, the reader's own included; verify current rules before treating any regulatory treatment as fixed.

**Trap — Choosing the CFD does not choose away counterparty risk**

§16.1 drew the custody line precisely: a self-custodied wallet is not anyone else's liability; an exchange balance is. A CFD position is the second case again, one further remove — a claim on OANDA, not on any crypto exchange and not on the underlying asset itself, since no private key exists for a CFD position at all. Trading BTC and ETH through the platform's CFD wrapper does not escape counterparty risk; it exchanges a crypto exchange's counterparty risk, in a young venue class carrying the market-quality and solvency concerns just described, for a retail broker's counterparty risk, in the same mature, regulated relationship the reader already extends to the other nineteen instruments in the book. That is a considered trade, not a free upgrade: regulatory oversight and one familiar counterparty, against giving up direct ownership of the underlying asset. Chapter 4's A-book/B-book description of what OANDA actually does with the other side of a client position applies to a crypto CFD exactly as it applies to a EURUSD one.

## 16.6 Exercises

**Exercise 16.1 — Funding across a week, with a sign flip**

A trader holds a long BTC perpetual position with constant notional $N = \$80{,}000.00$ throughout a full calendar week, funding settling three times a day as in Worked Example 16.1. From Monday through Thursday inclusive (4 days, 12 settlements), funding prints at a constant $F_1 = +0.0080\%$ per settlement. From Friday through Sunday inclusive (3 days, 9 settlements), positioning flips and funding prints at a constant $F_2 = -0.0060\%$ per settlement.

(a) Compute the total funding paid during the Monday–Thursday phase.

(b) Compute the total funding received during the Friday–Sunday phase.

(c) Compute the week's net funding P&L, and express it as a percentage of the position's notional.

▷ working space

**Solution — Exercise 16.1**

**(a)** Per (16.1), each Monday–Thursday settlement pays

$$N \times F_1 = \$80{,}000.00 \times 0.0000800 = \$6.40,$$

for 12 settlements:

$$\$6.40 \times 12 = \$\mathbf{76.80} \text{ paid.}$$

**(b)** Each Friday–Sunday settlement now pays the long side (funding negative):

$$N \times |F_2| = \$80{,}000.00 \times 0.0000600 = \$4.80,$$

for 9 settlements:

$$\$4.80 \times 9 = \$\mathbf{43.20} \text{ received.}$$

**(c)** Net funding P&L for the week:

$$-\$76.80 + \$43.20 = -\$\mathbf{33.60},$$

a net cost despite the sign flip, since the Monday–Thursday phase both ran at a higher per-settlement rate and lasted more settlements. As a fraction of notional,

$$\frac{-\$33.60}{\$80{,}000.00} = -\mathbf{0.0420}\%$$

for the week — a small figure in isolation, and exactly the kind of running cost that a strategy priced only on spread, and blind to funding, would never see coming out of its own P&L attribution.

**Exercise 16.2 — Vol-target sizing drill: FX versus BTC, same risk budget**

A strategy targets $\sigma^* = 12\%$ annualised with a leverage cap $w_{\max} = 2.0$. An FX pair's forecast volatility is $\hat{\sigma}_{\text{FX}} = 7.5\%$; BTC's is $\hat{\sigma}_{\text{BTC}} = 30\%$, four times the FX figure. Base notional at $w = 1$ is \$70,000.00; BTC's illustrative price is \$56,000.00.

(a) Compute $w_{\text{FX}}$ and $w_{\text{BTC}}$ by (16.2), confirming neither cap binds, and state the ratio $w_{\text{BTC}}/w_{\text{FX}}$.

(b) Convert both weights to notional, and convert the BTC notional to units of BTC.

(c) Suppose BTC's forecast volatility had instead been exactly twice the FX pair's, $\hat{\sigma}'_{\text{BTC}} = 15\%$, everything else unchanged. Recompute $w'_{\text{BTC}}$ and the resulting BTC notional and units.

▷ working space

**Solution — Exercise 16.2**

**(a)**

$$w_{\text{FX}} = \frac{0.12}{0.075} = \mathbf{1.600000}, \qquad w_{\text{BTC}} = \frac{0.12}{0.30} = \mathbf{0.400000}.$$

Both are below $w_{\max} = 2.0$; neither cap binds. Ratio:

$$\frac{w_{\text{BTC}}}{w_{\text{FX}}} = \frac{0.400000}{1.600000} = \mathbf{0.250000} = \frac{1}{4},$$

exactly the reciprocal of the four-times volatility gap.

**(b)**

$$\text{notional}_{\text{FX}} = 1.600000 \times \$70{,}000.00 = \$\mathbf{112{,}000.00},$$
$$\text{notional}_{\text{BTC}} = 0.400000 \times \$70{,}000.00 = \$\mathbf{28{,}000.00}.$$

In BTC units: $\$28{,}000.00 / \$56{,}000.00 = \mathbf{0.500000}$ BTC.

**(c)**

$$w'_{\text{BTC}} = \frac{0.12}{0.15} = \mathbf{0.800000}, \qquad \text{notional}'_{\text{BTC}} = 0.800000 \times \$70{,}000.00 = \$\mathbf{56{,}000.00},$$

which at \$56,000.00 per BTC is exactly **1.000000** BTC. Halving the volatility gap (four times to two times) exactly doubles both the weight and the BTC notional, precisely as (16.2)'s inverse-proportionality predicts — the relationship, not any one ratio, is what to carry away from either drill.

**Exercise 16.3 — Essay: what transfers from the FX playbook, and what does not**

In three or four paragraphs, assess how much of the FX-specific discipline built across Chapters 3–7 and 14 transfers to trading BTC and ETH unchanged, and how much needs genuine rework. Organise the answer around four themes: sessions, costs, counterparty, and regime/sizing.

▷ working space

**Solution — Exercise 16.3**

**Sessions do not transfer, and need a genuine rebuild.** Chapter 4's entire liquidity clock — Sydney, Tokyo, London, the London–New York overlap, `rollover_hour` — is built on FX's continuous-but-weekly-closed week. BTC and ETH have no session boundary of any kind (§16.2), so the bucket names themselves do not apply; what replaces them is an hour-of-day/day-of-week partition fitted on crypto's own spread and volume history, plus a `weekend_flag` for a thinner-not-absent regime, and gap handling drops out entirely, since there is no closure for a gap to be the record of.

**Costs transfer structurally but gain a genuinely new component.** The taxonomy of Chapter 3 — spread, commission where charged, financing — still applies to the platform's own CFD, and financing is still a notional-times-rate-times-days-held construction. What is new is the funding-rate mechanism (§16.3) underneath the broader crypto market even though it is not literally what the platform's own CFD charges: a covariate and a cost analogue with no FX equivalent, settling several times a day rather than once, set endogenously by order-book positioning rather than a central-bank differential.

**Counterparty logic transfers almost unchanged, once translated.** Chapter 4's "your quote stream is one dealer's view" and the A-book/B-book description of what a broker does with client flow apply to a crypto CFD exactly as they apply to a EURUSD one (§16.5): OANDA is still the counterparty, on every instrument in the book. What is genuinely different is the venue landscape one layer further out, where the platform's own spot composite is priced: younger exchanges, less consistent regulatory oversight, and a documented, unresolved wash-trading concern with no FX analogue at anything like this scale.

**Regime and sizing transfer completely, by construction.** [Course Book §14.4]'s volatility-targeting rule and Chapter 14's correlation-instability thesis were never FX-specific to begin with; they take an instrument's own $\hat{\sigma}_t$ and its own correlation history as inputs and require no branch for crypto at all (§16.4). The one genuine dependency is upstream of the rule, not in it: $\hat{\sigma}_t$ still has to be estimated well on BTC and ETH's own return history, at their own clustering and their own drifting equity beta, before the sizing and regime machinery downstream can do anything useful with it.

**Key idea — What to remember**

BTC and ETH are two CFD instruments among twenty-one, and most of what makes them different from the other nineteen is structural, not exotic: the underlying asset settles on a blockchain rather than through a clearing house, which makes an exchange balance — unlike an on-chain wallet — exactly the kind of intermediary liability every other instrument in this book already is; the venue landscape is fragmented and tape-less for reasons distinct from FX's dealer-network fragmentation but with the identical practical consequence, no single "the" price; and the market never closes, which forces a genuine rebuild of session bucketing and gap handling rather than a relabelling of Chapter 4's. The funding-rate mechanism anchoring the crypto market's dominant instrument, the perpetual future, is a carry cost in every arithmetic sense — notional times rate times periods, capable of flipping sign inside a single week — even though it is not literally what the platform's own spot-composite-referenced CFD charges. Higher volatility and a drifting equity correlation, by contrast, need nothing new at all: [Course Book §14.4]'s sizing rule and Chapter 14's regime thesis already take an instrument's own forecast and its own correlation history as given, and absorb a threefold or fourfold volatility gap into a proportionally smaller position without a single crypto-specific line of code. What remains genuinely open is counterparty quality one layer out — younger venues, thinner oversight, a documented manipulation concern this book will not pretend to quantify — which is precisely why the CFD wrapper is a considered trade

of one counterparty for another, not an escape from Chapter 1's question: whose money you are taking, and why they keep giving it to you. The appendices that follow close the book with exactly that discipline in reference form — pip values, session clocks, value-date rules and this chapter's funding-rate arithmetic, collected as quick lookups rather than taught again.

# A

# Market Conventions: A Quick Reference

*The formulas and conventions of Chapters 1–16, stripped of derivation and indexed by table rather than by argument — for the desk, not the armchair. Every figure that drifts carries a vintage or a verify-current instruction; every figure stated plainly is structure, exactly as the chapters cited alongside it derived or documented it to be.*

**Definition — Reading this appendix**

Two kinds of fact sit side by side in every table below. A figure marked *indicative*, or carrying the instruction *verify current specifications*, drifts — it has moved before and will move again, and this appendix's job is to show how it is built, not to freeze its current value. A figure stated plainly, with no such flag, is institutional structure this book's chapters argue is stable enough to state as fact and cite where possible. Neither kind of entry is a live feed: treat this appendix as a map of the machinery, not a data source to trade from directly.

## A.1 Pip sizes and pip values

Full derivation and worked arithmetic: Chapter 4, §4.3.

**Table A.1.** Quoting increment by instrument class.

| Class | Quoting increment | Finer increment | Illustrative pairs |
|---|---|---|---|
| FX, non-JPY quote | pip = 0.0001 | pipette = 0.00001 (5th decimal place) | EURUSD, GBPUSD, AUDNZD |
| FX, JPY quote | pip = 0.01 | pipette = 0.001 (3rd decimal place) | USDJPY, EURJPY, GBPJPY |
| Metals (XAU, XAG) | provider-set tick, quoted like an FX pair against USD | — | XAUUSD, XAGUSD |
| Crypto (BTC, ETH) | provider-set tick | — | BTCUSD, ETHUSD |
| Index CFDs | provider-set index point, own $-per-point value | — | the platform's two index CFDs |

The two FX rows are structure (Chapter 4, §4.3) and stable. Metal, crypto and index tick/point values are provider-set and drift: verify current specifications before sizing a live position (Chapters 12, 16, spec).

Standard lot sizes are likewise structural across the retail industry: a standard lot is 100,000 units of the base currency, a mini lot 10,000, a micro lot 1,000 (Chapter 4).

**Result — Pip value, three conventions — restated from Chapter 4**

For a position of $N$ units of the base currency, pip size $\pi$ (0.0001, or 0.01 if the quote currency is JPY), and spot rate $S$ (quote per base):

- account currency = quote currency: pip value = $\pi N$, no conversion;
- account currency = base currency: pip value = $\pi N/S$;
- account currency is neither leg: pip value = $\pi N \cdot X$, where $X$ is the prevailing rate between the quote currency and the account currency, triangulated (§A.1's companion identity below) when not streamed directly.

Derivation and three fully worked cases (EURUSD, USDJPY, AUDNZD): Chapter 4, Result 4.1 and Worked Example 4.1.

**Result — Triangulation — restated from Chapter 4**

For currencies A, B and a shared currency C: if C is the quote currency in both given pairs, $A/B = (A/C)/(B/C)$; if C is the base currency in both, $A/B = (C/B)/(C/A)$; if C is the quote in one and the base in the other, $A/B = (A/C) \times (C/B)$. A directly-quoted cross trading outside the band this identity implies (net of the two legs' combined spread) admits a riskless round trip, which is exactly what keeps it pinned (Chapter 4, Result 4.2).

## A.2 The session clock

Full derivation: Chapter 4, §4.4.

**Table A.2.** The FX trading week: UTC and the reader's AEST clock (no daylight saving).

| Boundary | UTC | AEST |
|---|---|---|
| Week opens (Wellington/Sydney) | Sunday 22:00 | Monday 08:00 |
| Week closes (New York) | Friday 22:00 | Saturday 08:00 |
| Daily rollover (5pm New York) | 22:00 (NY on EST) / 21:00 (NY on EDT) | 08:00 next day (EST) / 07:00 next day (EDT) |

Four session regimes recur daily, deepest to thinnest: the London–New York overlap (the day's liquidity peak); London on its own (the deepest single session); Asia (comparatively thin and rangey); New York late, after London has closed (thin again). Spreads widen, in the platform's own measured cost surface, in the thin buckets and at the rollover instant; this ordering is structure (Chapter 4, §4.4; [Course Book §18.3]).

**Table A.3.** Daylight-saving rules governing the UTC offsets above.

| Calendar | Daylight saving starts | Daylight saving ends |
|---|---|---|
| United States (New York) | 2nd Sunday of March | 1st Sunday of November |
| United Kingdom (London) | Last Sunday of March | Last Sunday of October |
| Australia (AEST, reader's own) | — no daylight saving — | — no daylight saving — |
| Japan (Tokyo) | — no daylight saving — | — no daylight saving — |

The rules are stable structure; the calendar *dates* they produce move every year and must be recomputed, never copied from a prior year. Because the US and UK windows do not start and end on the same date, the New York–London offset drops from five hours to four for one-to-three-week windows twice a year (Chapter 4, §4.4).

**Trap — Wednesday's triple swap is Wednesday in New York, not in Sydney**

At 5pm EST that is 22:00 UTC, which on the reader's own AEST clock (ten hours ahead, no daylight saving) is already Thursday 08:00. Checking a Sydney calendar for "Wednesday evening" and finding nothing unusual is not an exception to the rule; it is looking on the wrong day for an event timestamped in New York (Chapter 4).

## A.3 FX value dates and rollover

Full derivation: Chapter 4, §4.4; Chapter 3, §3.5.

**Table A.4.** Value-date roll and days financed, by weekday the 17:00 New York cut-off falls on.

| Roll fires on | Old value date | New value date | Days financed |
|---|---|---|---|
| Monday | Wednesday | Thursday | 1 |
| Tuesday | Thursday | Friday | 1 |
| Wednesday | Friday | Monday | **3 (tripled)** |
| Thursday | Monday | Tuesday | 1 |
| Friday | Tuesday | Wednesday | 1 |
| | | Weekly total | 7 |

Spot's T+2 value date cannot fall on a weekend, so the Wednesday-to-Thursday roll jumps three calendar days instead of one — the "triple swap Wednesday" convention, and nothing more than T+2 arithmetic meeting the calendar (Chapter 4). The triple redistributes a week's financing; it does not add to it: $1+1+3+1+1=7$, exactly the seven days a position held continuously through the week is charged. Public holidays shift the pattern and can produce double or quadruple rolls; these must be checked against the broker's published schedule, never assumed.

**Result — CFD overnight financing — restated from Chapter 3**

For a position of $N$ units of base currency at spot $S$, base/quote annualised rates $r_b, r_q$, broker markup $\alpha > 0$, and day-count basis $D$ (365 or 360 by currency):

$$f_{\text{long}} = \frac{NS\left[(r_b - r_q) - \alpha\right]}{D}, \qquad f_{\text{short}} = \frac{NS\left[(r_q - r_b) - \alpha\right]}{D}.$$

Holding both sides simultaneously nets to $-2\alpha NS/D$: the financing book is a spread business, and $\alpha$ is its spread. Non-FX CFDs (index, commodity, share) are financed on the same principle against a benchmark cash rate rather than a currency differential, typically charged on both sides rather than paid on one; share and index CFDs additionally carry the dividend adjustment of §A.6. Every rate and markup is provider-set and drifts: verify current specifications before pricing a live position (Chapter 3, §3.5).

## A.4 The COT calendar

Full treatment: Chapter 7, §7.1.

The Commitments of Traders report is dated to *Tuesday's* close of business and released the following *Friday* afternoon, US Eastern time — a structural three-calendar-day lag between reference date and release date, extended by one business day whenever the release Friday is a US federal holiday (CFTC). A backtest must join on the release timestamp, never the reference date ([Course Book §18.2]).

**Table A.5.** The three COT report families. Every covered contract carries Legacy, plus exactly one of Disaggregated or TFF, never both (CFTC, Explanatory Notes).

| Family | Covers | Categories |
|---|---|---|
| Legacy | Every covered contract | Commercial; Non-Commercial; Non-reportable |
| Disaggregated | Physical commodities (this platform's metals and WTI, among the instruments it holds) | Producer/Merchant; Swap Dealers; Managed Money; Other Reportables |
| Traders in Financial Futures (TFF) | Financial futures: FX, rates, equity indices (this platform's FX pairs and its two index CFDs, read via their CME tickers) | Dealer/Intermediary; Asset Manager/Institutional; Leveraged Funds; Other Reportables |

**Trap — COT is a futures-market proxy, not an OTC feed**

FX spot has no COT report of its own: this platform's "FX COT" reads CME currency-*futures* positioning, a centrally-cleared market materially smaller than, and populated by a different participant mix from, the OTC spot market OANDA's CFD prices actually reference. A cross with no CME contract of its own is approximated by differencing two USD-leg readings; a non-US equity index inherits no COT proxy at all (Chapter 7, §7.1).

## A.5 Contract structure: futures, options, and CFDs

CFD structure is drafted in full: Chapter 1, §1.1; Chapter 3, §3.5; Chapter 4, §4.6; Chapter 6 in full. Futures structure is drafted in full in Chapter 10; listed-options structure in Chapter 11 — both checked against their own finished text below.

**Table A.6.** What is standardised, what is provider- or venue-specific, and who stands as counterparty, by wrapper.

| Wrapper | Standardised | Not standardised (verify current) | Counterparty |
|---|---|---|---|
| Futures | Underlying, contract multiplier, tick size/value, expiry cycle, settlement method, margining regime (Chapter 10) | Exact multiplier/tick value per contract; margin levels; fee schedule | Central counterparty (CCP); novation removes bilateral credit risk |
| Listed options | Strikes, expiry cycle (monthlies/weeklies/0DTE), contract multiplier, exercise style by class (Chapter 11) | Multiplier and exercise style vary AU vs US and by underlying; liquidity by moneyness/tenor | OCC/CCP clears; AU and US structures differ |

continued overleaf

| Wrapper | Standardised | Not standardised (verify current) | Counterparty |
|---|---|---|---|
| CFD (this platform's instruments) | Nothing exchange-wide: the broker sets its own contract size, tick, margin and financing terms | Pip/tick size, spread, margin rate, financing markup, rollover instant, dividend-adjustment treatment: all provider-set (Chapters 1, 3, 4, 6) | The broker itself (OANDA); no exchange or CCP is ever party to the trade |

The structural constant underneath every futures and options row is central clearing: a CCP interposes itself after the trade is agreed, becoming counterparty to both sides, so neither an exchange trader nor OANDA's own upstream hedges carry the other's direct default risk the way a CFD client carries the broker's (Chapter 1, §1.1; Chapter 10).

## A.6 Index-CFD dividend adjustment

Full derivation and worked arithmetic: Chapter 9, §9.2.

> **Result — Dividend adjustment on an index CFD — restated from Chapter 9**
>
> On each ex-date, a long index CFD position is credited, and a short position debited, an amount equal to the index points the constituent's dividend contributes to the (price) index level, converted at the position's exposure per point. The credit and debit are equal in magnitude and opposite in sign. Whether a provider pays the long side the full gross amount or applies a withholding-style haircut is provider-specific: verify current terms.

An index CFD is priced off the price ("clean") index via the futures chain of Chapter 9, §9.1, never off a total-return series. Applying the credit/debit rule *and* simulating against a total-return series double-counts the same dividend; ignoring the rule while simulating against the price index misprices both the long side (understated) and the short side (overstated) of every backtest (Chapter 9). The identical logic, one level down, prices a single share's ex-date drop (Chapter 8, §8.4), with a franked Australian dividend's tax attribute riding on top of it as a separate, holder-specific layer (Chapter 8).

## A.7 Crypto funding-rate mechanics

This section is drafted in full in Chapter 16, §16.3; restated here and checked against that chapter's own finished text.

A perpetual future carries no expiry and is instead kept close to its spot reference by a periodic *funding* payment exchanged directly between longs and shorts: when the perpetual trades above spot, longs pay shorts; when it trades below spot, shorts pay longs — a mechanism giving every holder a continuous incentive to trade the perpetual back toward spot. The payment on one funding date is

$$\text{funding paid} = \text{position notional} \times \text{funding rate} \times (\text{funding periods held}),$$

with the funding *rate* itself typically quoted per period (commonly a small fraction of one percent, paid every eight hours on many venues, i.e. three periods a day) rather than per annum. The exact interval, rate-setting formula and reference spot composite are venue-specific and drift: verify current specifications before using any figure operationally. The platform's own BTC/ETH CFD instruments reference a spot composite rather than one exchange's perpetual directly; whether, and how, funding-style economics pass through to the CFD's own overnight financing charge is a

provider-specific structural question, answered the same way Chapter 3's general CFD-financing framework answers it for every other instrument class: notional times a rate, per day held, marked up.

# B

# Glossary

*The market terms this book uses most, alphabetically, one to two lines each, each pointing back to the chapter that earns it. All sixteen chapters are now drafted, and every pointer below is checked against each chapter's own finished text. §A states the formulas this glossary only points to.*

**A-book / B-book** Two retail dealing models: A-book passes client flow through to the broker's own liquidity providers for a markup; B-book internalises it, making the broker the client's direct counterparty. *(Chapters 4, 6.)*

**ADV (average daily volume)** The average dollar or share volume a name trades per day; the standard liquidity floor screened before a leg is sized. *(Chapter 9.)*

**Adverse selection** The risk a liquidity provider bears that a counterparty who chooses to trade against its quote knows something it does not, so accepted trades move price against the quoter on average. *(Chapters 1, 2.)*

**Arbitrage, triangular** A riskless round trip exploiting a directly-quoted cross rate that has drifted outside the band implied by its two constituent USD pairs. *(Chapter 4.)*

**Ask (offer)** The price a dealer, or the lowest resting sell order in the book, will sell at. *(Chapter 2.)*

**Auction (call auction)** A mechanism that accumulates orders over a window and clears them at the single price maximising matched volume; opens and closes every exchange session. *(Chapters 1, 2, 8.)*

**Backwardation** A futures curve sloping downward with maturity; in commodities, read as a scarcity or convenience-yield signal. *(Chapter 12.)*

**Base currency** In a BASE/QUOTE pair, the currency the quoted price is expressed per one unit of. *(Chapter 4.)*

**Basis** The difference between a derivative's price and its underlying's cash price, $F - S$; converges to zero at expiry. *(Chapters 9, 10, 12.)*

**Basis point** One hundredth of one percentage point; the standard fine-grained unit for rates, yields and differentials. *(Chapter 5.)*

**Bid** The price a dealer, or the highest resting buy order in the book, will buy at. *(Chapter 2.)*

**Bid-ask spread** The gap between the best bid and best ask; the price of immediacy, decomposing into order-processing, inventory and adverse-selection components. *(Chapter 2.)*

**Borrow (securities lending)** The arrangement supplying shares to a short seller, priced as an annualised fee on the position's notional; the fee tightens exactly when short demand is heaviest. *(Chapter 8.)*

**Breakeven spread (Glosten–Milgrom)** The minimum spread a riskless, costless dealer must still charge, derived from the probability an arriving trader is informed and the value gap between good and bad states. *(Chapters 1, 2.)*

**Carry (carry trade)** A position funded in a low-yield currency and invested in a high-yield one, harvesting the interest differential as the systematic reward for bearing the risk uncovered interest parity says should not be free. *(Chapters 3, 5.)*

**CCP (central counterparty)** A clearing entity interposed between the two sides of a trade after execution, becoming counterparty to both, so neither original party bears the other's default risk. *(Chapters 1, 10.)*

**CFD (contract for difference)** A bilateral OTC contract between client and broker settling in cash against the

change in a reference price, with the reference instrument itself never delivered. *(Chapter 1.)*

**Chain (options chain)** The full listing of a name's traded strikes and expiries, with each contract's quoted premium and implied volatility. *(Chapter 11.)*

**CIP (covered interest parity)** The no-arbitrage identity pinning the forward rate to the spot rate and the two currencies' interest rates, enforced by the FX swap. *(Chapter 5.)*

**Circuit breaker** A rule-based, market-wide trading pause triggered by a large, fast decline in a benchmark index; the direct regulatory descendant of the 1987 crash. *(Chapters 8, 15.)*

**Contango** A futures curve sloping upward with maturity; the normal shape when storage cost exceeds convenience yield. *(Chapter 12.)*

**Convenience yield** The non-monetary benefit of holding a physical commodity rather than a futures claim on it; the term that can flip a curve from contango into backwardation. *(Chapter 12.)*

**COT (Commitments of Traders)** The CFTC's weekly report of large-trader futures and options positions, dated to Tuesday's close and released the following Friday at the earliest. *(Chapter 7.)*

**Cross rate** A currency pair quoted between two non-USD currencies, pinned by triangulation to the two USD pairs sharing a leg with it. *(Chapter 4.)*

**Curve (term structure)** The relationship between an instrument's price or yield and its maturity or tenor. *(Chapters 10, 12, 13.)*

**Days-to-cover** Aggregate short interest divided by average daily volume; the standard proxy for how disorderly a rush to cover a crowded short could become. *(Chapter 8.)*

**Dealer** A participant that continuously quotes both sides of a market and stands ready to trade either direction, without necessarily carrying a formal exchange market-making obligation. *(Chapter 1.)*

**Depth** The quantity resting at, or within a stated distance of, the best prices on either side of the book. *(Chapter 2.)*

**Dividend adjustment** The credit (long) or debit (short) a CFD provider applies on a constituent's ex-date, replicating the cash entitlement a holder of the physical basket or share would receive or owe. *(Chapters 3, 9.)*

**Dollar smile** A practitioner framework in which the US dollar strengthens both in global crisis (haven demand) and in a US-specific boom (yield/growth demand), and weakens in the synchronised-growth middle. *(Chapter 14.)*

**Duration (modified duration)** The sensitivity of a bond's price to a small yield change, $dP/P = -D\,dy$; the channel through which a rate differential move is also a bond-price move. *(Chapter 13.)*

**Effective spread** Twice the gap between the actual execution price and the prevailing mid at the moment of the trade; what a trade actually cost relative to the decision price. *(Chapters 2, 3.)*

**Ex-date (ex-dividend date)** The first date a share trades without entitlement to a declared dividend; the price mechanically drops by approximately the dividend on this date. *(Chapter 8.)*

**Exercise (option exercise)** Invoking an option's right to buy (call) or sell (put) the underlying at the strike; American style allows this before expiry, European style only at it. *(Chapter 11.)*

**Expiry** The date an option's or futures contract's rights and obligations lapse or are settled. *(Chapters 10, 11.)*

**Fair value (cost of carry)** The no-arbitrage price a future or forward must trade at given the spot price, the financing rate, and any yield the underlying pays: $F = S(1 + (r - q)T)$ for a financial underlying. *(Chapter 9.)*

**Financing (overnight financing)** The daily charge or credit a leveraged CFD position accrues past the daily cut-off: notional times a rate differential (or benchmark rate) plus an administrative markup. *(Chapters 3, 6.)*

**Fix (the London 4pm fix)** A benchmark FX rate computed from trades executed in a short window around a fixed time, used to value and execute large flows at one common, published rate. *(Chapter 4.)*

**Flash crash** A sudden, severe, rapidly-reversed price dislocation driven by a rapid withdrawal of liquidity rather than new information; 6 May 2010 is the canonical case. *(Chapter 15.)*

**Forward points (swap points)** The pip difference between the forward rate and the spot rate for a tenor, equal to the interest-rate differential expressed in price terms. *(Chapters 4, 5, 6.)*

**Franking credit** An Australian tax credit attached to a dividend, certifying company tax already paid on the underlying profit; a shareholder attribute, not part of the ex-date price mechanics itself. *(Chapter 8.)*

**Funding rate (perpetual funding)** The periodic payment perpetual-future longs make to shorts (or vice versa) that anchors the perpetual's price to spot. *(Chapter 16.)*

**Gamma flows** The mechanical hedging trades option market makers must transact as spot moves, driven by the sign of their aggregate gamma position: long dealer gamma dampens moves, short dealer gamma amplifies them. *(Chapter 11.)*

**Gap** A jump between one price and the next with no trading in between, typically across a session close, a weekend, or a halt. *(Chapters 4, 9.)*

**Guaranteed stop** A stop order a broker contractually fills at its stated level regardless of the price the market actually gaps to, priced as insurance via a premium or a wider spread. *(Chapter 15.)*

**Haven (haven currency/asset)** An instrument that tends to appreciate in risk-off episodes because participants seek its liquidity or funding role rather than its yield; JPY, CHF, USD and, conditionally, gold are the standing examples. *(Chapters 7, 14.)*

**Hedger** A participant holding, or about to hold, a business-driven exposure who trades to reduce its variance, generally accepting a worse expected price in exchange. *(Chapter 1.)*

**HFT (high-frequency trading)** Electronic market making and arbitrage conducted at very short holding periods and high message rates; now a dominant share of quoted liquidity on most electronic venues. *(Chapters 1, 2.)*

**Immediacy** The service a liquidity supplier sells by continuously standing ready to trade; the economic reason a spread exists at all. *(Chapters 1, 2.)*

**Implementation shortfall** The total gap between the price that motivated a trading decision and the average price actually achieved, including the opportunity cost of any unfilled portion. *(Chapter 3.)*

**Implied volatility (IV)** The volatility level that, put into an option pricing model, reproduces the option's traded premium; the quantity dealers actually think and quote in. *(Chapter 11.)*

**Index divisor** The index-construction constant keeping a published index level continuous through a corporate action or constituent change, reset at the event rather than at every session. *(Chapter 8.)*

**Internalisation** A broker or wholesaler filling client order flow from its own book or inventory rather than routing it to a lit exchange. *(Chapters 2, 4, 8.)*

**Inventory risk** The risk a liquidity supplier bears from holding an unwanted position while waiting to lay it off; scales with volatility and the expected time to flatten. *(Chapters 1, 2.)*

**Kyle's lambda** The coefficient of linear price impact per unit of net signed order flow; its reciprocal is the market's depth. *(Chapter 2.)*

**Last look** A dealing-model practice in which a liquidity provider reviews a trade request against its own price before confirming or rejecting the fill. *(Chapter 6.)*

**Leverage** The ratio of notional position size to the margin or capital actually posted; retail caps are set by regulation, vary by jurisdiction, and drift. *(Chapters 4, 6.)*

**Liquidity clock** The recurring daily and weekly pattern of thin and deep liquidity a continuous market exhibits, driven by which trading centres are open and overlapping at a given moment. *(Chapter 4.)*

**Liquidity spiral (margin spiral / loss spiral)** The mechanism by which a forced deleveraging tightens the funding conditions that force further deleveraging. *(Chapters 7, 14.)*

**Locate** The regulatory requirement that a broker have a reasonable basis for believing shares can be borrowed and delivered before executing a short sale. *(Chapter 8.)*

**Lot** The standard FX position-size unit: 100,000 units of base currency (standard), 10,000 (mini), 1,000 (micro). *(Chapter 4.)*

**Manufactured dividend** The cash payment a short seller owes the lender of borrowed shares on an ex-date, equal to the dividend the shares would otherwise have earned. *(Chapter 8.)*

**Margin call** A broker's demand for additional funds when posted margin falls below the required maintenance level on a leveraged position. *(Chapters 6, 10.)*

**Market maker** A participant that continuously quotes both a bid and an ask and stands ready to trade either

direction, earning the spread across many trades net of inventory and adverse-selection costs. *(Chapters 1, 2.)*

**Microprice** The quantity-weighted average of the best bid and ask, weighted by the *opposite* side's size; a finer fair-value estimate than the simple mid. *(Chapter 2.)*

**NBBO (National Best Bid and Offer)** The US linkage regime requiring listed venues to honour the best displayed price across all of them. *(Chapter 8.)*

**NDF (non-deliverable forward)** A forward on a restricted or non-convertible currency that cash-settles the rate difference at maturity rather than delivering the currency itself. *(Chapter 4.)*

**Open interest** The number of futures or options contracts currently outstanding; distinct from volume, which measures turnover. *(Chapter 10.)*

**Order flow (signed order flow)** The net direction and size of executed trades, signed positive for buyer-initiated and negative for seller-initiated. *(Chapter 7.)*

**Order-processing cost** The per-trade operating cost of running a continuously quoting desk; one of the three components of the quoted spread. *(Chapter 2.)*

**Peg (currency peg)** A policy commitment to hold an exchange rate at or near a fixed level, defended by central bank intervention; transfers tail risk rather than removing it. *(Chapter 15.)*

**Perpetual (perpetual future)** An exchange-traded futures-like contract with no expiry, kept close to spot by the funding rate. *(Chapter 16.)*

**Pip** The standard FX quoting increment: 0.0001 of the quote currency for most pairs, 0.01 for pairs quoted in Japanese yen. *(Chapter 4.)*

**Pipette** One tenth of a pip, the finer increment brokers stream prices in. *(Chapter 4.)*

**Point-in-time discipline** Joining any data series to a model on the timestamp it became publicly knowable, never on the date it describes. *(Chapter 7.)*

**Prime broker** An intermediary extending its own credit and settlement relationship to a client, letting that client trade with many dealers while every trade settles against the prime broker alone. *(Chapter 4.)*

**Price impact** The permanent component of a trade's cost: the difference between the post-trade mid and the pre-trade mid, distinct from the payment made to the liquidity supplier. *(Chapters 2, 3.)*

**Quoting conventions (base/quote, direct/indirect)** The market's rule for which currency is priced per unit of the other, and the American-terms/European-terms labelling built on the US dollar's role as vehicle currency. *(Chapter 4.)*

**Realised spread** The effective spread minus price impact; what a liquidity supplier actually keeps after the market has moved following the trade. *(Chapters 2, 3.)*

**Reference Tuesday / release Friday** The COT report's data-as-of date and its earliest possible publication date; the origin of its structural reporting lag. *(Chapter 7.)*

**Roll (futures roll, roll yield)** Closing an expiring futures position and opening the equivalent position in a later expiry to maintain continuous exposure; the resulting gain or drag, depending on the curve's shape, is roll yield. *(Chapters 10, 12.)*

**Rollover** The nightly process, conventionally anchored at 5pm New York, by which an open CFD position's implicit value date is advanced and financing is assessed. *(Chapters 3, 4, 6.)*

**RORO (risk-on/risk-off)** A cross-asset co-movement regime, not a tradeable signal in itself, describing how havens, growth-sensitive currencies, equities and, conditionally, gold move together under shifts in global risk appetite. *(Chapter 7.)*

**Seasonality** A recurring calendar pattern in price or demand driven by a genuine physical cycle; real in some commodities, weak to absent in others such as mined metals. *(Chapter 12.)*

**Session** A market's defined trading hours; equities and listed derivatives close daily and reopen with an auction, while FX and crypto trade continuously with a liquidity-thinning pattern rather than a scheduled close. *(Chapters 1, 4.)*

**Settlement** The point at which the two sides of a trade actually exchange value (or, for a CFD, never do, replaced by daily financing); conventions vary from T+1 to T+2 by market. *(Chapters 4, 8.)*

**Skew (volatility skew)** The pattern of implied volatility across strikes at a given expiry; index puts persis-

tently bid above calls is the standing equity-index example, read as crash-insurance demand. *(Chapter 11.)*

**Slippage** The difference between an expected execution price and the price actually achieved; distinct from, and not fully explained by, the quoted spread alone. *(Chapters 2, 3.)*

**Spot** The shortest standard settlement convention a market quotes (T+2 for most FX pairs), not a claim of instantaneous delivery. *(Chapter 4.)*

**Spread decomposition** The division of the quoted spread into order-processing cost, inventory risk and adverse selection, each responding to different conditions. *(Chapters 1, 2.)*

**Square-root (impact) law** An empirical rule of thumb that price impact scales roughly with the square root of order size relative to volume; documented at institutional equity size, not calibrated for small retail CFD trades. *(Chapter 3.)*

**Stablecoin** A crypto-asset designed to hold a fixed value against a reference currency, typically USD; the structural quote leg for most crypto trading pairs. *(Chapter 16.)*

**Stop order** An order that sits inactive until a trigger price trades, then converts to a market (or limit) order; a resting stop is filled at the first available price after a gap, not guaranteed at its stated level. *(Chapters 1, 15.)*

**Storage (theory of storage)** The framework explaining a commodity futures curve's shape from the physical cost of storing the underlying against the convenience yield of holding it. *(Chapter 12.)*

**Tick** The minimum price increment a venue allows; for a futures or options contract, a fixed monetary tick value converts one tick of price movement into a dollar amount. *(Chapters 2, 10.)*

**Time-in-force** An order qualifier governing how long it remains live: day, good-till- cancelled, immediate-or-cancel, or fill-or-kill. *(Chapter 1.)*

**Total-return index** An index variant reinvesting every constituent dividend on the ex-date, so no permanent ex-date drop appears in the level; not the object an index CFD is priced off. *(Chapter 9.)*

**Touch** The best bid and best ask together, also called the BBO; the two prices that must satisfy bid $<$ ask. *(Chapter 2.)*

**Triangulation** The identity chaining two USD-quoted pairs through their shared leg to derive, and arbitrage-pin, the cross rate between the other two currencies. *(Chapter 4.)*

**UIP (uncovered interest parity)** The hypothesis that expected currency depreciation offsets the interest-rate differential exactly; its systematic empirical failure is what the carry trade harvests. *(Chapter 5.)*

**Uncrossing** The mechanism by which a call auction's accumulated buy and sell orders are matched at the single price maximising executable volume. *(Chapters 1, 2, 8.)*

**Value date** The calendar date on which the two currencies in an FX transaction actually change hands; T+2 for spot in most pairs. *(Chapter 4.)*

**Variation margin** The daily cash settlement of a futures position's mark-to-market gain or loss; a realised cash flow, not paper profit or loss. *(Chapter 10.)*

**Vega** An option's sensitivity to a one-point change in implied volatility; the conversion factor between a dollar premium move and a vol-point move. *(Chapter 11.)*

**VIX** The Cboe's index of 30-day implied volatility on the S&P 500, constructed from a strip of listed option prices. *(Chapter 11.)*

**Volume** The quantity of contracts or shares actually traded over a period, distinct from open interest, which measures positions outstanding. *(Chapter 10.)*

**Weighting scheme (index)** The rule — price-weighted, cap-weighted, float-adjusted — by which an index combines constituent prices into one level. *(Chapter 8.)*

**Z-score (positioning)** A trader category's current net position expressed in standard deviations from its own trailing average; the standard way of reading positioning "extremes" comparably across time. *(Chapter 7.)*

C

# Annotated References

*Every work REFERENCES3.md whitelists for inline citation, regrouped by the five parts of this book rather than by subject, with a one-line note on why each earns its place here. Nothing below is new: this is REFERENCES3.md itself, reformatted for the desk. A reference used across several parts is listed once, under the part it does the most load-bearing work in.*

## Part I — What a Market Is

**Harris (2003).** *Trading and Exchanges: Market Microstructure for Practitioners.* Oxford UP. — The standing citation for order types, auction clearing and session structure; cited throughout Chapters 1, 2 and 8.

**O'Hara (1995).** *Market Microstructure Theory.* Blackwell. — General theoretical backing for the three-component spread decomposition of Chapter 2.

**Hasbrouck (2007).** *Empirical Market Microstructure.* Oxford UP. — The empirical decomposition of trading cost into realised spread and price impact used in Chapters 2 and 3.

**Kyle (1985).** Continuous Auctions and Insider Trading. *Econometrica* 53. — Linear price impact per unit of order flow; the $\lambda$ that prices size in Chapter 2.

**Glosten, Milgrom (1985).** Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders. *J. Financial Economics* 14. — The adverse-selection model deriving the breakeven spread from an informed-trader probability, worked numerically in Chapters 1 and 2.

**Amihud, Mendelson (1986).** Asset Pricing and the Bid-Ask Spread. *J. Financial Economics* 17. — Illiquidity priced in the cross-section of expected returns, cited alongside the spread's order-processing and inventory components in Chapter 2.

**Roll (1984).** A Simple Implicit Measure of the Effective Bid-Ask Spread. *J. Finance* 39. — Recovers the effective spread from the serial covariance of transaction prices alone; derived and worked in full in Chapter 2.

**Almgren, Chriss (2001).** Optimal Execution of Portfolio Transactions. *J. Risk* 3. — The temporary/permanent impact trade-off framing execution scheduling in Chapter 3 and the liquidity screen for single-name pairs legs in Chapter 9.

**Perold (1988).** The Implementation Shortfall: Paper versus Reality. *J. Portfolio Management* 14. — The full decision-to-fill cost decomposition derived and worked in Chapter 3.

**Duffie (2012).** *Dark Markets: Asset Pricing and Information Transmission in Over-the-Counter Markets.* Princeton UP. — The search-and-bargain account of OTC market structure underlying every non-exchange venue in this book, from Chapter 1's exchange/OTC taxonomy to Chapter 4's dealer tiers.

**Menkveld (2013).** High Frequency Trading and the New Market Makers. *J. Financial Markets* 16.

— Electronic market making as the modern liquidity supplier, cited in Chapters 1 and 2 wherever "dealer" becomes "electronic dealer".

## Part II — The Foreign Exchange Market

**Lyons (2001).** *The Microstructure Approach to Exchange Rates.* MIT Press. — Order flow as a price-relevant object in FX; the frame Chapters 1 and 4 build the dealer-tier structure on.

**Evans, Lyons (2002).** Order Flow and Exchange Rate Dynamics. *J. Political Economy* 110. — Signed interdealer order flow explains a large share of daily FX moves; the central empirical result recapped in Chapters 4 and 7.

**BIS (2022).** Triennial Central Bank Survey of Foreign Exchange and OTC Derivatives Markets. (figures indicative, as-of vintage) — The only methodologically consistent measure of global FX turnover; every scale figure drawn from it in Chapter 4 carries its April 2022 vintage explicitly.

**Fama (1984).** Forward and Spot Exchange Rates. *J. Monetary Economics* 14. — The forward-premium puzzle: the empirical failure of uncovered interest parity that Chapter 5 builds the carry trade on.

**Meese, Rogoff (1983).** Empirical Exchange Rate Models of the Seventies. *J. International Economics* 14. — Structural exchange-rate models fail to reliably out-forecast a random walk at short horizons; the honesty check Chapter 5 applies to its own economic drivers.

**Menkhoff, Sarno, Schmeling, Schrimpf (2012).** Carry Trades and Global Foreign Exchange Volatility. *J. Finance* 67. — Carry returns turn sharply negative exactly in high-volatility regimes; the volatility-interaction result Chapters 5 and 7 tie to the platform's RORO/VolRegime covariates.

**Lustig, Roussanov, Verdelhan (2011).** Common Risk Factors in Currency Markets. *RFS* 24. — The cross-sectional counterpart to carry crash risk, cited alongside Fama (1984) in Chapter 5.

**Brunnermeier, Nagel, Pedersen (2008).** Carry Trades and Currency Crashes. *NBER Macro Annual* 23. — "Up the stairs, down the lift": the crowded-positioning unwind mechanism Chapters 5 and 7 both invoke, and one of three literature results Chapter 14 ties directly to carry's negative skew.

**Sarno, Taylor (2001).** Official Intervention in the Foreign Exchange Market. *J. Economic Literature* 39. — The catalogue of central bank intervention styles and their effectiveness used in Chapters 4 and 5.

**Rime, Schrimpf (2013).** The Anatomy of the Global FX Market through the Lens of the 2013 Triennial Survey. *BIS Quarterly Review.* — The FX market's tiered dealer anatomy, from inside the BIS survey; the structural backbone of Chapters 1 and 4.

**CFTC.** Commitments of Traders: Explanatory Notes. (official documentation; structure stable) — The primary-source definition of every COT category, threshold and calendar rule Chapter 7 states as fact.

## Part III — Equities and Indices

**Shleifer (1986).** Do Demand Curves for Stocks Slope Down? *J. Finance* 41. — Index inclusion moves price; the demand-curve evidence behind Chapter 8's treatment of rebalance flows.

**Harris, Gurel (1986).** Price and Volume Effects Associated with Changes in the S&P 500. *J. Finance* 41. — The companion event-study evidence for the same index-effect discussion in Chapter 8.

**Fama (1970).** Efficient Capital Markets. *J. Finance* 25. — Efficiency as a testable hypothesis, and the benchmark every named anomaly in Chapter 8's information-cycle section is measured against.

**French (1980).** Stock Returns and the Weekend Effect. *J. Financial Economics* 8. — A documented, re-tested calendar anomaly, cited in Chapter 8 as the "test, do not trust" family's second standing

example.

## Part IV — Derivatives and Commodities

**Hull (2021).** *Options, Futures, and Other Derivatives*, 11th ed. Pearson. — The standing reference for derivatives pricing structure this book assumes and does not re-derive, used directly for Chapter 9's cost-of-carry relationship and referenced throughout Chapters 10 and 11.

**Natenberg (1994).** *Option Volatility and Pricing.* McGraw-Hill. — The practitioner's volatility-trading vocabulary; whitelisted for Chapter 11's market-maker and hedging-flow territory, though the finished chapter is not found to cite it by name.

**Sinclair (2013).** *Volatility Trading*, 2nd ed. Wiley. — Volatility trading as a systematic discipline; whitelisted for Chapter 11's vol-trader workflow section, though the finished chapter is not found to cite it by name.

**Gatheral (2006).** *The Volatility Surface.* Wiley. — The implied-volatility surface as an object with its own structure (skew, term structure), the frame Chapter 11 reads a chain against.

**Cboe (2019).** VIX White Paper: Cboe Volatility Index. (methodology; figures indicative) — The VIX construction Chapter 11 describes structurally as the index of 30-day S&P 500 implied volatility.

**Black (1976).** The Pricing of Commodity Contracts. *J. Financial Economics* 3. — Prices a derivative on the forward/futures price itself; the convention Chapter 10 uses to price off the curve without separately identifying storage cost and convenience yield, and that Chapter 15's negative-WTI case study shows breaking down once the underlying itself turns negative.

**Keynes (1930).** *A Treatise on Money*, Vol. II (normal backwardation). Macmillan. — The hedging-pressure reading of the futures curve, cited in Chapter 7's positioning discussion and derived properly in Chapter 12.

**Working (1949).** The Theory of Price of Storage. *American Economic Review* 39. — The storage-cost/convenience-yield framework Chapter 12 uses to explain contango and backwardation.

**Gorton, Rouwenhorst (2006).** Facts and Fantasies about Commodity Futures. *Financial Analysts Journal* 62. — Collateralised commodity futures as a priced asset class; one side of Chapter 12's honest return-decomposition debate.

**Erb, Harvey (2006).** The Strategic and Tactical Value of Commodity Futures. *Financial Analysts Journal* 62. — The roll-yield critique of naive commodity-index returns; the other side of the same Chapter 12 debate.

## Part V — The System View

**Bernanke, Kuttner (2005).** What Explains the Stock Market's Reaction to Federal Reserve Policy? *J. Finance* 60. — Only the policy surprise, not the level, moves asset prices; derived in Chapter 5 and generalised across pillars in Chapter 13.

**Gurkaynak, Sack, Swanson (2005).** Do Actions Speak Louder Than Words? *International J. Central Banking* 1. — Target and path factors in monetary policy surprises, used alongside Bernanke-Kuttner in Chapters 5 and 13.

**Rey (2015).** Dilemma not Trilemma: The Global Financial Cycle and Monetary Policy Independence. *NBER WP* 21162. — A common global factor dominating local monetary independence; the macro-level statement of the RORO mechanism in Chapters 7 and 14.

**Campbell, Shiller (1991).** Yield Spreads and Interest Rate Movements. *Review of Economic Studies* 58. — The expectations-hypothesis reading of the yield curve underlying Chapter 13's forward-

rate arithmetic.

**Brady Commission (1988).** Report of the Presidential Task Force on Market Mechanisms (1987 crash). — The official account of portfolio insurance's mechanical selling feedback; the origin of the circuit breakers described in Chapter 8 and the first case study in Chapter 15.

**Lowenstein (2000).** *When Genius Failed: The Rise and Fall of Long-Term Capital Management.* Random House. — Leverage and correlated positioning turning being right into being insolvent first; Chapter 15's second case study.

**Brunnermeier (2009).** Deciphering the Liquidity and Credit Crunch 2007–2008. *J. Economic Perspectives* 23. — The funding-crisis anatomy underlying Chapter 14's liquidity-spiral discussion and Chapter 15's 2007–08 case study.

**Brunnermeier, Pedersen (2009).** Market Liquidity and Funding Liquidity. *RFS* 22. — The formal margin-spiral/loss-spiral mechanism cited in Chapter 7's crowding discussion and derived in full in Chapter 14.

**CFTC-SEC (2010).** Findings Regarding the Market Events of May 6, 2010 (flash crash report). — The official account behind Chapter 8's limit-up/limit-down discussion and Chapter 15's flash crash case study.

**Kirilenko, Kyle, Samadi, Tuzun (2017).** The Flash Crash: High-Frequency Trading in an Electronic Market. *J. Finance* 72. — The trade-level mechanics of the same episode, paired with CFTC-SEC (2010) in Chapters 8 and 15.

**Swiss National Bank (2015).** Press release of 15 January 2015 discontinuing the minimum exchange rate. (public record) — The primary source for Chapter 15's centrepiece case study, the CHF depeg.

**BIS (2020).** Bulletins on the March 2020 market turmoil. (as-of vintage) — The dash-for-cash account behind Chapter 7's correlation-instability warning, Chapter 12's negative-WTI mechanics, and Chapter 15's sixth case study.

**Nakamoto (2008).** Bitcoin: A Peer-to-Peer Electronic Cash System. (whitepaper) — The origin document for the asset class underlying the platform's two crypto instruments, cited at Chapter 16's opening and again in its regulatory-drift discussion.

### Citation mechanics across the three volumes

Two further notes from REFERENCES3.md govern citation practice rather than naming a single work, and are preserved here rather than folded into either list above. First, the Course Book (this library, 2026) is cited throughout this volume via the gold [Course Book §X.Y] chip, never as an author-year pair; every chip in Chapters 1–16 has been checked against the Course Book's own table of contents. Second, the six 2026 programme-brief arXiv identifiers this library draws on may be cited, where they appear, as "cited in the programme brief" — the same convention the Course Book itself uses, and not repeated as separate whitelist entries here.

# D

# The Platform's Market Map

*Twenty-one OANDA CFD instruments, one table: where each is treated in this book, how its session clock behaves, what the cost surface must measure for it, what mechanics beyond price it carries, and which programme version trades it. A closing note places this volume against its two companions.*

**Your project — Your project — the whole book, one table**

Every row below is one of the 21 instruments the SST platform actually holds: 14 FX pairs, XAU, XAG, BTC, ETH, two index CFDs, and WTI (Chapter 1). The table compresses fifteen chapters of structure into the form a build actually needs — which chapter to reopen when a mechanic is unclear, not a substitute for having read it once.

**Table D.1.** The platform's 21 instruments, by class: chapter(s), session/clock treatment, cost-surface notes, special mechanics, and programme version.

| Instrument | Ch. | Session/clock | Cost surface | Special mechanics | Version |
|---|---|---|---|---|---|
| FX majors (e.g. EURUSD, GBPUSD, AUDUSD, NZDUSD, USDCAD, USDCHF; illustrative, not the full roster) | 4, 5, 6, 7 | Continuous week, Sun 22:00–Fri 22:00 UTC (§A.2); 4 liquidity buckets + rollover/weekend flags | Session-bucketed spread surface; widens at news/rollover/thin hours | Nightly rollover; T+2 value date; Wednesday triple swap (§A.3) | v1, primary |
| FX crosses (e.g. EURGBP, AUDNZD; illustrative) | 4, 5, 6, 7 | As majors | As majors; baseline spread typically wider | As majors; pip value needs triangulation when the account currency is neither leg | v1, primary |
| FX JPY pairs (e.g. USDJPY, EURJPY, GBPJPY and any other JPY cross the book carries) | 4, 5, 6, 7 | As majors | As majors | As majors; 0.01 pip convention (§A.1) | v1, primary |
| XAU (gold) | 12 | Quoted like an FX pair vs USD; thins at FX rollover and weekend | Session- bucketed surface extended to metals cells | Financing embeds lease/carry; real-yield and haven driver (Chapters 7, 13, 14) | v4, commodities |
| XAG (silver) | 12 | As XAU | As XAU | As XAU; materially higher indicative vol; dual monetary/industrial character | v4, commodities |

continued overleaf

| Instrument | Ch. | Session/clock | Cost surface | Special mechanics | Version |
|---|---|---|---|---|---|
| BTC | 16 | True 24/7, no weekly close; thinner weekend liquidity (documented pattern) | Dedicated crypto session cells, no FX-style rollover clock | Vol-targeted sizing absorbs the higher vol without special-casing; funding-rate mechanics apply if the CFD references a perpetual (§A.7) | v1* |
| ETH | 16 | As BTC | As BTC | As BTC | v1* |
| Index CFDs (two, generic) | 8, 9 | Near-continuous off the future; cash session (open/-close auctions) nested inside it; cash-open bar flagged separately | Sessioned by cash-hours vs overlapping-but-outside-cash vs low-liquidity windows | Dividend credit/debit on ex-dates; fair-value/basis vs future; overnight gap decomposition (§A.6) | v3* |
| WTI | 10, 12 | Futures-referenced; roll calendar sets the "front" contract | Roll-aware cost accounting; provider absorbs the mechanical roll | Provider roll policy; margin/spread regime shifts around expiry (negative front-month episode, Chapter 15) | v4, commodities |

*Version tags for BTC, ETH and the two index CFDs are inferred from cross-chapter context — Chapter 16's claim that the platform's vol-targeting machinery handles crypto without special-casing; Chapter 9's framing of the index CFDs alongside the v3 equity-pairs programme — rather than stated verbatim in the platform canon paragraph. The canon names four versions (v1 FX primary, v2 AU/US options, v3 equity pairs, v4 commodities); v2 has no row here because AU/US listed options trade on separate exchange venues, not as one of the platform's 21 OANDA CFDs. Confirm all three flagged tags against the live programme documentation.

## How to use these three volumes together

The three volumes this library comprises answer three different questions about the same platform, and confusing which volume answers which is the most common way to waste an afternoon. The Course Book *derives*: it builds, from stated assumptions to proved conclusions, the statistical and financial machinery the platform's models actually run — distributional forecasting, calibration, the deflation-adjusted evaluation gate, the pairs-trading statistics, the options-pricing engine, the leakage taxonomy that makes any of the above trustworthy. Open it when the question is *why does this formula take this form, and what does it assume*. The Playbook *executes*: it is the concrete, versioned build — the actual point-in-time data pipeline, the cost-surface calibration, the Stage −1..4 programme, the code turning Course Book theory into a running system trading 21 real instruments. Open it when the question is *what does the platform actually do, right now, and how is it built*. This volume is the *terrain*: it describes the markets themselves — how an order book matches, why a spread exists, what a COT report actually contains, how a rollover charge is computed, what a crisis does to correlation — independent of any model or any platform, the way the ground exists independent of whoever is walking on it.

The three are ordered here by how directly they touch the platform's own code, not by importance. A model can be perfectly derived and perfectly implemented and still lose money if it is blind to the terrain this volume maps: a spread is a market maker's compensation for adverse selection, not a friction to be minimised away by a sharper entry rule; a COT report is three days stale by construction, not a live feed; a CHF-style peg's break is a tail no amount of correct statistics would have priced, because a peg transfers tail risk rather than removing it. Equally, terrain knowledge alone trades nothing: knowing that carry crashes "up the stairs, down the lift" is not a position, a sizing rule, or a stop — it becomes tradeable only once the Course Book's distributional and evaluation machinery and the Playbook's actual execution stack are both built around it.

Read the shelf in order when learning; consult in whichever order the question demands once built. This appendix's own four sections — conventions, glossary, references, map — are written for the second mode: open at the table, not at the beginning, and let the chapter pointer send you back to the argument only when the quick answer is not enough.

**Key idea — What this back matter adds to the platform**

Sixteen chapters and four appendices have covered the same ground the platform trades every day, from a different angle: not what to predict, but what the thing being predicted actually is. Costs are not friction to be modelled away; they are compensation for a real economic function, and understanding whose function explains why they cluster exactly where the platform's own cost surface finds them. Positioning, dividends and rollovers — Wednesday's triple included — are not noise to be filtered; they are structure with a clock and a rule, knowable in advance to anyone who reads this book's tables rather than assumes them. And every crisis this book named worked the same handful of mechanisms Chapter 14 derived, at full amplitude — the reason the platform carries a stressed cost surface, CVaR sizing and a regime router is not caution for its own sake, but the terrain itself insisting on it. Keep this back matter open; the rest of the shelf assumes you have.