Chapter 22

Relative Value and the Basis Trade

In these notes we look at what the other side of the industry does with the same machinery. A relative value desk is not trying to price a derivative correctly; it is trying to find two things that should be worth the same and are not. We derive the decomposition of a fixed income return into carry, roll-down and yield change, construct the curve trades that isolate one factor from another, and then work through the cash-futures basis in enough detail to see both why the trade exists and why it periodically destroys the people doing it.

22.1 What This Chapter Is

Everything so far has been derivation. A model was posited, a measure found, a price computed, and the result was true given the assumptions. This chapter describes trades that people put on to make money.

The subject here is not whether these trades work. It is what they *are*: what a position consists of, how its profit and loss decomposes, and — the part that is genuinely mathematical and genuinely useful — where the risk actually sits, which is frequently not where it appears to sit. A trade whose P&L looks like a small steady income and whose risk is a rare enormous loss is a recognisable object, and recognising it is worth more than any view about whether it is currently cheap.

The reader who wants to know whether to put a trade on will not find it here. The reader who wants to know what they are holding will.

Structure (What makes a spread a trade)**.** A model tells you a spread is wide. That is not a signal, and treating it as one is the characteristic error of the subject. Two further things are required before a wide spread is a trade, and neither is a modelling output.

A structural reason. Somebody has to be on the other side for a reason that is not an opinion — a regulatory constraint that forces a pension fund to hold a particular instrument, an index rule that requires a bond to be sold on a date, a balance sheet cost that makes a dealer unwilling to warehouse a position, a settlement convention that nobody can arbitrage away. When such a reason exists, the spread is a price being paid for a service and can be collected. When it does not, the wide spread is more likely to be the model's error than the market's, and §22.8's negative number is the case where telling the two apart is the whole problem.

A horizon. Convergence is not a date. A spread that is two standard deviations wide and genuinely mean reverting still takes longer to close than the intuition suggests, and goes further against the position first. That is a question about first passage times rather than about forecasting, it is answerable, and the next section answers it — because the usual way of getting it wrong is not misjudging direction but misjudging how long being right takes.

The model's job in all of this is narrower than it looks. It supplies the spread and the decomposition of the P&L; it has nothing to say about either of the two conditions above. Which is why the discipline in relative value work lives in chapter 20's estimation rather than in the pricing.

22.2 How Long Being Right Takes

Take the simplest possible model of a converging spread: an Ornstein-Uhlenbeck process, as in chapter 20,

$$dX_t = -\kappa X_t\,dt + \sigma\,dW_t, \tag{22.1}$$

entered when X sits two stationary standard deviations from zero, held until it returns. The mean reversion is certain — this is not a case where the trade might be wrong about direction. Everything below is what happens when it is right.

Calculation 22.1 (First passage is not the half-life)**.** The half-life $\ln 2/\kappa$ is the number everyone quotes and it answers a different question. It says how fast an *expectation* decays: $\mathbb{E}[X_t] = X_0 e^{-\kappa t}$ halves in that time. It does not say how long a path takes to reach zero, and the two are not close.

Simulating (22.1) from two standard deviations, with a one-year half-life:[1]

half-life $\ln 2/\kappa$	1.00 years
mean time to first reach the mean	2.10 years
mean worst level reached first, in deviations	2.48

So the trade takes more than twice the half-life, and before converging it goes about half a deviation further against the position than the level it was entered at. Neither number is available from the half-life, and both are what size the trade.

Remark (Where a stop turns a winning trade into a loss)**.** The mean worst excursion is an average, so a stop placed at it is hit about half the time. Placing it closer is worse:

Stop, in deviations	Fraction stopped out before converging
2.5	39%
3.0	12%
4.0	under 1%

Entered at two deviations, a stop half a deviation away loses a trade that was right about direction two times in five. At twice the entry width the stop stops binding altogether, and the trade's risk becomes the holding period rather than the loss.

That is the regime to be in, and reaching it is a sizing decision rather than a stop-placement decision: the position has to be small enough that sitting through twice the entry width is tolerable. A desk that sizes to its stop rather than to its horizon has built a trade that loses when it is right.

[1] `estimation::ConvergenceTrade`.

Structure (The same first passage problem as the doubling strategy)**.** This is chapter 4's doubling strategy in respectable clothing.

There, a strategy on a martingale was certain to reach its target and the drawdown before it had a tail so heavy that its mean was infinite — which is why admissibility has to bound the loss rather than its expectation. Here the process is mean reverting rather than a martingale, so the tail is far lighter and the mean excursion is finite. But the structure is the same: a bet that is certain to win, whose risk lives entirely in what happens before it does, and whose sizing is therefore governed by a first passage distribution rather than by an expected return.

The difference is what makes relative value a business and doubling a fallacy. Mean reversion makes the excursion's distribution thin enough to survive with finite capital, and (22.1)'s κ is exactly what controls how thin. Which puts an uncomfortable amount of weight on knowing κ.

Remark (And κ is the parameter that is estimated badly)**.** Chapter 20 measures that mean reversion estimated from a finite sample comes out too fast, and chapter 14 explains why from the spectrum: a deviation decays more quickly than the slowest mode at first, so fitting one exponential to a whole sample returns a rate above the true gap. The bias is one-signed.

Follow that through to the trade. An overstated κ is an understated half-life, which is an understated horizon. With the true half-life at one year and the estimate thirty per cent fast — comfortably inside what a decade of data delivers — the planned holding period is 1.63 years against a realised 2.10.[2] The desk expects to be out in twenty months and is still in the position at twenty-five.

Every part of that error points the same way. The horizon is longer than planned, so the funding cost is larger than budgeted; the position is held through more of the excursion distribution, so the drawdown is deeper than modelled; and the capital is committed longer, so the return on it is lower than advertised. There is no compensating error in the other direction.

22.3 Whether the Spread Converges at All

The section above began by granting that the mean reversion is certain, and everything in it followed from that. What happens when it is not granted?

Chapter 20 has already ruled out the obvious approach. Fitting (22.1) to a history returns a positive $\hat{\kappa}$ whether or not there is any mean reversion to find — on a random walk the estimate is positive with probability approaching one — so "the regression found convergence" establishes nothing. What is needed is not a better estimate but a *test*, with the random walk as the null hypothesis rather than as an alternative nobody considered.

That is the Dickey-Fuller test, and it is the natural tool here rather than an imported one: regress the change on the level, and ask whether the coefficient is distinguishable from zero. The one subtlety is that the usual critical value does not apply, because under the null the regressor is a random walk and the t-ratio is not t distributed; the five per cent value is about -2.86 instead of -1.65.[3]

Calculation 22.2 (What the test can actually see)**.** The test is correctly sized — a genuine random walk passes for mean reverting about five per cent of the time, as asked. The question is the other error: how often does the test find mean reversion that is really there?

[2] `measured`.

[3] `estimation::unit_root_test`.

Half-life	**1 year**	**3 years**	**5 years**	**10 years**
6 months	6%	11%	17%	52%
1 year	6%	7%	8%	18%
2 years	6%	5%	6%	8%

Read the middle row. A spread that genuinely reverts with a one-year half-life, watched for five years, is identified as reverting eight times in a hundred.[4] The test is not broken; it is being asked to separate two hypotheses that a sample of that length barely distinguishes.

And the pattern across the table is the same shape chapter 20 found for the estimation bias. Power depends on the number of *half-lives* the sample spans and not on its length or its resolution: a one-year half-life over five years and a two-year half-life over ten give the same answer, though one history is twice the other.[5] Sampling the spread hourly would add observations and no power at all.

Structure (Which is why the structural condition does the work)**.** Put Calculation 22.2 beside the two conditions this chapter opened with and the two halves of its argument fit together.

A test that finds real convergence eight times in a hundred cannot be what a desk relies on. Worse, it is not merely uninformative but dangerous under search: given n instruments there are many combinations to weight, and looking through enough of them will produce one that passes at five per cent whether or not anything converges. A screen that tests a thousand spreads finds fifty by construction. The statistic that was weak as evidence becomes actively misleading as a filter.

So the structural reason demanded earlier is not a piece of good practice sitting alongside the statistics. It is what the statistics cannot supply. A butterfly is a curvature by construction, a cash-futures basis is tied to financing by an arbitrage that must close at delivery, an on-the-run spread is a liquidity premium with a named mechanism — and each of those is a reason to expect convergence that does not come from the sample, so it is not consumed by having searched the sample. The test is then a check on a prior rather than a way of forming one, which is the only role its power supports.

That is also the honest difference between this and the equity pairs trading the same mathematics is usually taught with. There the combination is typically *discovered* by search over a universe, and the discovery is exactly the procedure Calculation 22.2 says will manufacture false positives. Here the combination is usually written down first, for a reason, and the data is asked only whether it disagrees.

22.4 Carry, Roll-Down, and the Rest

Start with the simplest possible question. I buy a bond, fund it, and hold it for three months. Where does the money come from?

Write $y(T)$ for the yield of a T-maturity bond, D for its duration and h for the horizon. Over the horizon two things happen: the bond gets older, and the curve moves. Expanding the price change,

$$\underbrace{\frac{\Delta P}{P}}_{\text{total return}} \approx \underbrace{y(T)\,h}_{\text{coupon}} \underbrace{-\,r\,h}_{\text{funding}} \underbrace{-\,D\big(y(T-h)-y(T)\big)}_{\text{roll-down}} \underbrace{-\,D\,\Delta y}_{\text{yield change}} + \tfrac{1}{2}C(\Delta y)^2. \tag{22.2}$$

[4] measured.

[5] measured.

The first two terms are the *carry*: what the bond pays less what the funding costs. The third is the *roll-down*: even if the curve does not move at all, a five year bond becomes a four-and-three-quarter year bond, and on an upward sloping curve that means its yield falls and its price rises. The fourth is the only term involving an actual change in the market, and the fifth is the convexity of chapter 7.

Calculation 22.3 (How the known part compares with the unknown)**.** Equation (22.2) separates what is known from what is not and says nothing about their sizes, which is the comparison that decides whether a carry trade is a harvest or a bet. Both are available: the known part from today's curve, and the risk from the realised volatility of the same maturity's yield over the same horizon.

Taking the Treasury curve and a quarter's holding period, with the one-month bill as funding and yield volatilities measured from the year's daily changes:[6]

Maturity	**yield**	**carry + roll**	**risk over the quarter**	**ratio**
1y	4.01%	7bp	20bp	0.37
2y	4.19%	17bp	62bp	0.28
5y	4.35%	19bp	162bp	0.12
10y	4.65%	32bp	263bp	0.12
30y	5.19%	35bp	444bp	0.08

Remark (The long end pays more and is worth less)**.** Every ratio is well below one. The part of the return that is known today is a fraction of a single standard deviation of the part that is not, at every maturity, so a carry position held for a quarter is not an income stream with noise attached — it is a directional bet with a small tilt in its favour. That is consistent with what the two conditions above claim and is a good deal more concrete: the tilt at ten years is about an eighth of the quarter's uncertainty.

And the ratio *falls* with maturity while the carry rises. Thirty year bonds pay five times the carry and roll of one year bonds and are the worst risk-adjusted place to collect it, because duration multiplies the uncertain term and adds only roll to the known one. A desk reaching down the curve for carry is buying a bigger number and a worse trade, and the number is the one that appears in a pitch.

The important structural feature of (22.2) is that the first three terms are *known today.* They are properties of today's curve and the passage of time, not forecasts. Only the fourth is uncertain.

That is what makes carry-and-roll trades attractive and what makes them dangerous. A position with positive carry and roll makes money on every day the market does not move. It loses money when the market moves against it, and the losses are proportional to duration, which is to say much larger than the daily income. A carry trade is therefore short a large, rare loss and long a small, steady gain — which is the payoff diagram of a sold option, assembled without buying or selling one.

Remark. Note how much of this survives without any model at all. The decomposition is arithmetic and the figure is the curve of chapter 7 with a subtraction applied. No volatility, no measure, no calibration. That is characteristic of relative value work: most of the analysis is careful book-keeping, and the modelling enters only when an option is involved.

[6] `risk::carry_and_roll`, against the panel in `public/marketdata`.

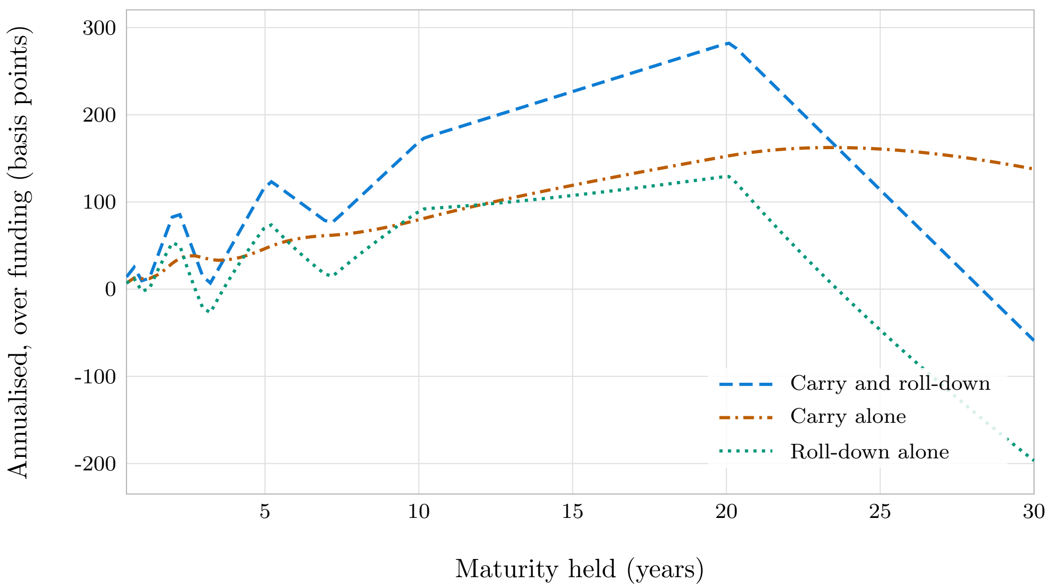

Figure 22.1: Carry and roll-down along the US Treasury curve, annualised, over the cost of funding for three months. Carry rises steadily with maturity: a longer bond simply yields more over the funding rate. Roll-down does not — it depends on the *slope* of the curve where the bond sits, so it is largest where the curve is steepest and turns negative where the curve inverts. The total is the sum, and it is not monotone in maturity: the point of the curve that pays best to hold is not the longest one.

Drawn by `bootstrap_par` and `Curve::df`. US Department of the Treasury, daily par yield curve rates, as of 2026-08-07 (par yield, semiannual coupon, actual/actual). Retrieved from https://home.treasury.gov/interest-rates-data-csv-archive.

22.5 Curve Trades

A directional position in one bond is a bet on the level of rates, which is a macroeconomic view rather than a relative value one. The relative value version isolates the shape.

Chapter 12 noted that a principal component analysis of curve changes finds three factors: a level factor moving all rates together, a slope factor steepening or flattening, and a curvature factor moving the middle against the ends. Between them they explain almost all of the variance.

This immediately suggests what to trade. Pick three points on the curve, and choose notionals w_1, w_2, w_3 so that the position has no exposure to the first two factors:

$$\sum_i w_i D_i \, \ell_i = 0, \qquad \sum_i w_i D_i \, s_i = 0,$$

where ℓ and s are the loadings of the level and slope factors on each point and D_i the durations. Two equations in three unknowns leave a one-dimensional family, and fixing the overall size picks one. The result is a *butterfly*: long the middle against the wings, or the reverse.

What is left is exposure to curvature and to nothing else — so if the middle of the curve is dear relative to the two ends, this position monetises exactly that, whatever the level and slope do.

That is the claim, and on a real curve it is only approximately true.

Calculation 22.4 (How neutral the neutral position is)**.** Take the equal-weighted $1 : -2 : 1$ fly on the two, five and ten year points — the version that is used because it needs no estimate at all, being exact for any move that is linear in maturity — and measure how much of its daily variance the first two principal components still explain, on the committed Treasury history.[7]

	explained by level and slope
Outright five year	98.1%
$1 : -2 : 1$ butterfly	26.4%

So the weighting does most of what it claims and not all of it. An outright position is the level factor and essentially nothing else; the fly removes three quarters of the factor exposure, and a quarter of what it does is still level and slope. Its own volatility is around two basis points a day, so the residual is not a rounding error — it is a directional position of real size sitting inside a trade that is described as non-directional.

The gap is the difference between the weighting that is exactly right for a straight-line move and the weighting that is right for the moves this curve actually makes. Fitting the weights to the estimated loadings closes it, at the cost of the next remark.

Remark (The weights are a modelling choice)**.** The loadings come from a covariance matrix estimated over some window, and the estimate depends on the window. A butterfly hedged with loadings from a calm period is not hedged in a volatile one, because the factors themselves rotate.

This is chapter 7's lesson about interpolation, and chapter 24's about model risk, arriving in a third place: the hedge is a property of an estimate, and the estimate is a choice. A desk running these trades is exposed to the estimation window in a way that no risk report shows, because the risk report uses the same window.

[7] `risk::factor_neutrality`, on the same daily curves as the factor decomposition above.

Calculation 22.5 (Not choosing a window)**.** There is no window that is right, and that can be shown rather than argued. Take a hedge ratio that genuinely wanders and estimate it two ways, over a month and over a year. When the ratio moves slowly the long window wins comfortably; when it moves five times faster the short window wins instead.[8] The better choice depends on a rate of rotation that is not observed, so choosing a window is guessing at it while pretending to be doing something else.

The alternative is to stop choosing. Write the ratio as a state that drifts and the regression as a noisy view of it,

$$\beta_t = \beta_{t-1} + \eta_t, \qquad y_t = \beta_t\, x_t + \varepsilon_t,$$

and this is chapter 21's filter with a different state in it — no new machinery, and the same recursion. Its gain settles at a level fixed by the ratio of the two variances, so the estimate is exponentially weighted with a memory implied by how fast the ratio is thought to move rather than declared in advance. Told the truth about that rate, it beats both windows at every rate.[9]

Remark (What the filter has actually changed)**.** Not as much as the last paragraph suggests.

The filter needs the rate of rotation, which is the window question in different clothes. Give it a rate within about a factor of two of the truth and its advantage survives; tell it the ratio moves five times more slowly than it does and it is *worse* than having simply taken the long window.[10]

So the gain is not that a parameter has been removed. It is that the parameter has been replaced by a better posed one. "How many days of history should the weights use" is a question about the estimator with nothing outside the data to answer it; "how fast do the factors rotate" is a question about the curve, and chapter 12's factor analysis is the sort of thing that could answer it. That is a real improvement and a modest one, and it is the honest description of what filtering buys here.

22.6 The Cash-Futures Basis

Now the trade that made the news, and the one where the mathematics is richest.

A Treasury futures contract does not reference a single bond. The short may deliver any bond from a defined basket, and to make bonds of different coupons and maturities comparable the exchange applies a *conversion factor* CF_i to each. Delivering bond i against a futures priced at F earns the short the invoice amount

$$\mathrm{Invoice}_i = \mathrm{CF}_i \times F + \text{accrued interest}_i.$$

Definition 22.6 (The basis)**.** For a deliverable bond with clean price P_i,

$$\mathrm{Basis}_i = P_i - \mathrm{CF}_i \times F.$$

The short delivers whichever bond costs least to buy and deliver, so the futures price is set by the *cheapest to deliver* — the bond minimising the basis. That is an option held by the short, and like every option in these notes it has value, is worth more when volatility is higher, and changes

[8] `measured`.

[9] `measured`.

[10] `measured`. Being told too fast is the more forgiving of the two errors, which is the usual asymmetry: a filter that distrusts its own history recovers, one that will not update does not.

hands as the market moves: a large enough rally changes which bond is cheapest, so the futures contract's effective underlying switches.

The standard way to see whether the basis is rich or cheap is to ask what financing rate makes the trade break even.

Definition 22.7 (Implied repo rate)**.** The implied repo rate is the return earned by buying the bond today, funding it to delivery, and delivering it into the futures:

$$\mathrm{IRR} = \frac{\text{Invoice at delivery} - \text{Purchase price today}}{\text{Purchase price today}} \times \frac{1}{\tau},$$

with τ the time to delivery.

Compare it with the actual repo rate at which the bond can be financed. If the implied repo rate exceeds the actual, the package — buy the bond, sell the future, fund in repo — earns more than it costs, and that difference is the basis trade.

Calculation 22.8 (What the basis trade actually is)**.** Set it out as a portfolio, which is the only way to see the risk.

- Buy N face of the cheapest to deliver bond, at price P.
- Finance it in the repo market: post the bond as collateral, receive $NP(1-h)$ in cash, where h is the haircut. Put up NPh of your own money.
- Sell $N \times \mathrm{CF}$ of futures — the amount that delivers exactly the bonds held, since the invoice pays CF per unit of futures price. Selling futures requires initial margin rather than the notional, so this leg costs little cash.

The futures position is pinned by the bonds held and by nothing else: it is set so that the delivery obligation is exactly what is owned.

With the legs matched, the position is close to hedged: a rise in yields loses money on the bond and makes it on the future. What remains is the basis itself, which converges to zero at delivery — so the profit is the difference between the implied and actual repo rates, earned with near certainty if the position is held to delivery.

So the leverage does not come from the futures at all. It comes entirely from h, the haircut on the bond leg, which is what lets a large holding of bonds be carried against a small amount of capital; the futures margin is a further call on capital rather than a multiplier of the position. With a haircut of 1%, \$100 of bonds requires \$1 of equity to hold, so a basis of two basis points annualised becomes a return on that equity of two hundred. *The trade is not attractive because the edge is large. It is attractive because the edge is small and the leverage is enormous.*

Why does the gap exist at all? Because the two sides of it are wanted by different people. Asset managers who want duration without balance sheet buy futures, pushing them rich. Someone must take the other side, hold the cash bond, and finance it, and that is a use of balance sheet that banks have been progressively less willing to supply since capital rules began charging for it. Hedge funds stepped into that gap. The basis is, in effect, the price of balance sheet.

22.7 Why It Breaks

The trade is hedged against the thing it looks exposed to, and exposed to two things it does not look exposed to. Both are about funding rather than about rates.

The margin on the futures leg is settled daily; the gain on the cash leg is not. Chapter 8 derived this asymmetry to compute the future-forward basis, where it was worth a basis point or two. Here it is a cash flow: if yields fall, the short futures position loses and must post variation margin *today*, while the matching gain on the bond is unrealised. A hedged position with no economic exposure can therefore demand cash without limit in the interim.

The haircut is not constant. It is set by the lender and rises when markets are volatile. A haircut moving from 1% to 3% triples the capital the position requires, and a fund that does not have it must sell. Selling the cash leg widens the basis, which marks every other holder of the trade to a loss and raises their margin calls too.

That is the whole mechanism. At 1% haircut the leverage is a hundred to one. A move in the basis of ten basis points — unremarkable in a stressed week — is ten percent of capital. A move of a hundred is the fund.

Remark (This is chapter 4's admissibility condition, in the wild)**.** There is a genuine and not merely rhetorical connection to the theory.

Chapter 4 met a strategy that was an arbitrage under the definition then in use: the doubling strategy, which reaches a certain profit almost surely by being willing to sustain unbounded losses along the way. The repair was to require *admissibility* — that the wealth process stay above a fixed level — and we observed that without it no market is arbitrage-free, because the mathematics permits a trade that no counterparty would fund.

The basis trade is that condition being tested by reality. Held to delivery with no financing constraint it converges, and the profit is nearly certain. Held with a hundred times leverage and daily margin, the interim losses are exactly what admissibility rules out, and the strategy fails not because the arbitrage was not there but because surviving to collect it was not financed.

The lesson chapter 4 drew from a piece of measure theory is the same one: a profit you cannot fund your way to is not a profit.

Exercise (Sizing the spiral)**.** A fund holds \$100 of the basis trade at a 2% haircut, so \$2 of capital. The haircut rises to 4%. Show that, holding the position, the fund needs \$2 more capital; and that if instead it deleverages to its existing capital it must sell half the position. Then argue that if enough funds do this at once the basis widens, and show that a widening of 20 basis points on the remaining position wipes out a further tenth of the capital. This is the loop, and its speed is set by the haircut, not by rates.

22.8 Swap Spreads, and a Negative Number That Should Not Exist

One more spread, because it makes a point about what these trades measure.

The swap spread is the swap rate of chapter 7 less the yield of the matched-maturity government bond. Naively it should be positive: a swap has bank credit behind it and a government bond does not, so the swap should pay more.

At long maturities it has been persistently *negative*, which appears to say that lending to a bank is safer than lending to the government. It does not. The swap is collateralised, and chapter 7 derived what that means: a fully collateralised trade carries essentially no credit exposure, and is discounted at the collateral rate. The bond is not collateralised — it must be bought, funded and held on a balance sheet that is charged for it.

So the swap spread is not a credit spread at all. It is the price of balance sheet again, the same quantity the basis measures, seen from another direction. Once the capital rules made holding bonds expensive, the spread that compensated for that could and did go negative, and stayed there.

Remark. The general lesson: when a spread that theory says should have one sign persistently has the other, the usual explanation is not that the market is wrong. It is that the theory omitted a constraint that binds — here, that balance sheet is finite and charged for. Chapter 7's collateral derivation and chapter 24's capital reserves are two faces of that constraint, and this spread is its price.

22.9 Selling Volatility, and What That Is Short

Everything so far has been a trade in a rate or a spread between rates. There is a second family, at least as large, in which the instrument is an option and the quantity being traded is volatility itself.

The position is a swaption or a cap sold and then delta hedged, so that the direction of rates is removed and what is left is the difference between the volatility charged and the volatility that arrives. The structural reason is on the other side and is the one chapter 23 describes: the flow is one-directional. Borrowers want caps and issuers want the option to call, so the dealer community is structurally short optionality and is paid to be. Chapter 10 locates the same premium inside a model, where it sits in the drift parameters κ and θ and not in the vol-of-vol η, which is measure-invariant.

Calculation 22.9 (What a hedged short option actually earns)**.** Sell a call at an implied volatility σ_i, hedge it with the delta computed at σ_i, and let the world realise σ_r. Over one rebalancing interval the position is delta neutral, so by chapter 5's expansion the profit and loss over that interval is the theta collected less the gamma paid,

$$-\Theta\, dt - \tfrac{1}{2}\Gamma\,(dS)^2,$$

and the theta-gamma identity of that chapter, $\Theta = -\frac{1}{2}\sigma_i^2 S^2\Gamma$ at zero rates, substitutes for the first term. Since $(dS)^2 = \sigma_r^2 S^2 dt$ along the realised path, the two collapse into

$$\text{P\&L} \;=\; \int_0^T \tfrac{1}{2}\,\Gamma_t\, S_t^2\left(\sigma_i^2 - \sigma_r^2\right) dt. \tag{22.3}$$

Simulating the hedge and accumulating (22.3) along the same paths are different computations — the first forms the payoff and uses the delta, the second never forms a payoff and uses the gamma — and they agree.[11]

[11] `measured`.

Read (22.3) carefully, because it does not say what the trade is usually described as saying. It is not a bet on realised variance. It is a bet on realised variance *weighted by gamma*, and gamma is large only near the strike and near expiry. A move of a given size contributes according to where the underlying was when it happened, so two paths delivering identical realised variance pay different amounts.

How different is worth measuring. Selling a one-year at-the-money option at 22% into a world realising 18% earns 1.59 on average, which is exactly the difference between the two Black-Scholes premiums and is the sense in which the edge is real.[12] But restricting to the paths that realised 18% to four decimal places, the outcomes still run from 0.49 to 2.94.[13] The dispersion remaining after the forecast has been made *exactly right* is wider than the entire average edge.

Structure (Short the path, not the variance)**.** A desk selling volatility describes itself as short variance and forecasts variance accordingly, but (22.3) says the exposure is to a gamma weighted average, and the weighting is a fact about the path rather than about its total. Being right about volatility and wrong about where the moves land is a losing position.

So the trade is short the *shape* of the realised path, and the chapter's usual question has the usual answer: it is short a constraint, namely that the hedge is discrete and the gamma is concentrated. Chapter 5 has already drawn the same object from the other side, where two markets with identical total variance and different jump content produced identical premiums and entirely different distributions of outcome. This is that figure's trading counterpart.

It also explains the instrument. A variance swap pays realised variance with no gamma weighting at all, which is exactly the exposure the hedged option fails to give, and that — not any view on volatility — is what it is for. Chapter 15 has in fact already built the object, though not under that name: the second moment of the swap rate comes out as an integral of swaption prices with *equal* weight at every strike, which is a variance swap on the rate in all but the contract. The uniform weighting is not an accident — expanding a payoff in options gives each strike the weight $g''(K)$, and the second moment has $g'' = 2$ everywhere.

The familiar $1/K^2$ weighting is the same construction asking for a different variance. Buying the variance of the *logarithm* means replicating $-2\ln(F_T/F_0)$, whose second derivative is $2/K^2$, and that is the equity convention because equity variance is quoted in proportional terms. A rates desk buying basis points squared weights uniformly. The two are not competing formulas for one instrument; they are the same formula for two.

Three further trades in this family:

The volatility term structure. Selling short-dated and buying long-dated volatility is a trade on how implied volatility is interpolated between expiries, and what it is short is the event calendar. Chapter 9 shows why: a smooth interpolation through a payroll or a meeting date reads the lump as a term structure and marks every expiry between the quotes wrong. The trade collects the marking error and is short the possibility that the event is bigger than the lump priced for it.

Skew. Selling a receiver and buying a payer against it is a position in the smile's slope, and chapter 10 says the slope is a statement about the correlation between the rate and its volatility. What that is short is not a rate move but a change in that correlation — and the mortgage hedging flow of the next section is precisely a mechanism that changes it, which is why the two trades are less independent than a risk report shows.

[12] `measured`.

[13] `measured`.

Conditional and curve volatility. A swaption on a spread is worth an amount that depends on the correlation between the two rates, and chapter 13 measures how much: marking correlation moves a price by percentage points where the choice between two well-specified models moves it by hundredths. A trade in spread volatility against outright volatility is therefore short a correlation marking, which is an estimate from chapter 21 rather than a quote, and the least observable input in the book.

22.10 Six More, and What Each One Is Short

The two trades worked through above are not a representative sample, and a reader who saw only them would take away that relative value is about bonds against futures. It is not. What follows is a catalogue rather than a derivation — each entry is a trade that has been put on at scale, with the structural reason that made it collectible and the exposure that made it dangerous.

On-the-run against off-the-run. The most recently auctioned Treasury of a given maturity trades rich to an almost identical bond issued a few months earlier. The reason is not credit and not cash flows, which are nearly the same; it is that the new bond is the benchmark, is what hedges are executed in, and finances more cheaply in repo because everyone wants to borrow it. The spread is therefore the price of liquidity and of financing, and it can be collected by buying the old bond and selling the new one. What the position is short is liquidity itself — and liquidity premia widen precisely when a position needs to be unwound. The trade was a signature of Long Term Capital Management, and in 1998 the spread went from rich to richer while the leverage that made it worth doing forced the unwind. Chapter 4's admissibility condition is the abstract version: a trade financed by an unbounded drawdown is not a trade, and the drawdown here was a mark-to-market on a position whose thesis was correct.

The squeeze. A short position in a specific deliverable bond is not short a price. It is short the ability to borrow that bond, and if someone corners the supply, the cost of maintaining the short is set by them. Salomon Brothers demonstrated this at the two year Treasury auction of May 1991 by acquiring a dominant share of the issue, after which the repo rate on the note collapsed and the shorts paid. The general point is structural and worth carrying: for a deliverable instrument the funding leg has a different risk profile from the price leg.

Inflation-linked against nominal. A Treasury inflation-protected bond plus an inflation swap that converts its cash flows to fixed replicates a nominal Treasury exactly — a static replication of the kind chapter 15 builds, with no model in it at all. The replication has been persistently cheaper than the nominal bond it replicates, at times by a wide margin, and the mispricing survived for years rather than minutes. It is as close to a textbook arbitrage as the market offers, and it persisted for the reason everything else in this section persists: capturing it requires holding a levered, balance-sheet-intensive position to maturity, and the institutions with the balance sheet had better uses for it.

Covered interest parity. Borrowing dollars directly and borrowing yen and swapping them into dollars should cost the same, since the forward rate is a ratio of numeraires and chapter 17 derives it with no freedom in it. Since 2008 they have not cost the same, the gap widens reliably at quarter ends when balance sheets are reported, and it is a cross-currency basis rather than a mistake. The trade is to supply the scarce thing — dollars, on a balance sheet — and what it is short is the regulatory constraint that made them scarce, which is to say a rule that can change.

Mortgage convexity hedging. A homeowner holds an option to prepay, so a mortgage portfolio is short an option and its duration shortens when rates fall. Whoever holds that exposure must hedge it, and the hedge is mechanical: rates fall, duration shortens, receive fixed to make it up; rates rise, pay. This is a flow proportional to the change in the aggregate duration of a very large portfolio, it is forced rather than discretionary, and it therefore amplifies rate moves and supports swaption implied volatility. The trades built on it — being long volatility ahead of the move, or trading the amplification — are short the stability of prepayment behaviour, which is a model output and one of the least reliable ones in the industry.

The auction cycle. Dealers must absorb each new issue, and their capacity to warehouse it is finite. Yields have tended to rise into an auction and fall after it, a pattern with a named mechanism rather than a curve-fit, and it is the closest thing in this chapter to a pure statistical arbitrage with a structural reason attached. What it is short is dealer capacity: the pattern is strongest when balance sheet is scarce, which is also when the position cannot be financed.

Structure (The same short, six times)**.** Set the six side by side and one thing is common to all of them. None is short a price. Each is short a *constraint* — financing, borrowing, balance sheet, regulation, a behavioural model, warehousing capacity — and each pays a premium that exists because the constraint binds on somebody.

That is not a coincidence and it is the reason these trades exist at all. A genuine mispricing of a price, in a liquid market with many participants, is arbitraged away in the time it takes to notice. What is not arbitraged away is a premium paid for absorbing a constraint, because absorbing it requires the very thing that is scarce. So the surviving trades are exactly the ones where the return is compensation for a service rather than a reward for being right.

Why is a mispriced *price* removed so fast?

Not because it is hard to see. A cash-futures basis, a triangular inconsistency in currencies, a bond trading through its own strip — each is arithmetic on quoted numbers, computed identically by everyone with the same feed. Detection is not the scarce input and confers no advantage, which is why no desk earns anything from noticing.

It is removed fast because it is a *race for a finite quantity.* A mispricing is not a pool of free money but a bounded amount available at a stale price, and it goes to whoever reaches it first. That makes the return to speed winner-take-most rather than proportional, and a payoff of that shape is competed to whatever the physical limit happens to be — which is why the participants who do this have spent their capital on the distance between two machines rather than on models.

And that is exactly what separates the two classes. A mispricing removable by a fast trade needs no balance sheet: the position is opened and closed, nothing is carried, and the only scarce input is time, which is bought once and then owned. A premium for absorbing a constraint cannot be competed away by speed at all, because the scarce input is the ability to *hold* — collateral, capital, warehousing, the willingness to sit through §22.1's excursion — and no amount of speed supplies it. Speed removes the arbitrages that can be closed in a moment and is powerless against the ones that must be carried for months, which is why the surviving trades all look alike.

The fast trader's profit is the slow quoter's adverse selection as seen in chapter 23: a quote that can be picked off before it is pulled is exactly the toxic flow that chapter measures and prices into the spread.

It follows that the risk of the whole class is one risk. All six lose money in the same circumstance — when the constraint tightens rather than relaxes, which is when financing is withdrawn, balance sheets contract and everyone holding the premium needs to stop. That is why relative value books that appear diversified across trades are not, and why chapter 24's stress work has to be run on the constraint rather than on the instrument.

22.11 Where an Edge Lives: Information as a Filtration

Everything so far in this book has fixed a filtration $\mathcal{F}_t$ at the outset and asked what prices must be, given it. That construction has been so uniform that it is easy to miss what it assumes: that everyone sees the same thing. An edge is what happens when they do not, and the machinery to say so precisely is already in place.

Let $\mathcal{G}_t$ be the market's filtration, the one under which the discounted price is a martingale by chapter 4, and let $\mathcal{F}_t \supseteq \mathcal{G}_t$ be a trader's, containing something extra. A martingale in the smaller filtration need not be a martingale in the larger one. Under a technical condition it remains a semimartingale, and acquires a drift:

$$\underbrace{dS_t = \sigma_t \, dW_t}_{\text{in } \mathcal{G}} \qquad \longrightarrow \qquad \underbrace{dS_t = \sigma_t \alpha_t \, dt + \sigma_t \, d\tilde{W}_t}_{\text{in } \mathcal{F}}. \tag{22.4}$$

That drift α is called the *information drift.* The alpha a trader is looking for is not a metaphor borrowed from regression. It is a drift term, it exists only relative to a filtration, and its magnitude is the amount by which one information set disagrees with another about where the price is going. The literature that works this out is the literature on enlargement of filtrations, developed to study insider trading; the mathematics is indifferent to how the extra information was obtained.

Example 22.1 (A signal, and what it is worth)**.** Take the tractable case, where the extra information is a single number seen at the outset. Let W drive the price over $[0, 1]$ and let the trader observe

$$L = \rho W_1 + \sqrt{1-\rho^2}\, Z,$$

with Z independent noise — a view on the terminal value, correlated ρ with it.

Fix t and work conditionally on $\mathcal{F}_t$, which is to say on W_t. Split the signal at t,

$$L \;=\; \underbrace{\rho W_t}_{\text{known}} \;+\; \underbrace{\rho\,(W_1 - W_t) + \sqrt{1-\rho^2}\, Z}_{\text{not}},$$

so the unknown part has variance $\rho^2(1-t) + (1-\rho^2) = 1 - \rho^2 t$. Write $V = L - \rho W_t$ for it. This is the piece of the signal the trader has not yet seen confirmed, and it shrinks as t runs on.

Now take the next increment ΔW over $[t, t+h]$. Given $\mathcal{F}_t$, the pair $(\Delta W, V)$ is jointly Gaussian with

$$\mathrm{Var}(\Delta W) = h, \qquad \mathrm{Var}(V) = 1 - \rho^2 t, \qquad \mathrm{Cov}(\Delta W, V) = \rho h,$$

the covariance because L loads on every increment of W with weight ρ. Gaussian conditioning is then one line,

$$\mathbb{E}\big[\Delta W \,\big|\, \mathcal{F}_t, L\big] \;=\; \frac{\mathrm{Cov}(\Delta W, V)}{\mathrm{Var}(V)}\, V \;=\; \frac{\rho\, h}{1-\rho^2 t}\,\big(L - \rho W_t\big),$$

and dividing by h gives the drift the enlarged filtration sees,

$$\alpha_t \;=\; \frac{\rho\,(L - \rho W_t)}{1-\rho^2 t}. \tag{22.5}$$

So the information drift is a *regression coefficient* — the slope from projecting the next increment on the part of the signal not yet realised. That is the same object chapter 4 found when it wrote the hedge ratio as $d\langle V, S\rangle / d\langle S\rangle$ and read it as a regression slope. Alpha and a hedge ratio are the same kind of thing seen from two sides.

The second moment follows without further work. Unconditionally $V \sim N(0, 1-\rho^2 t)$, so squaring (22.5) and taking expectations,

$$\mathbb{E}[\alpha_t^2] \;=\; \frac{\rho^2\,(1-\rho^2 t)}{(1-\rho^2 t)^2} \;=\; \frac{\rho^2}{1-\rho^2 t},$$

which rises as $t \to 1$: the signal is worth most at the end, when little of it is left unconfirmed and the remaining uncertainty has collapsed onto it. What that drift is worth needs one standard fact. Put a fraction π of wealth in an asset with $dS/S = \mu\,dt + \sigma\,dW$ and the rest in cash at zero, so that $dX/X = \pi\mu\,dt + \pi\sigma\,dW$. By Itô,

$$d\ln X_t = \Big(\pi\mu - \tfrac{1}{2}\pi^2\sigma^2\Big)dt + \pi\sigma\,dW_t,$$

so the long-run growth rate is the bracket, and maximising it over π gives $\pi^* = \mu/\sigma^2$ and

$$g^* \;=\; \tfrac{1}{2}\Big(\frac{\mu}{\sigma}\Big)^2. \tag{22.6}$$

The achievable growth rate is half the squared Sharpe ratio, and nothing about the horizon or the wealth enters.

Now read (22.4) through (22.6). In the enlarged filtration the price acquires an extra drift of $\sigma\alpha_t$, so its Sharpe ratio rises by exactly α_t. If the uninformed Sharpe was θ, the informed investor's growth is $\frac{1}{2}(\theta+\alpha_t)^2$ against $\frac{1}{2}\theta^2$, and the difference is $\theta\alpha_t + \frac{1}{2}\alpha_t^2$. The signal is unbiased — $\mathbb{E}[\alpha_t] = 0$, since L and W_t both have mean zero — so the cross term vanishes in expectation and the extra growth is $\frac{1}{2}\mathbb{E}[\alpha_t^2]$, whatever the uninformed investor was already earning.

The value of the signal is second order in it: it comes from the *variability* of the perceived drift and not from its level, which is why what emerges below is an information quantity rather than a return. Integrating over the horizon,

$$\frac{1}{2}\int_0^1 \mathbb{E}[\alpha_t^2]\,dt = \frac{1}{2}\int_0^1 \frac{\rho^2\,dt}{1-\rho^2 t} = -\tfrac{1}{2}\ln\big(1-\rho^2\big), \tag{22.7}$$

which is exactly the mutual information of the jointly Gaussian pair (W_1, L).

Kelly proved the general statement in 1956, for a gambler receiving tips over a noisy wire: the growth rate a log-optimal bettor can achieve from a side channel equals that channel's mutual information, whatever the channel is. The derivation above is the continuous-time instance — half a squared Sharpe ratio, integrated — and it arrives at an information quantity because that is what the theorem says it must. Which also fixes the units in which a signal should be valued. Not in basis points of edge, which depend on how the position is sized, but in nats, which do not.[14]

[14] `estimation::information_value` simulates the drift and confirms both (22.7) and the pointwise second moment, which is where an algebra error would hide.

Example 22.2 (What the datasets actually are, in rates)**.** In fixed income the enlargement usually has a specific and mundane shape: the market is waiting for a scheduled government statistic, and something observable earlier is correlated with it.

Payrolls and unemployment are released monthly and are among the largest scheduled moves in the front end of the curve. Job postings, payroll-processor records and staffing volumes are observable continuously and are correlated with the release. Consumer spending arrives in retail sales and in the consumption component of output; card transaction panels and point-of-sale feeds see much of it first. Inflation is released monthly and scraped prices, freight rates and commodity fixings move before it. Wages appear in the employment cost series and in earnings, and are visible earlier in posted salaries. Activity shows up in satellite imagery of shipping, in electricity load, in traffic. Geopolitical and policy events are read from filings, transcripts and news well before they are quantified anywhere official.

Each of these is an enlargement $\mathcal{F}_t \supseteq \mathcal{G}_t$ of exactly the kind (22.4) describes, and (22.7) says what each is worth. That reframing does real work, because it replaces the question people usually ask — is this dataset novel, is it large, is it hard to obtain — with the only question that determines the value: how much does it move the conditional distribution of the release, *given* everything already public. A card panel covering a small and unrepresentative slice of spending may be genuinely novel and worth nothing. A widely available series may still carry information if nobody has bothered to condition on it correctly.

Two features specific to this setting are worth naming. The information has a known expiry: the release date is when $\mathcal{G}$ catches up completely and the drift goes to zero by construction, so the horizon in (22.7) is not a modelling choice but a published calendar. And the release itself is a measurement of the same underlying quantity with its own error, so a signal can be right about the economy and wrong about the print — the tradeable object is the statistic, not the world, and the conditional mutual information that matters is with respect to the number that gets published.

Structure (The value of a dataset is a mutual information)**.** Equation (22.7) is a special case of a theorem: the additional expected logarithmic utility available to a trader whose filtration is enlarged by a random variable equals the mutual information between that variable and the market. The question "what is this dataset worth" therefore has an answer, in nats, and the answer does not depend on what the dataset is made of.

Three consequences.

Novelty is not the criterion; conditional information is. What (22.7) measures is the information in L *given* what is already known. A dataset that is a function of the observable price history has zero conditional mutual information and is worth exactly nothing, however unusual its provenance — and this is a testable condition rather than a matter of judgement. Most datasets sold as alternative fail it.

The value is quadratic in the correlation at the low end. Expanding (22.7), a signal with $\rho = 0.1$ is worth about a quarter of one with $\rho = 0.2$, not a half. Weak signals are worth very much less than their correlation suggests, and the last increment of correlation is worth the most, which is the mathematical form of the observation that being slightly better informed than the market is nearly useless.

Alpha decays by ceasing to exist, not by being crowded. The drift in (22.4) is defined relative to $\mathcal{G}$. As others acquire the same data, $\mathcal{G}$ grows to contain it, and α does not shrink — it becomes zero, because the information is no longer information. That is a sharper statement than the usual one about crowding, and it predicts the right shape: the decay is driven by who else has the data rather than by how much capital is deployed against it.

Two honest limits on all of this. The mutual information is an upper bound, attained by an investor with no transaction costs, no position limits and no financing constraint, and the two conditions above's structural conditions are exactly the reasons that investor does not exist; the earlier sections of this chapter are a catalogue of what the bound loses to reality. And the binding difficulty in practice is not acquiring information but estimating α from it, which is chapter 20's identification problem in yet another guise. Equation (22.7) tells you the size of the prize. It does not tell you that you can find it, and chapter 19 argues that this — rather than compute — is what the constraint has always been.

References

- Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, 35(4), 917–926.

- Burghardt, G., & Belton, T. (2005). *The Treasury Bond Basis*. McGraw-Hill.

- Barth, D., & Kahn, J. (2021). Hedge funds and the Treasury cash-futures disconnect. *OFR Working Paper* 21-01.

- Schrimpf, A., Shin, H. S., & Sushko, V. (2020). Leverage and margin spirals in fixed income markets during the Covid-19 crisis. *BIS Bulletin* 2.

- Litterman, R., & Scheinkman, J. (1991). Common factors affecting bond returns. *Journal of Fixed Income*, 1(1), 54–61.

- Klingler, S., & Sundaresan, S. (2019). An explanation of negative swap spreads. *Journal of Finance*, 74(2), 675–710.

- Krishnamurthy, A. (2002). The bond/old-bond spread. *Journal of Financial Economics*, 66(2–3), 463–506.

- Jegadeesh, N. (1993). Treasury auction bids and the Salomon squeeze. *Journal of Finance*, 48(4), 1403–1419.

- Fleckenstein, M., Longstaff, F. A., & Lustig, H. (2014). The TIPS-Treasury bond puzzle. *Journal of Finance*, 69(5), 2151–2197.

- Du, W., Tepper, A., & Verdelhan, A. (2018). Deviations from covered interest rate parity. *Journal of Finance*, 73(3), 915–957.

- Perli, R., & Sack, B. (2003). Does mortgage hedging amplify movements in long-term interest rates? *Journal of Fixed Income*, 13(3), 7–17.

- Lou, D., Yan, H., & Zhang, J. (2013). Anticipated and repeated shocks in liquid markets. *Review of Financial Studies*, 26(8), 1891–1912.

- Amendinger, J., Imkeller, P., & Schweizer, M. (1998). Additional logarithmic utility of an insider. *Stochastic Processes and their Applications*, 75(2), 263–286.

- Pikovsky, I., & Karatzas, I. (1996). Anticipative portfolio optimization. *Advances in Applied Probability*, 28(4), 1095–1122.