Chapter 23

Market Making

Everything so far has priced instruments. This chapter quotes them, which is a different problem: a price is a number and a quote is a pair of numbers offered to somebody who gets to choose which side to take. We derive the one genuinely solvable quoting model, and find that the substitution which solves it is the same that solved the affine models — exponential utility makes a nonlinear control problem linear. Then we look at what that model cannot see, which is that the counterparty may know something, and at how a rates desk measures whether its flow is toxic. The chapter ends where a desk's day ends: with a position nobody chose, because hedging it completely was not worth what it cost, which is what chapter 24 has to carry.

23.1 What the Spread Is Payment For

A market maker undertakes to quote both sides of a price on request, and is paid the difference between them. The undertaking is the product: a client who wants to hedge a liability in size, now, does not want to wait for a natural counterparty to appear. Immediacy is the service and the spread is its fee.

That much is uncontroversial and explains nothing about how wide the spread should be. Two separate costs set it.

- *Inventory risk.* Having bought, the maker holds a position it did not choose, and the position has variance. This cost is symmetric in the sign of the flow, grows with volatility, and would exist even if every counterparty were a coin flip.

- *Adverse selection.* The counterparty chose to trade, and chose which side. If they know something, the maker is on the wrong side of it. This cost is asymmetric by construction — it exists precisely because the client selects — and would exist even if the maker could hedge instantly and hold no inventory at all.

The two point in the same direction on the spread and for opposite reasons, which is why a desk that models one and neglects the other can be badly wrong while appearing to have a theory.

The next section builds the model of the first. The one after that is about the second, which no tractable model of the first contains.

23.2 A Solvable Quoting Problem

The canonical model is Avellaneda and Stoikov's. It is the only quoting model in this book that is solvable in the sense of chapter 14, and *why* it is solvable turns out to be a result we already have.

Definition 23.1 (The quoting problem)**.** A mid price diffuses with no drift, $dS_t = \sigma\, dW_t$. The maker continuously chooses distances $\delta_t^a, \delta_t^b \geq 0$ and stands ready to sell at $S_t + \delta_t^a$ and buy at $S_t - \delta_t^b$. Fills of one unit arrive as Poisson processes whose intensity falls with the distance,

$$\lambda(\delta) = Ae^{-k\delta}, \tag{23.1}$$

independently on the two sides. Writing X_t for cash and q_t for inventory, the maker maximises exponential utility of terminal wealth,

$$u(t,x,q,s) = \sup_{\delta^a,\delta^b} \mathbb{E}\Big[-\exp\big(-\gamma(X_T + q_T S_T)\big)\Big]. \tag{23.2}$$

Equation (23.1) says the further from the mid one quotes, the less one trades, with a sensitivity k. It is the only place the market's willingness to deal enters, and it is exogenous: the arrival rate depends on where the maker quotes and on nothing else. In particular it does not depend on where the price is about to go, and §23.3 is about the consequences.

The exponential utility in (23.2) looks like a taste and is a tractability assumption. It will make the problem solvable, and nothing else would.

And the mid price is a martingale, so the maker has no view. That is the right convention: a maker with a view should express it as a position, not by leaning the quote, and the two decisions are cleanly separable in this model precisely because the drift is zero.

The equation

Calculation 23.2 (The Hamilton-Jacobi-Bellman equation)**.** Over $[t, t+dt]$ three things can happen: nothing, a sale, or a purchase. Collecting the diffusion of s and the two jump terms, the value function satisfies

$$\begin{aligned} 0 = \partial_t u + \tfrac{1}{2}\sigma^2 \partial_{ss} u &+ \sup_{\delta^a} \lambda(\delta^a)\Big[u\big(t, x+s+\delta^a, q-1, s\big) - u\Big] \\ &+ \sup_{\delta^b} \lambda(\delta^b)\Big[u\big(t, x-s+\delta^b, q+1, s\big) - u\Big]. \end{aligned} \tag{23.3}$$

A sale moves cash up by the price received, $s + \delta^a$, and inventory down by one; a purchase does the reverse. The suprema sit *inside* the equation, which is what makes it a control problem rather than a valuation.

Equation (23.3) is unpromising. It is nonlinear, it contains two optimisations, and the state has four dimensions. What rescues it is that the objective was chosen to make a particular family invariant.

Theorem 23.3 (The substitution that solves it)**.** *Write*

$$u(t,x,q,s) = -\exp\big(-\gamma(x+qs)\big)\cdot v(t,q)^{-\gamma/k}. \tag{23.4}$$

Then (23.3) *holds if and only if* v *satisfies the* linear *system*

$$-\frac{\partial v}{\partial t}(t,q) = -\alpha\, q^2\, v(t,q) \;+\; \eta\Big(v(t,q-1)+v(t,q+1)\Big), \tag{23.5}$$

with $v(T,q)=1$ *and constants*

$$\alpha = \frac{k\gamma\sigma^2}{2}, \qquad \eta = A\left(1+\frac{\gamma}{k}\right)^{-(1+k/\gamma)}. \tag{23.6}$$

The optimal distances are then

$$\delta^a(q) = \frac{1}{k}\ln\frac{v(t,q)}{v(t,q-1)} + \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{k}\right), \qquad \delta^b(q) = \frac{1}{k}\ln\frac{v(t,q)}{v(t,q+1)} + \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{k}\right). \tag{23.7}$$

Proof. Substitute (23.4) and divide through by the positive quantity $\exp(-\gamma(x+qs))$, which appears in every term. Writing $\theta(t,q) = \frac{1}{k}\ln v(t,q)$, so that $u = -\exp(-\gamma(x+qs+\theta))$, the derivatives are

$$\partial_t u = \gamma\,\partial_t\theta\cdot|u|, \qquad \partial_s u = \gamma q\,|u|, \qquad \partial_{ss}u = -\gamma^2 q^2\,|u|.$$

A sale takes u to $-\exp(-\gamma(x+qs+\delta^a+\theta(t,q-1)))$, since the cash gained is $s+\delta^a$ and the inventory term loses one s. So the jump bracket is

$$u(\text{after}) - u = |u|\Big[1-e^{-\gamma(\delta^a+\Delta^a)}\Big], \qquad \Delta^a = \theta(t,q-1)-\theta(t,q).$$

Dividing (23.3) by $|u|$ leaves an equation in θ alone — the cash, the price and the utility scale have all cancelled, which is the point of the exponential form:

$$\gamma\,\partial_t\theta - \tfrac{1}{2}\gamma^2\sigma^2 q^2 + \sup_{\delta^a} G(\delta^a,\Delta^a) + \sup_{\delta^b} G(\delta^b,\Delta^b) = 0, \tag{23.8}$$

where $G(\delta,\Delta) = Ae^{-k\delta}\big[1-e^{-\gamma(\delta+\Delta)}\big]$.

Now do the optimisation, which is one line. Setting $\partial G/\partial\delta = 0$ and writing $z = e^{-\gamma(\delta+\Delta)}$,

$$-k(1-z)+\gamma z = 0 \quad\Longrightarrow\quad z = \frac{k}{k+\gamma} \quad\Longrightarrow\quad \delta^* = \frac{1}{\gamma}\ln\left(1+\frac{\gamma}{k}\right) - \Delta,$$

which is (23.7) once Δ is written in terms of v. Substituting back, G at its optimum is $Ae^{-k\delta^*}\gamma/(k+\gamma)$, and

$$e^{-k\delta^*} = \left(1+\frac{\gamma}{k}\right)^{-k/\gamma}\frac{v(t,q\mp1)}{v(t,q)},$$

so both suprema in (23.8) become multiples of $v(t,q\mp1)/v(t,q)$. Finally $\partial_t\theta = \partial_t v/(kv)$; multiplying (23.8) through by kv/γ clears every denominator and gives (23.5) with the constants (23.6). □

Structure (The same trick as the affine models)**.** A control problem and a pricing problem have just turned out to be the same kind of object.

Chapter 14 showed that a nonlinear equation becomes linear when the unknown is an exponential whose parameters the equation can absorb — there the family was $e^{\phi+\psi x}$, the parameters moved by a Riccati equation, and the nonlinearity came from the second derivative bringing a parameter down twice. Here the family is $e^{-\gamma(x+qs+\theta)}$, and γ, the risk aversion, plays the role ψ played there. Dividing (23.3) by $|u|$ is the step that removes the wealth and the price from the problem entirely, exactly as dividing by u removed the state there.

The difference is instructive. In the affine case the reduction left a *Riccati* equation, quadratic in the parameter. Here it leaves a *linear* one, and the reason is visible in the proof: the price enters the utility linearly, through qs, so the second derivative $\partial_{ss}u$ contributes a term in q^2 that is a coefficient rather than an unknown. The nonlinearity that would have made it Riccati was spent on the optimisation instead, and the optimisation closed in one line.

So exponential utility is not a convenience here in the sense of making the algebra shorter. It is the choice that makes an invariant family exist, and Definition 23.1 would be a numerical problem with any other utility. That is the same trade chapter 14 describes throughout: a modelling restriction adopted because it is what makes the mathematics close.

What it says

Solving (23.5) numerically and reading off (23.7) gives the quotes.[1] At $\sigma = 30\%$, $\gamma = 0.1$, $A = 140$, $k = 1.5$ and a one-year horizon:

Inventory q	**bid distance**	**ask distance**	**total spread**	**skew**
-10	0.574	0.725	1.300	$+0.075$
-5	0.613	0.685	1.298	$+0.036$
0	0.649	0.649	1.298	0
5	0.685	0.613	1.298	-0.036
10	0.725	0.574	1.300	-0.075

Remark (Inventory skews the quote, it does not widen it)**.** The fourth column is nearly constant and the fifth is not, and that is the model's central practical statement. A maker who is long does not quote a wider market; it quotes a *lower* one — a better ask and a worse bid — and keeps the width. The position is managed by moving both quotes down together, so that the flow it attracts is the flow that reduces the position.

This is the analogue of an indifference price. A maker holding q values the asset below the mid, by

$$\text{reservation offset} = -q\gamma\sigma^2(T-t), \tag{23.9}$$

and quotes symmetrically about *that* rather than about the mid. The offset is linear in the inventory, in the risk aversion and in the variance still to be borne, all three of which are as one would hope.

The width itself is the other term of (23.7), and it has a floor. As $T-t \to 0$ the inventory term vanishes and the spread does not: it tends to $(2/\gamma)\ln(1+\gamma/k)$, which for these parameters is 1.291

[1] `marketmaking::Quoting`, which also solves (23.3) directly without the substitution. The two agree to the discretisation — the gap halves each time the time step does and extrapolates to nothing — which is how one checks that a substitution linearises a problem rather than approximating it.

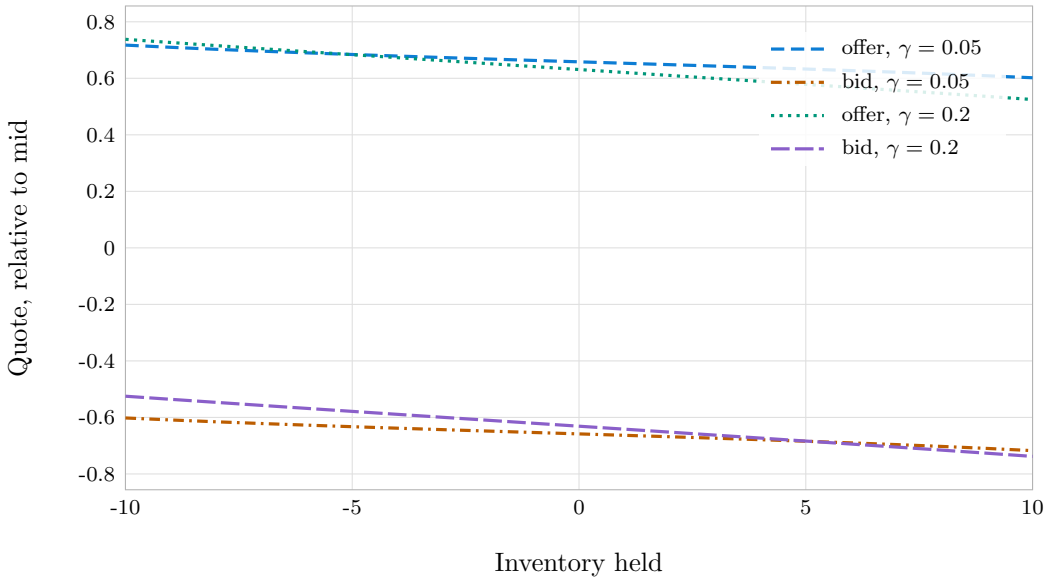

Figure 23.1: The solved quotes against the inventory held, at two levels of risk aversion. Read the pair rather than either line: as the position grows the bid and the offer slide *down together*, keeping their width, so a maker who is long shows a better offer and a worse bid rather than a wider market. Raising γ steepens the slide without changing the width. The width is set by the arrival decay k and the risk aversion; the slide is set by the position, and the two are independent, which is the model's one genuinely practical statement.

Drawn by `Quoting::linear_solution` and `Quoting::quotes`.

of the 1.298 quoted at a year. Almost all of the spread here is not compensation for risk at all. It is the margin implied by (23.1) — the maker quotes wide because it can, and k is what limits how wide.

Remark (Which parameters a desk can actually observe)**.** Of the four, σ is estimable from the market and γ is a policy choice, however uncomfortably. The other two are the difficulty. A and k describe how the arrival of client business responds to the price shown, they are properties of a franchise rather than of a market, and estimating them requires the desk's own historical quotes and fills — data no external party has, and which is heavily censored, since a quote that lost tells you the trade happened elsewhere but not at what level.

This is chapter 20's identification problem in a new place. The parameter that matters most for the answer, k, is the one estimated from the least reliable data, and (23.7) depends on it directly rather than through a second-order term.

23.3 What the Model Cannot See

Return to (23.1). The intensity depends on the distance quoted and on nothing else. In particular the arrival of a fill carries no information about the subsequent path of S: a buy and a sell are equally likely at symmetric distances, whatever is about to happen.

This is not a small idealisation. It removes adverse selection from the problem entirely, and adverse selection is the thing a real desk spends most of its attention on. The model's inventory is a risk — a symmetric bet that happens to be unwanted. Real inventory is worse than that, because of *how* it was acquired: the maker is long precisely because somebody chose to sell.

Example 23.1 (The minimal model of the other cost)**.** Glosten and Milgrom's model isolates what Avellaneda-Stoikov omits exactly as cleanly as that model isolates inventory.

An asset is worth either V^+ or V^-, each with probability one half. A fraction π of arriving traders know which; the rest trade for reasons of their own, buying or selling with equal probability. The maker holds no inventory — it can offload instantly — and is competitive, so it quotes at zero expected profit.

The informed always buy at the ask if the value is V^+ and sell at the bid if it is V^-. So a buy order is evidence for V^+, and the ask must be the conditional expectation of the value *given* that a buy arrived:

$$\text{ask} = \mathbb{E}\big[V \,\big|\, \text{buy}\big], \qquad \text{bid} = \mathbb{E}\big[V \,\big|\, \text{sell}\big],$$

and the spread between them is entirely the information content of the order. Working the conditional expectations out,

$$\text{spread} = \pi\big(V^+ - V^-\big), \tag{23.10}$$

linear in the fraction of informed flow and in how much there is to know.

Not that there is no inventory anywhere in (23.10) and no risk aversion, so this spread is not compensation for bearing anything — a risk-neutral maker with instant hedging quotes it too. And the maker earns nothing on average: it loses to the informed exactly what it makes from the rest. The spread is not profit, it is the price of not being able to tell the two apart.

Structure (Two costs, two models, no model with both)**.** The two models are complementary in a way that is more than tidy.

Avellaneda-Stoikov has inventory and no information: its flow is exogenous, its cost is variance, and its answer is a skew. Glosten-Milgrom has information and no inventory: its flow is endogenous, its cost is being picked off, and its answer is a width. Each is solvable, and each is solvable *because* it omits the other's mechanism — allowing the arrival intensity of (23.1) to depend on the future price would destroy the invariance theorem 23.3 needs, and giving Glosten-Milgrom's maker an inventory it must hold removes the zero-profit condition that pins its quotes.

Models with both exist and none is tractable. A desk therefore does not use a single model; it uses one model for the skew, another argument for the width, and measurement for everything else. The rest of this chapter is about the measurement, because that is where the actual work is.

23.4 Measuring Toxicity

If a fill is followed by the price moving against the maker, the spread it earned was an illusion. The standard measurement makes exactly that comparison.

Definition 23.4 (Markout)**.** For a fill of signed size Q (positive when the maker bought) at price p against a mid m_0, and the mid m_h observed h later, the *markout* at horizon h is

$$\operatorname{markout}(h) = Q\big(m_h - p\big). \tag{23.11}$$

At $h = 0$ it is the captured spread, $Q(m_0 - p)$, which is positive by construction.

The object of interest is the whole curve of markout against h, and its shape is the diagnosis. A flat curve means the maker keeps what it earned. A curve decaying towards zero means the price drifts against the fill and the spread is being given back. A curve going negative means the flow is worth less than nothing: the maker would have done better to decline.

Remark (The decomposition this implies)**.** Writing $s_{\text{eff}} = Q(m_0 - p)$ for the spread captured and $I(h) = Q(m_0 - m_h)$ for the price impact over the horizon, (23.11) rearranges to

$$\underbrace{\operatorname{markout}(h)}_{\text{what is kept}} = \underbrace{s_{\text{eff}}}_{\text{what is charged}} - \underbrace{I(h)}_{\text{what moves against}}. \tag{23.12}$$

This is microstructure's effective, realised and impact triple, and only the middle term is a decision. The desk sets s_{eff}; the market determines I; what the desk keeps is the residual. Widening the quote raises the first term and, through (23.1), reduces the number of fills — but it also changes *which* fills arrive, and generally not in the desk's favour, since the clients most willing to pay a wide spread are the ones most confident about direction.

Remark (Horizons, and what each one is measuring)**.** The choice of h is not a detail. A few seconds measures whether the desk is being arbitraged by faster participants. A few minutes measures whether the client had a short-lived signal. A day or more measures whether the client was simply right — which, for a pension fund hedging a liability, it may well be without anything improper having occurred.

Flow that is directional but slow is not toxic in the sense that matters; it is a hedging cost that the desk can price and manage by holding the position, and refusing it damages a franchise for no gain. Toxic flow is flow whose information decays faster than the desk can hedge. The relevant comparison is not markout against zero but markout against the desk's own time to flatten.

23.5 The Race the Short Markout Measures

A markout at a few seconds was described above as measuring whether the desk is being arbitraged by faster participants. The taxonomy of chapter 22 turns on it: the mispricings that vanish immediately are the ones somebody is racing for, and this is what the racing consists of.

The section is description rather than derivation, and is marked as such. Nothing in it is measured here — doing so honestly would need tick and venue data this book does not carry — so it is offered as an account of a mechanism, with the argument at the end being the part that is not merely reportage.

What the race is physically made of

The winning trade is decided by the time between a message arriving and an order leaving, and the industry has spent two decades compressing it.

The path is shortened first. Machines sit in the exchange's own data centre, and links between venues follow the straightest available line — microwave and millimetre wave rather than fibre, because light moves about half as fast again through air as through glass and the towers can run closer to a great circle than cable trenches do. Then the machine is shortened: the operating system is taken out of the path, and the decision logic is put into programmable hardware so that a packet can be parsed and an order emitted without a general purpose processor being involved. Then the message itself is shortened, by decoding the exchange's binary protocol directly, acting on a partial message before the remainder has arrived, and taking whichever of two redundant feeds happens to arrive first.

And on a price-time priority book there is a further asset that is not speed but is bought with it: queue position. Being early at a price level determines whether a resting order is filled at all, it accrues only by resting, and it is forfeited by cancelling — which is what makes the decision to pull a stale quote expensive rather than free.

Two roles, one race

The strategies divide into two.

A taker watches a fast instrument and trades against a stale quote in a slower related one — a future against its underlying, one venue against another. Its race is against the maker's cancel.

A maker posts quotes and tries not to be the stale one. Its race is against the taker's take, and every fill it loses shows up as a negative short-horizon markout: exactly the $I(h)$ of (23.12) at small h. So Definition 23.4 is not merely diagnostic of this activity, it is the scoreboard for it.

Why the spending does not stop, and who pays for it

Here is the part that is an argument rather than a description.

A mispricing is a bounded quantity available at a stale price, so the fastest participant takes essentially all of it and the second fastest takes nothing. A payoff of that shape has a known consequence: competitors bid for the advantage until the cost of holding it consumes the rent. The profit pool is fixed by how many stale quotes the market produces, and it does not grow when everyone gets faster — so the spending converts a flow of trading profit into a stock of fixed cost, and the participants end up no better off than before while the equipment gets more expensive.

That equilibrium is not a fact about human nature. It is a consequence of the market's *design*. A continuous limit order book resolves simultaneous orders by arrival time, so it makes an infinitesimal speed advantage decisive, and Budish, Cramton and Shim showed that this is what manufactures the race. Batching orders and crossing them at a single price every fraction of a second removes it, because orders arriving within a batch are simultaneous by construction and speed stops being the tiebreak. The arms race is a choice of matching rule, and a different rule would end it.

Meanwhile the cost is not borne by the participants who lose the race. A maker that is picked off must widen until the spread covers the expected loss, so the expenditure reappears in s_{eff} and is paid by everyone who trades, including those who have never heard of any of this.

Remark (How much of this reaches a rates desk)**.** Less than one might think, and the next section says why: most rates risk changes hands by request for quote rather than in a lit book, and a request that arrives at several dealers at once is not won by microseconds.

Where it does reach the desk is in the hedge. The benchmark futures and on-the-run bonds a swaps book hedges itself with *are* traded in central limit order books, so the desk meets the race not when it prices the client but when it lays the risk off — which puts the cost in the hedging leg of §23.8 rather than in the quote, and makes it a reason the hedge is partial rather than a reason the spread is wide.

23.6 Quoting on a Rates Desk

The general theory above is agnostic about the market. Rates market making has a structure that changes the problem materially, and in four ways.

Requests, not a book

Most rates risk does not change hands in a limit order book. A client sends a request for quote to several dealers at once and trades with the best price. Benchmark futures and on-the-run government bonds have genuine order books; swaps, swaptions, and everything structured are quoted on request.

That inverts the adverse selection problem in a way, because it produces the cost *without any informed trader at all*.

Calculation 23.5 (The winner's curse)**.** Suppose n dealers each value an instrument at its fair value plus an independent error of standard deviation s, reflecting different curves, different models and different axes. Nobody is informed; nobody has an edge; the errors have mean zero. The client trades with the best price.

The dealer who wins is the one whose error was most aggressive, so conditioning on having won selects the minimum of n draws. Measuring it:[2]

Dealers	2	3	5	10
Expected error of the winner, in units of s	-0.56	-0.84	-1.16	-1.54

So against two competitors the winner has underpriced by more than half its own pricing error, and against ten by more than one and a half times it. Eventually the growth is like $\sqrt{2\ln n}$, which is slow: ten competitors cost less than three times what two do, not five times.

[2] `marketmaking::winners_curse`, with the two-dealer case checked against its closed form $-1/\sqrt{\pi}$.

Structure (Winning is information)**.** This is separate from the informed-trader story, and the remedy is different.

In Glosten-Milgrom the maker loses because the counterparty knows something. Here nobody knows anything, and the maker still loses, because *the act of winning* is evidence about its own price. The fill is a draw from a conditional distribution — conditional on having been the most aggressive quote — and that distribution has a worse mean than the unconditional one.

The consequence is that a dealer's own historical fills are a biased sample of its own pricing, and biased in a direction that makes the desk look better than it is: the trades it did are the ones it priced most aggressively, and the ones it declined or lost are absent.

Practically it means the correction is a function of how many dealers are in competition, which the client knows and the dealer often does not. A desk that quotes the same margin into a two-dealer and a ten-dealer request is systematically mispricing one of them.

Who is on the other side

The client's identity is the single most informative variable available, and unlike almost everything else in this book it is known before the quote is made. A desk's flow divides roughly as follows.

- *Real money* — pension funds, insurers, liability-driven investors. Large, directional, slow, and usually one-way for months at a time. Markout at a day is often against the desk, because these clients are right about direction on a horizon far longer than the desk's. This is a hedging cost to be priced, not toxicity to be refused.

- *Corporates.* Hedging an issuance or a floating exposure. Benign, infrequent, and price-insensitive in the sense of (23.1) — the low-k flow that pays for the franchise.

- *Fast funds.* Small in size, high in frequency, and selecting on exactly the horizons the desk cannot hedge within. Markout decays within minutes. This is the flow the word toxic was invented for.

- *Other dealers.* Interdealer trades are usually somebody laying off risk they just took, so a request is evidence about the market's aggregate position, and one dealer's hedge is another's inventory.

A quote is on a portfolio of factors

This is the structural feature with no equity analogue, and it is a direct consequence of chapter 8's picture of the state as a curve.

A request for a ten-year swap is not a request on a single asset. The instrument's value depends on the whole curve, and its risk decomposes onto the factors of chapter 12 — level, slope, and whatever else the model carries. So a desk that has bought ten-year risk has not acquired a position in the ten-year point; it has acquired a particular combination of level and slope, and whether that combination offsets or compounds the existing book depends on what the book already holds.

Two consequences follow immediately.

The first is that the inventory variable q of Definition 23.1 is the wrong object. The relevant state is a vector of factor exposures, and the reservation offset (23.9) generalises to one in which $\gamma\sigma^2$ becomes a covariance matrix and q a vector. The skew on any one instrument then depends

on the whole book through that matrix, which is why a desk long the front end may quote a tight offer in the long end: the trade that reduces its net slope exposure is not the trade that reduces its position in the instrument requested.

The second is that some risk is far cheaper to shed than the rest. The benchmark points are liquid; the points between them are not. A position in a ten-year swap can be hedged in size within seconds; the same risk expressed at seven years and three months cannot, and must be hedged with a combination of benchmarks that leaves a residual. The residual is a basis risk, it is small per trade, and it accumulates in one direction because clients ask for the dates their liabilities fall on rather than the dates the market is liquid at.

Volatility that cannot be hedged

The swaption surface makes the same point more sharply. A desk quoting a seven-year option on a five-year swap has taken vega in a bucket where nothing liquid trades. It can hedge the delta immediately and the vega approximately, using whatever combination of quoted expiries and tenors comes closest, and the remainder is warehoused — carried deliberately, because the alternative does not exist.

Chapter 12's parameters are exactly what determine whether that remainder is dangerous. If the model's mean reversion is wrong, the mapping from the liquid buckets to the illiquid one is wrong, and the desk's belief about what it has hedged is wrong in a way no amount of care with the delta repairs. This is the practical cost of the identification problem that chapter describes: a parameter nobody can see from European prices is what decides whether a warehoused vega position is offset or doubled.

23.7 A Signal, and Who It Is Worth More To

Nothing so far has let the maker have a view. Definition 23.1 makes the mid a martingale precisely so that quoting and forecasting stay separable. What happens when they are not?

Suppose the maker believes fair value is μ rather than the observed mid S_t. The machinery already derived handles it without modification: the quotes were never centred on the mid but on the mid plus the reservation offset (23.9), and a signal enters exactly where inventory does — as a displacement of the centre, so it is the *deviation* $\mu - S_t$ that appears rather than μ itself.

$$\text{centre of the quoted market} \;=\; \underbrace{S_t}_{\text{mid}} \;\underbrace{-\,q\gamma\sigma^2(T-t)}_{\text{inventory}} \;\underbrace{+\,(\mu - S_t)}_{\text{signal}} \;=\; \mu - q\gamma\sigma^2(T-t). \tag{23.13}$$

Remark (A signal skews the quote, it does not tighten it)**.** Equation (23.13) moves the *centre* and leaves the width alone, which is the same conclusion the inventory term reached and for the same reason. The width answers a different question — how price sensitive the flow is, and how much risk the position carries — and a forecast bears on neither.

So a maker who becomes bullish shows a better bid and a worse offer. Read from one side that looks like tightening, but it is not a symmetric tightening which would be the wrong response. Tightening both sides buys more flow of both kinds, which is a decision to trade more rather than a decision to express a view; the maker would acquire the position it wants and the position it does not in equal measure, and the signal would earn nothing. The whole content of acting on a forecast is the asymmetry.

Structure (The same signal is worth more to a maker than to a taker)**.** There is an asymmetry here that is easy to state and easy to miss, and it explains why the same research is used differently on the two sides of a market.

A hedge fund or proprietary trader with a signal of size e must *cross the spread* to express it. The signal is worth e less the round trip, and chapter 22's calculation applies: below a threshold set by the cost there is no trade at all, however good the forecast.

A market maker with the same signal does not cross anything. It shifts (23.13) and waits, and the position arrives through flow it was going to quote on anyway — while being *paid* the spread for taking it. The signal's value is therefore e *plus* the spread earned on the trades it selects, and the breakeven signal strength is near zero rather than a round trip away.

Two consequences follow. A signal too weak to trade on can still be worth acting on if one is quoting, which is why market making desks fund research that would not clear a proprietary trader's threshold. And the maker's version is capacity constrained in a way the taker's is not: the position accumulates only as fast as clients happen to ask, so a strong short-lived signal cannot be expressed in size, and a maker with a genuinely urgent view has to cross the spread like everybody else — at which point it is a taker, and pays.

The asymmetry is not an inefficiency. It is payment for the service of §23.3: the maker is exposed to adverse selection on every quote it shows, and the right to monetise a signal cheaply is part of what compensates that exposure.

23.8 The Partial Hedge

Everything above converges on one decision. The desk cannot hedge completely — sometimes because the instrument does not exist, always because trading costs money — and the question is how much to leave.

Calculation 23.6 (How much to leave)**.** Take a unit exposure with variance Σ over the horizon it would be held. Hedging a fraction h costs $c\,h$ in spread and fees and leaves residual variance $\Sigma(1-h)^2$, penalised at $\gamma/2$ as in (23.2). Minimising

$$c\,h + \frac{\gamma}{2}\Sigma\,(1-h)^2$$

over h gives $c = \gamma\Sigma(1-h)$, so

$$1 - h^* = \frac{c}{\gamma\Sigma}, \tag{23.14}$$

truncated at zero and one.[3]

Remark (What (23.14) says)**.** The fraction left unhedged is the transaction cost divided by the risk penalty.

It is zero only if hedging is free. Any positive cost leaves a residual, so a completely hedged book is not a well-run one — it is one that has paid too much. At $\Sigma = 4\%$ and $\gamma = 2$, a cost of one per cent of notional leaves an eighth of the risk standing; four per cent leaves half of it.

It is total when the cost is large enough. Past $c = \gamma\Sigma$ the optimum is to hedge nothing at all, which is the illiquid vega case above: the bid-ask on the only available hedge exceeds the risk it removes, and the position is warehoused as a matter of arithmetic rather than of nerve.

[3] `marketmaking::optimal_hedge_fraction`, checked against a direct minimisation.

And it scales the wrong way for comfort. The residual is proportional to the cost and inversely proportional to the variance, so the positions a desk is most tempted to leave unhedged — illiquid, wide, and therefore expensive to trade — are exactly the ones (23.14) says to leave the most of. The formula agrees with the temptation, which is not a reason to distrust it, but is a reason to notice that the discipline has to come from a limit rather than from the optimisation.

Structure (Hedging is a projection, and the residual is what is left over)**.** Equation (23.14) is one-dimensional and the real problem is not, but the shape of the answer survives.

With several hedging instruments, the exposure is a vector and the hedgeable directions span a subspace. The best hedge ignoring costs is the orthogonal projection onto that subspace, and the unhedgeable part is the residual — which is exactly chapter 6's construction, where the market price of risk was a projection of excess returns onto the span of the volatility matrix and the residual was an arbitrage. Here the residual is not an arbitrage, because these instruments are not available at any price. It is a position.

Adding costs tilts the projection, shrinking it towards zero by the factor (23.14) in each direction according to what that direction costs to trade. So a desk's residual risk has two quite different components: the part that is unhedgeable because no instrument spans it, and the part that is unhedged because spanning it was not worth the fee. The first is a modelling constraint and cannot be reduced by spending more. The second is a choice, and can.

23.9 What Is Handed On

A day of market making ends with a book that nobody designed. Its composition is the sum of what clients asked for, which is not a portfolio anybody would choose; its hedges are approximate by decision, per (23.14); and its residual concentrates in the illiquid corners, because those are the corners where hedging was least worth the cost.

That object is what chapter 24 is about, and the chain should be stated in one place since it runs through the whole book. The quote came from a model, so it inherits the model's assumptions — chapter 20's question of whether the model has a view on what the trade is a bet on. The hedge came from the model's sensitivities, so it inherits the errors chapter 10 measures, where a wrong backbone produces a delta wrong by five per cent of notional in a consistent direction. And the residual is real, unavoidable and one-signed.

So the risk that arrives at the end of the chain is not the risk of a chosen position. It is the risk of a position accumulated by a process, hedged by a model, and left partly open on purpose. Measuring it requires knowing which of those three contributed what, which is why chapter 24 begins by insisting that a risk number without a decomposition is not a risk number.

References

- Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, 8(3), 217–224.

- Gueant, O., Lehalle, C.-A., & Fernandez-Tapia, J. (2013). Dealing with the inventory risk: a solution to the market making problem. *Mathematics and Financial Economics*, 7(4), 477–507.

- Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1), 71–100.

- Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6), 1315–1335.

- Easley, D., Lopez de Prado, M., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. *Review of Financial Studies*, 25(5), 1457–1493.

- Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: frequent batch auctions as a market design response. *Quarterly Journal of Economics*, 130(4), 1547–1621.

- Almgren, R., & Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of Risk*, 3(2), 5–39.