GRADUATE LECTURE NOTES

Lectures on Market Microstructure

Prices, Information, and Liquidity

Aryan Ayyar

www.aryanayyar.com

With a mathematical supplement on
fractional kernels and rough volatility

September 2026

Preface

These lectures grew out of the detailed notes I made while studying finance during my Master's degree at The Hong Kong Polytechnic University. Their common subject is the formation of prices when information, trading opportunities and financial resources are unevenly distributed. A price contains news, but it also records an agreement between people facing different constraints. Understanding that agreement requires following the mechanism that produces it.

The volume begins with the trading process itself. Rational-expectations models then explain how investors learn from equilibrium prices; transaction-price models distinguish that learning from bid–ask bounce and transitory trading costs. The strategic-trading lecture asks how an informed investor uses an information advantage without immediately revealing it. The final two core lectures examine the uncertain execution of limit orders and the dependence of market liquidity on the capital that intermediaries can commit. A short mathematical supplement develops the fractional-kernel material from the original collection.

The exposition is intended for a graduate reader who has seen calculus, elementary optimization, probability and regression. Conditional expectations and Gaussian updating recur throughout. The dynamic Kyle and rough-volatility sections additionally use stochastic-process ideas, introduced where they enter. The aim is to make the economic problem legible before solving the algebra: who chooses, what they know, what constrains them, and which equilibrium conditions connect their decisions.

The original notes have been reorganized, expanded and corrected. Definitions, sign conventions, timing and derivations are stated explicitly; worked examples and exercises with solutions have been added. Numerical examples are teaching constructions unless expressly identified otherwise. Discussions of research extensions distinguish their additional assumptions from the benchmark models. The cited papers remain the sources for the original contributions. These are personal lecture notes, not an official course publication of the university.

How to use the lectures

Read the first chapter before using the later models: its distinctions among payoff, public valuation, quote, transaction price, trade flow and resting depth prevent several common misunderstandings. Lectures 2–4 form a sequence on information and prices. Lecture 5 can be studied after Lecture 1 and an introduction to probability; Lecture 6 is most useful after the equilibrium lectures. The mathematical supplement is self-contained relative to the market-microstructure sequence.

For each model, reconstruct the timeline and derive the first-order or clearing condition before substituting numbers. Use the exercises to test the interpretation as well as the arithmetic. A coefficient is informative only after its units, information set and counterfactual have been named.

Contents

Lecture 1

How trading turns information into prices

A share has a payoff, but it does not have a transaction price until somebody agrees to trade it. The agreement depends on who arrives, what that person knows, how urgently the trade must be completed, and which orders are already available. Market microstructure studies this passage from investors' demands to actual transactions. Its central economic question is therefore concrete: why does this quantity trade at this price, through this mechanism, at this moment?

To answer this question, we need both a mechanism and an account of traders' incentives. We begin with the distinction between value, quotes and transaction prices, then follow an order through a book and a call auction. Those examples explain how submitted orders become trades. Learning from trade direction explains why dealers revise their quotes, and a stochastic model finally describes how the book evolves between transactions.

1.1 Value, information and the price of immediacy

Let V be the terminal cash payoff of one share. Set the risk-free interest rate to zero for this discussion, and let $\mathcal{H}_t$ denote the public information available by time t. Define the public conditional valuation

$$m_t = \mathbb{E}[V \mid \mathcal{H}_t]. \tag{1.1}$$

This is a belief about a payoff. It need not be a price at which an investor can execute an order. A dealer may quote a bid b_t and an ask a_t, with $b_t \leq a_t$. An incoming seller receives the bid; an incoming buyer pays the ask. The quoted midpoint and spread are

$$M_t = \frac{a_t + b_t}{2}, \qquad s_t = a_t - b_t.$$

The midpoint M_t and the conditional valuation m_t coincide only under additional assumptions about the quoting rule. We keep their notation separate.

For a fixed integrable payoff V and nested information sets, the tower property gives

$$\mathbb{E}[m_{t+1} \mid \mathcal{H}_t] = \mathbb{E}[\mathbb{E}[V \mid \mathcal{H}_{t+1}] \mid \mathcal{H}_t] = m_t.$$

Thus m_t is a martingale under these assumptions. If its increments are square-integrable, innovations at different dates are uncorrelated. Neither independence nor normality follows from the martingale property. With risk premia, dividends, discounting or a changing payoff horizon, one must first specify the correct valuation object before applying this argument. Likewise, the logarithm of a conditional mean is generally different from the conditional mean of a logarithm.

A useful elementary observation equation is

$$p_t = m_t + cD_t, \qquad D_t \in \{-1, +1\}, \quad c > 0, \tag{1.2}$$

where p_t is the transaction price and $D_t = +1$ denotes a buyer-initiated trade. With symmetric quotes, c is the half-spread. Even if m_t is a martingale, changes in D_t can make transaction-price changes negatively autocorrelated. Chapter 3 develops this distinction.

1.1.1 What a market maker earns

A dealer who buys one share at $m - c$ and later sells it at $m + c$, while value stays fixed, earns $2c$ before costs. This is a completed round trip consisting of two transactions. It is not a guaranteed profit on each trade. If value falls after the dealer buys, the inventory loss can exceed the spread; if an informed investor buys before good news, the dealer may sell too cheaply. Fees, hedging expenses and capital requirements also matter.

The economic spread can compensate for order processing, adverse selection and inventory-bearing costs. These mechanisms can coexist, but a quoted spread alone does not identify their respective contributions. Competitive, risk-neutral dealers can face an adverse-selection spread even while earning zero expected profit conditional on each trade direction. This point will emerge directly from Bayes' rule below.

1.2 The order book and the cost of trading size

The spread measures the cost of demanding a small amount of immediacy. A larger order must also confront how much is offered at each price. The order book makes this quantity dimension visible.

A buy limit order specifies a maximum purchase price; a sell limit order specifies a minimum sale price. A nonmarketable order rests in the book until it is executed, cancelled or expires. A marketable limit order can execute immediately against standing orders at prices satisfying its limit. The adjective "limit" therefore describes a price restriction, not a promise to wait.

Consider the illustrative book in Table 1.1. Quantities are shares; prices are currency units per share. The tick size is 0.05.

The best bid is 236.30, the best ask 236.35, the midpoint 236.325, and the quoted spread 0.05. A market sell of 2,000 shares consumes 1,141 shares at the best bid and 859 at the

Table 1.1: An illustrative limit order book before a trade

Bid quantity	Bid price	Ask price	Ask quantity
1,141	236.30	236.35	600
1,479	236.25	236.40	900
2,000	236.20	236.45	1,500

next bid, assuming no concurrent arrivals or cancellations. Its average execution price is

$$\bar{p}_{\text{sell}} = \frac{1141(236.30) + 859(236.25)}{2000} = 236.278525.$$

The last execution occurs at 236.25, while 620 shares remain at that bid. The seller's average shortfall from the initial midpoint is 0.046475 per share, or approximately 1.967 basis points. By contrast, the last execution is approximately 3.174 basis points below the initial midpoint. Average execution cost and last-price movement measure different things.

The surviving best bid is now 236.25. If the ask remains 236.35, the new midpoint is 236.30: its movement is only 0.025, because only one side moved. A comparison of the initial and final transaction prices would require knowing the initial transaction price, which the initial book does not supply. This is why an empirical price-impact calculation must name its benchmark.

Figure 1.1 places the execution schedule beside these benchmarks. The average price summarizes the whole order; the last price records only the least favorable executed portion.

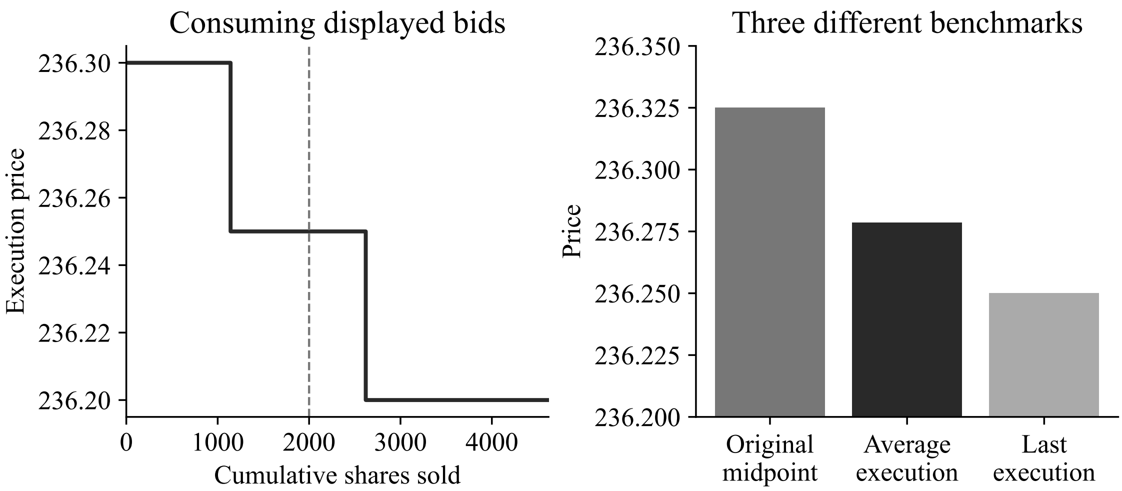

Figure 1.1: Selling 2,000 shares into the fixed displayed book. Left: the price available at successive cumulative quantities; the dashed line marks the order size. Right: the pretrade midpoint, volume-weighted average execution price and last execution price. They answer different cost and price-movement questions even for the same trade.

1.2.1 Depth is a schedule, not just a total

Let q_j^b denote bid quantity at price p_j. The cumulative quantity available to a seller at prices at least p is

$$B(p) = \sum_{j:p_j \geq p} q_j^b.$$

Similarly, the quantity available to a buyer at prices no greater than p is $A(p) = \sum_{j:p_j \leq p} q_j^a$. Walking through these step functions determines the mechanical cost of consuming the displayed book. A large total quantity far from the best quotes offers little protection to a small order at the touch. Conversely, a narrow spread can coexist with very little depth.

These calculations hold the book fixed. They do not include hidden orders, dealer replenishment, cancellations, strategic responses or information-driven quote revisions. A static depth curve measures a particular execution opportunity, not the full counterfactual market response to a large trading program.

1.2.2 A call auction

A call auction accumulates orders and selects one clearing price under an announced rule. For limit orders, define cumulative willingness to buy and sell at candidate price p by

$$D(p) = \sum_{j:\ell_j^b \geq p} q_j^b, \qquad S(p) = \sum_{j:\ell_j^a \leq p} q_j^a.$$

The executable quantity is $Q(p) = \min\{D(p), S(p)\}$. A common rule first maximizes $Q(p)$, then resolves ties using imbalance and reference-price rules. The exact tie-breaker belongs to the venue's auction specification; it cannot be inferred from the equations alone.

For example, suppose bids are 100 shares at 101 and 150 at 100, and offers are 120 at 99 and 100 at 100. At prices 99, 100 and 101, executable quantities are respectively 120, 220 and 100. The volume-maximizing clearing price is 100, and 220 shares can trade. Eligible demand is 250, so a further allocation rule determines which 30 shares of demand remain unfilled.

The uncrossed book in Table 1.1 is a continuous-market snapshot: every bid is below every offer. Without additional orders it has no positive executable quantity in a call. A clearing-price calculation therefore needs orders whose price limits actually overlap. Displayed quantities alone do not determine a subsequent price movement; their side, price and interaction with incoming orders matter.

1.3 Signed trades and book events

The book example tells us where a trade executes. To ask what the trade reveals, we must also identify which side chose to execute immediately.

Every completed transaction has a buyer and a seller. "Buyer-initiated" identifies the side demanding immediacy, not an excess count of buyers over sellers. For trade i of size $v_i > 0$, let $D_i = +1$ for buyer initiation and -1 for seller initiation. Signed trade volume over interval I is

$$Q_I = \sum_{i \in I} D_i v_i.$$

This is a flow accumulated over an interval. Displayed bid or ask depth is a stock measured at an instant. The two need not have the same sign convention.

When the exchange does not report the initiating side, a quote test compares the transaction price with the prevailing midpoint. Above-midpoint trades are classified as buys; below-midpoint trades as sells. A tick test can classify midpoint trades using the most recent nonzero price change. Lee and Ready [29] study the classification and timestamp-alignment problems behind this approach. An inferred sign remains an estimate, especially when quotes and trades are not synchronized. A delay chosen for an old reporting system is not automatically appropriate for a modern event feed.

Signed executed volume should also be distinguished from order-book imbalance measures that include limit additions and cancellations. Cancelling a large bid can weaken displayed buying interest without producing a trade. The distinction becomes essential whenever a model moves from transaction data to book-event data.

1.4 Adverse selection: learning from the direction of a trade

The sequential-trade logic of Glosten and Milgrom [17] can be taught with a two-value example. Let $V \in \{v_L, v_H\}$, with $v_H > v_L$, and let the current public probability of the high value be $\pi \in (0, 1)$. An arriving trader is informed with probability $\eta \in [0, 1)$, independently of value. The informed trader knows V and buys in the high state and sells in the low state. An uninformed trader buys or sells with equal probability, independently of value. All trades are for one share.

The probability of a buy is

$$\Pr(B) = \pi \frac{1+\eta}{2} + (1-\pi)\frac{1-\eta}{2} = \frac{1-\eta+2\eta\pi}{2}, \tag{1.3}$$

$$\Pr(S) = \frac{1+\eta-2\eta\pi}{2}. \tag{1.4}$$

Equivalently, with low-state prior probability $\delta = 1-\pi$, one has $\Pr(B) = 1/2+\eta(1-2\delta)/2$. The underlying calculation is unchanged when the state used to express the prior changes.

Bayes' rule gives the high-state posteriors

$$\pi_B = \frac{\pi(1+\eta)}{1-\eta+2\eta\pi}, \qquad \pi_S = \frac{\pi(1-\eta)}{1+\eta-2\eta\pi}.$$

Competitive risk-neutral dealers quote conditional expected payoffs:

$$a = v_L + (v_H - v_L)\pi_B, \qquad b = v_L + (v_H - v_L)\pi_S.$$

For an interior prior and a positive informed fraction, $\pi_B > \pi > \pi_S$. A buy raises the dealer's valuation; a sell lowers it. The dealer must condition on the event that the quote is accepted, because acceptance is informative.

For the symmetric prior $\pi = 1/2$, these formulas simplify to

$$a = \frac{v_H + v_L}{2} + \frac{\eta(v_H - v_L)}{2}, \quad b = \frac{v_H + v_L}{2} - \frac{\eta(v_H - v_L)}{2}, \quad a - b = \eta(v_H - v_L).$$

Take $v_L = 90$, $v_H = 110$ and $\eta = 0.2$. The public mean is 100, the ask is 102, and the bid is 98. The dealer's expected payoff from selling at the ask is zero because $\mathbb{E}[V \mid B] = 102$. An informed buyer in the high state nevertheless expects a profit of 8. The dealer's losses to informed trades are offset by trades made for noninformational reasons.

The next trader arrives after the public belief is updated. A sequence of trades therefore changes both the center of the quotes and the uncertainty that generates the spread. This connects trading to price discovery. It does not require the dealer to know the payoff in advance, nor does it imply that every investor with a useful signal is legally an insider.

1.5 Permanent revision and transitory trading cost

The preceding example gives a trade two roles: it reveals information and supplies immediacy. These roles suggest a useful empirical question. How much of a transaction-price movement persists in the market's valuation, and how much disappears when the next trade arrives from the other side? In a simple illustrative specification, let

$$m_t - m_{t-1} = z_0 D_t + z_1 D_t v_t + u_t, \tag{1.5}$$

$$p_t = m_t + c_0 D_t + c_1 D_t v_t. \tag{1.6}$$

Here z_0, z_1 govern persistent valuation revisions, while c_0, c_1 govern contemporaneous transitory deviations. Differencing yields

$$\begin{aligned}\Delta p_t &= z_0 D_t + z_1 D_t v_t + u_t \\ &\quad + c_0(D_t - D_{t-1}) + c_1(D_t v_t - D_{t-1} v_{t-1}).\end{aligned} \tag{1.7}$$

The equation is obtained by subtraction; an empirical interpretation additionally needs assumptions about innovations and order-flow predictability. It illustrates the distinction emphasized by Glosten and Harris [16]: a trade can alter the valuation and also occur away from that valuation.

When order flow is predictable, it is natural to separate its innovation. Set $q_t = D_t v_t$ for signed volume at trade t, and let L be the number of lags used for prediction. Write a linear forecasting equation

$$q_t = \alpha + \sum_{j=1}^{L} \beta_j \Delta p_{t-j} + \sum_{j=1}^{L} \gamma_j q_{t-j} + e_t.$$

The residual e_t is unexpected relative to the chosen linear forecasting variables. It equals the innovation in the full conditional expectation only when that linear specification is correct. An illustrative price equation is

$$\Delta p_t = \lambda e_t + c(D_t - D_{t-1}) + u_t.$$

The sign-difference term is transitory in this specification: summing it over time telescopes to $c(D_T - D_0)$. The innovation term accumulates. Only the latter is assigned a persistent effect in this equation. The coefficient λ measures a response to the particular order-flow innovation being modeled; identifying it with the structural Kyle coefficient would additionally require matching information sets, quantities and equilibrium assumptions.

1.6 A stochastic representation of the book

Trade-based equations leave out events that change available liquidity without producing a transaction. To include those events, treat the book itself as the state, evolving through additions, executions and cancellations. The following unit-order representation follows the modeling logic of Cont, Stoikov and Talreja [8].

Let prices lie on a grid $p_j = j\delta_p$, $j = 1, \dots, n$. Define $X_j(t)$ to be positive for resting sell quantity and negative for resting buy quantity. Then

$$j_A(t) = \inf\{j : X_j(t) > 0\}, \qquad j_B(t) = \sup\{j : X_j(t) < 0\}.$$

When both sides exist, $a_t = \delta_p j_A(t)$ and $b_t = \delta_p j_B(t)$. Formal empty-book conventions such as $j_A = n + 1$ and $j_B = 0$ are useful boundary markers, but they are not executable quotes. A reported midpoint should be marked unavailable when either side is empty.

For unit orders, the state transitions are as follows. A cancellation acts at its own resting level j, not automatically at the best quote.

Event	State change
Resting buy added at $j < j_A$	$X_j \mapsto X_j - 1$
Resting sell added at $j > j_B$	$X_j \mapsto X_j + 1$
Market buy, if ask exists	$X_{j_A} \mapsto X_{j_A} - 1$
Market sell, if bid exists	$X_{j_B} \mapsto X_{j_B} + 1$
Resting buy cancelled at j	$X_j \mapsto X_j + 1$
Resting sell cancelled at j	$X_j \mapsto X_j - 1$

Suppose limit orders at distance i ticks from the opposite best quote arrive at rate λ_i, market orders arrive at rate μ, and each resting unit cancels at rate θ_i. The total cancellation intensity at a level containing q eligible units is $\theta_i q$. A specification such as $\lambda_i = ki^{-\alpha}$ is a power law in distance, not an exponential law. This distance convention presumes that the opposite quote exists; an implementation covering empty-side states must separately specify its boundary submission rules.

If event types have intensities $r_e(x)$ and state increments Δ_e, the infinitesimal generator is

$$(\mathcal{L}f)(x) = \sum_e r_e(x)\{f(x + \Delta_e) - f(x)\}.$$

This equation collects the probability-weighted changes of a function of the book over a short interval. It is the bridge from an event description to probabilities of quote moves. The probability that a particular resting order executes before a quote change additionally requires its priority or queue position and the associated transition rules; aggregate displayed depth alone does not specify that tagged-order event.

1.6.1 Estimating rates with the correct exposure

The event description becomes empirically useful once its rates can be estimated. An arrival rate needs a denominator that measures opportunities for arrival; a per-unit cancellation rate needs a denominator that also measures how much resting quantity was exposed.

For a Poisson count N_i observed over eligible exposure time T_i, the log-likelihood, up to terms independent of λ_i, is $N_i \log \lambda_i - \lambda_i T_i$. Maximization gives $\widehat{\lambda_i} = N_i/T_i$. The denominator is the time during which an event at that relative level could have been observed under the specified definition. It need not equal the full trading day when quotes move, data are missing or the relevant state is absent.

The distance profile can then be fitted to these estimated rates. For example, the unweighted least-squares specification in the study model chooses

$$(\widehat{k}, \widehat{\alpha}) \in \operatorname*{arg\,min}_{k \geq 0,\, \alpha \geq 0} \sum_{i \in \mathcal{D}} \big(\widehat{\lambda}_i - k i^{-\alpha}\big)^2,$$

where $\mathcal{D}$ contains observed positive tick distances. At least two distinct distances with informative rate estimates are needed to distinguish scale from slope. For fixed α, differentiation gives

$$\widehat{k}(\alpha) = \frac{\sum_{i \in \mathcal{D}} \widehat{\lambda}_i i^{-\alpha}}{\sum_{i \in \mathcal{D}} i^{-2\alpha}}.$$

One can substitute this expression back into the objective and search over the slope. Unequal exposures make the rate estimates unequally precise, so unweighted least squares is a particular fitting criterion, not an automatic maximum-likelihood estimator for the original counts.

For per-unit cancellations, the analogous estimator is

$$\widehat{\theta}_i = \frac{N_{c,i}}{\int_0^T q_i(t)\, dt}.$$

Dividing by clock time alone estimates an aggregate cancellation rate, not a per-resting-unit rate. The integral has units of unit-orders times time.

If a unit-order model uses typical limit-order size $\bar{v}_l$, an approximate market-order rate in comparable units is $(N_m/T)(\bar{v}_m/\bar{v}_l)$. This converts average volume flow; it does not reproduce the distribution of individual order sizes. Calibration should therefore report whether events, shares or normalized units are being counted. A fitted event model is a useful description of book mechanics, while strategic responses to changes in fees or priority require further economic structure.

1.7 From mechanism to equilibrium

In a sequential-trade model, information reaches the dealer through the direction of individual trades. In a strategic-trade model, an informed investor anticipates the price response to the size of the submitted order. If a dealer uses $p = \mu + \lambda(x + u)$ and noise demand u has conditional mean zero, an insider knowing $V = v$ has expected profit

$$x\{v - \mu - \lambda x\}.$$

For $\lambda > 0$, the second derivative is $-2\lambda < 0$, and the best response is $x = (v - \mu)/(2\lambda)$. The price-impact term reduces profit as the order grows. The dealer's Bayesian pricing condition must still be imposed to determine λ; the best response alone is not an equilibrium. Chapter 4 develops that fixed point and its dynamic extensions.

1.8 Exercises

Exercise 1.1 (Execution and benchmarks)**.** In Table 1.1, execute a market buy of 1,000 shares against the unchanged initial book. Find the average price, the last price and the average cost relative to the midpoint. Identify which displayed quantities remain.

Exercise 1.2 (A buy is news)**.** In the two-value model, take $v_L = 90$, $v_H = 110$, $\eta = 0.2$ and $\pi = 0.7$. Compute the buy probability, the two posterior probabilities, and the quotes. Is the quote midpoint equal to the prior mean?

Exercise 1.3 (Cancellation exposure)**.** A price-distance bin records 240 cancelled unit orders. Its average eligible resting depth over 120 minutes is 20 units. Estimate the per-unit cancellation rate and the aggregate rate. Explain why the numbers differ.

Exercise 1.4 (Persistent and transitory effects)**.** Sum equation (1.7) from $t = 1$ to T. Which terms telescope? What assumption would fail if every predictable buy generated a positive expected innovation in m_t?

1.9 Solutions and discussion

1.Buying 600 shares at 236.35 and 400 at 236.40 gives an average price of 236.37 and last price of 236.40. Average midpoint cost is 0.045 per share, approximately 1.904 basis points. The remaining asks are 500 shares at 236.40 and 1,500 at 236.45; bids are unchanged.

2.The buy probability is 0.54. The posteriors are $\pi_B = 0.84/1.08 = 7/9$ and $\pi_S = 0.56/0.92 = 14/23$. Hence $a \approx 105.556$ and $b \approx 102.174$. Their midpoint is approximately 103.865, not the prior mean of 104. Symmetry around the public mean depended on the symmetric prior; it is not a general identity.

3.The depth exposure is $20 \times 120 = 2400$ unit-minutes, giving $\widehat{\theta} = 0.1$ per minute per unit. The aggregate cancellation rate is $240/120 = 2$ units per minute. Multiplying the per-unit rate by average depth reconciles them.

4.The transitory contributions reduce to $c_0(D_T - D_0) + c_1(D_T v_T - D_0 v_0)$. The persistent contributions sum over all dates. If the valuation is a public conditional-expectation martingale, its innovation must have conditional mean zero. Predictable flow must already be priced or offset in the innovation specification.

Lecture 2

Rational Expectations: Learning from Equilibrium Prices

A price is both a term of exchange and a message. A lower price makes an asset cheaper, but it may also tell an investor that other traders have learned something unfavourable. The first effect encourages buying; the second can discourage it. Understanding how these two effects interact is the central task of a rational-expectations model.

This lecture moves from portfolio choice to information aggregation. We first derive the demand of a risk-averse investor and recover the covariance logic of the capital asset pricing model. We then introduce private signals, distinguish a price-clearing calculation from a rational-expectations equilibrium, and solve a noisy linear equilibrium. The final sections separate market depth, price volatility, and informational efficiency. These are related outcomes, but they are not interchangeable measures of market quality.

2.1 What must be consistent in equilibrium?

Suppose an investor knows a private signal s_i and can observe the market price p. Her relevant information is then $\mathcal{I}_i = \sigma(s_i, p)$, together with the public information understood throughout. The notation $\sigma(s_i, p)$ means the information generated by these observations. If she chooses a position using only s_i, even though p conveys additional information, her calculation does not use everything she knows.

Definition 2.1 (Competitive rational-expectations equilibrium)**.** A competitive rational-expectations equilibrium consists of a price function, conditional beliefs, and demand functions such that:

(i) each investor's demand maximizes expected utility conditional on her information, including the observed price;
(ii) these conditional beliefs agree with the probability law generated by the actual price function and the underlying uncertainty;
(iii) aggregate demands clear the market in every relevant state.

Each investor treats the quoted price as given when choosing her position.

The consistency requirement is about conditional beliefs and optimal decisions. It does not mean that everyone agrees about fundamental value, that everyone holds the same portfolio, or that no one regrets a decision after an uncertain payoff is realized. Two investors may rationally disagree because they observe different signals. An optimal risky investment can also lose money.

Price-taking and rational expectations address different questions. Price-taking asks whether an investor internalizes her own effect on price. Rational expectations asks whether her beliefs correctly incorporate the information she observes. A competitive investor can learn rationally from price without strategically manipulating it. Conversely, a strategic model requires an explicit description of how orders affect prices; merely writing down a market-clearing equation does not supply that description.

2.2 Portfolio choice with a risky endowment

Before asking what an investor learns from price, hold her beliefs fixed and ask what she wants to own. This portfolio decision is the building block of the equilibrium: different information sets will change the investor's conditional mean and variance, while the underlying risk-bearing calculation remains the same.

There are two dates, 0 and T. A risk-free asset costs one at date 0 and pays $R_f = 1 + r_f > 0$ at T. A risky asset costs p and pays a random amount F. The investor begins with I shares and B units of the risk-free asset. Let X denote her *net purchase* of risky shares. Her final risky holding is $I + X$, not X alone.

A purchase is financed by reducing the bond position. Thus initial marked-to-market wealth and terminal wealth are

$$W_0 = Ip + B, \tag{2.1}$$

$$W_T = (I + X)F + (B - Xp)R_f. \tag{2.2}$$

We allow borrowing and short selling. Position constraints would replace some equalities below with inequalities or corner solutions. The price is treated as fixed in this individual decision problem.

Let utility be constant absolute risk aversion, or CARA:

$$U(w) = -\exp(-Aw), \qquad A > 0.$$

This is a risk-*averse* investor. A risk-neutral investor instead maximizes expected wealth. The distinction matters because unconstrained risk-neutral demand is generally unbounded whenever expected payoff differs from purchase cost compounded at the risk-free rate.

Conditional on information $\mathcal{I}$, assume

$$F \mid \mathcal{I} \sim \mathcal{N}(\mu_{\mathcal{I}}, v_{\mathcal{I}}), \qquad v_{\mathcal{I}} > 0.$$

For a normally distributed random variable W with mean μ_W and variance v_W, the normal moment-generating function gives

$$\mathbb{E}[-e^{-AW}] = -\exp\left(-A\mu_W + \frac{A^2}{2}v_W\right).$$

Maximizing expected utility is consequently equivalent to maximizing the certainty equivalent $\mu_W - Av_W/2$. Using (2.2), the terms depending on X are

$$X(\mu_{\mathcal{I}} - R_f p) - \frac{A}{2}(I + X)^2 v_{\mathcal{I}}.$$

The first derivative is $\mu_{\mathcal{I}} - R_f p - A(I + X)v_{\mathcal{I}}$; the second derivative is $-Av_{\mathcal{I}} < 0$. We therefore obtain a unique unconstrained optimum:

$$X^*(p, \mathcal{I}) = \frac{\mu_{\mathcal{I}} - R_f p}{Av_{\mathcal{I}}} - I. \tag{2.3}$$

The numerator is expected excess payoff per share. The denominator combines risk aversion with payoff uncertainty. The final term subtracts what the investor already owns. An investor who receives ten shares as an endowment does not ordinarily want to buy the same additional quantity as an otherwise identical investor who begins with none.

Rearranging the first-order condition produces the investor's marginal valuation:

$$p = \frac{\mu_{\mathcal{I}} - A(I + X)v_{\mathcal{I}}}{R_f}. \tag{2.4}$$

For a positive final holding, risk reduces price below discounted expected payoff. This minus sign is economically essential: a buyer who must bear more risk demands a lower purchase price, other things equal. When $p > 0$, the expected simple return $r = (F - p)/p$ satisfies

$$\mathbb{E}[r \mid \mathcal{I}] = r_f + \frac{A(I + X)v_{\mathcal{I}}}{p}.$$

A positive risk premium in expected return and a negative risk adjustment in price are two descriptions of the same trade-off.

2.2.1 The same first-order condition through Stein's identity

The normal–CARA shortcut is convenient, but it is useful to see where the covariance enters. Differentiating expected utility directly gives

$$\mathbb{E}[U'(W_T)(F - R_f p) \mid \mathcal{I}] = 0.$$

For jointly normal (W_T, F), Stein's identity, under the usual differentiability and integrability conditions, implies

$$\mathrm{Cov}(U'(W_T), F \mid \mathcal{I}) = \mathbb{E}[U''(W_T) \mid \mathcal{I}]\,\mathrm{Cov}(W_T, F \mid \mathcal{I}).$$

Writing the expectation of a product as the product of expectations plus covariance yields

$$R_f p = \mu_{\mathcal{I}} + \frac{\mathbb{E}[U''(W_T) \mid \mathcal{I}]}{\mathbb{E}[U'(W_T) \mid \mathcal{I}]}\,\mathrm{Cov}(W_T, F \mid \mathcal{I}).$$

Under CARA, the ratio is $-A$, and the covariance is $(I + X)v_{\mathcal{I}}$. This recovers (2.4). The identity relates a nonlinear marginal-utility covariance to a payoff covariance; it does not identify those two objects without the derivative term.

Example 2.1 (Risk bearing and the purchase decision)**.** Suppose $\mu_{\mathcal{I}} = 110$, $v_{\mathcal{I}} = 100$, $A = 0.02$, $R_f = 1.05$, and the investor's desired final holding is two shares. The marginal price is

$$p = \frac{110 - 0.02(2)(100)}{1.05} = 100.9524.$$

Expected return is about 8.9623%: the 5% risk-free rate plus $4/100.9524 = 3.9623\%$. If the investor already owns one share, her net purchase is one share. The allocation and its financing, not merely expected payoff, determine her demand.

2.3 From individual risk bearing to the CAPM

An individual's willingness to pay depends on the risk of her portfolio. In equilibrium, however, all investors' portfolios must add up to the assets available. Aggregating their first-order conditions reveals which risk cannot be removed by trading among themselves, and therefore earns compensation in market prices.

Now let F and p be vectors of payoffs and prices for K risky assets. Suppose the payoff vector is jointly normal and investors share its mean vector μ and positive-definite covariance matrix Σ, face the same risk-free return, and can trade without constraints. Joint normality, rather than normal marginal distributions alone, preserves the normal–CARA certainty-equivalent calculation for every portfolio. Write h_i for investor i's final holdings and A_i for her CARA coefficient. The vector first-order condition is

$$\mu - R_f p = A_i \Sigma h_i.$$

Define total risk tolerance $T = \sum_i A_i^{-1}$ and aggregate risky supply $Q = \sum_i h_i$. Summing individual demand gives

$$Q = T\Sigma^{-1}(\mu - R_f p), \qquad \mu - R_f p = \frac{1}{T}\Sigma Q.$$

The market's effective absolute risk aversion is therefore $A_M = 1/T$. Asset j earns an expected excess payoff proportional to its covariance with the aggregate risky payoff $F_M = Q'F$:

$$\mu_j - R_f p_j = A_M \operatorname{Cov}(F_j, F_M).$$

Let $P_M = Q'p > 0$, and define simple returns $r_j = F_j/p_j - 1$ and $r_M = F_M/P_M - 1$, for positive individual prices. Dividing by p_j gives

$$\mathbb{E}[r_j] - r_f = A_M P_M \operatorname{Cov}(r_j, r_M).$$

Applying the same expression to the market portfolio gives

$$\mathbb{E}[r_M] - r_f = A_M P_M \operatorname{Var}(r_M).$$

Provided market-return variance is positive, division eliminates the common price of risk:

$$\mathbb{E}[r_j] - r_f = \beta_j\big(\mathbb{E}[r_M] - r_f\big), \qquad \beta_j = \frac{\operatorname{Cov}(r_j, r_M)}{\operatorname{Var}(r_M)}. \tag{2.5}$$

This is the familiar CAPM covariance restriction associated with [32]. The derivation here is a normal–CARA route to that restriction, not an assertion that these are the only assumptions under which a CAPM can be obtained.

Notice the economic aggregation step. The risk relevant to market pricing is covariance with the risky payoff that investors collectively must hold, not covariance with an arbitrarily selected individual's wealth. A highly volatile asset can have a modest risk premium if its fluctuations diversify aggregate risk. An asset that performs poorly precisely when aggregate wealth is low is less attractive to risk-averse holders and must offer compensation.

This symmetric-information benchmark is useful because it makes the next change explicit. Once investors observe different information, they need not share μ or Σ. The single common-belief aggregation above cannot simply be carried over unchanged. We must specify who knows what and how price itself changes those beliefs.

2.4 A private signal and Bayesian updating

Information changes demand by changing the payoff distribution the investor uses. We first isolate learning from a private observation. Learning from the market price will use the same probability calculation, but the observation will itself be generated by other investors' demands.

Begin with a signal-as-measurement model:

$$F \sim \mathcal{N}(\bar{F}, \sigma_F^2), \qquad S = F + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma_\eta^2),$$

where F and η are independent and both variances are positive. Conditional on F, S has mean F and variance σ_η^2. Unconditionally, its variance is $\sigma_F^2 + \sigma_\eta^2$. Confusing the conditional noise variance with the unconditional signal variance produces incorrect updating weights.

Bayes' rule says that the posterior density is proportional to likelihood times prior. Terms not depending on F can be absorbed into the normalizing constant:

$$f(F \mid S) \propto \exp\left[-\frac{1}{2}\left\{\frac{(F-\bar{F})^2}{\sigma_F^2} + \frac{(S-F)^2}{\sigma_\eta^2}\right\}\right].$$

Expanding the quadratic, the coefficient of F^2 is $\sigma_F^{-2} + \sigma_\eta^{-2}$ and the coefficient of $-2F$ is $\bar{F}\sigma_F^{-2} + S\sigma_\eta^{-2}$. Completing the square therefore gives

$$v_I = \mathrm{Var}(F \mid S) = \left(\frac{1}{\sigma_F^2} + \frac{1}{\sigma_\eta^2}\right)^{-1}, \tag{2.6}$$

$$\mu_I = \mathbb{E}[F \mid S] = v_I\left(\frac{\bar{F}}{\sigma_F^2} + \frac{S}{\sigma_\eta^2}\right) = \bar{F} + \kappa(S - \bar{F}), \tag{2.7}$$

$$\kappa = \frac{\sigma_F^2}{\sigma_F^2 + \sigma_\eta^2}. \tag{2.8}$$

Precision means inverse variance. Independent information adds precision; the posterior mean is an arithmetic average of the prior mean and observed signal, weighted by their precisions. The reciprocals belong to the variances, not to the levels being averaged. A more precise signal receives more weight. A more precise prior makes the investor less responsive to the same signal surprise.

Example 2.2 (A good signal is not a perfect signal)**.** If the prior mean is 100, prior variance is 16, signal-noise variance is 9, and the observed signal is 110, then $\kappa = 16/25 = 0.64$. The posterior mean is 106.4, not 110, and posterior variance is $144/25 = 5.76$. The investor moves toward the signal while retaining uncertainty about the payoff.

The same result can be written as a Gaussian projection:

$$\mathbb{E}[F \mid S] = \mathbb{E}[F] + \frac{\operatorname{Cov}(F,S)}{\operatorname{Var}(S)}(S - \mathbb{E}[S]).$$

Here the linear projection equals the conditional expectation because the variables are jointly normal. That equality will require more care in the transaction-price lecture, where trade signs are binary rather than Gaussian.

2.5 Clearing a market with naive expectations

We now bring informed and uninformed investors together. An intermediate calculation will show why clearing alone is insufficient: price can balance their orders while simultaneously revealing information that makes one group's original orders suboptimal.

Set $R_f = 1$ and initial risky endowments to zero. Suppose there are masses $N_I > 0$ and $N_U \geq 0$ of informed and uninformed competitive investors with common risk aversion A. Informed investors observe S; uninformed investors begin with the prior. Let z be aggregate exogenous net buying, with positive z representing a purchase demand that other investors must accommodate. Zero net risky supply gives

$$N_I X_I + N_U X_U + z = 0.$$

If both groups ignore what the price reveals beyond their initial information, their schedules are

$$X_I = \frac{\mu_I - p}{A v_I}, \qquad X_U = \frac{\bar{F} - p}{A\sigma_F^2}.$$

These are often called naive expectations in this particular comparison. The defect is not that investors are price-takers; the defect is that they omit informative price observations from their beliefs.

Multiply the clearing equation by A and collect terms in p. Define

$$D_N = \frac{N_I}{v_I} + \frac{N_U}{\sigma_F^2}.$$

Then

$$p = \omega\mu_I + (1-\omega)\bar{F} + \frac{A}{D_N}z, \qquad \omega = \frac{N_I/v_I}{N_I/v_I + N_U/\sigma_F^2}. \tag{2.9}$$

The weight on informed beliefs rises with their mass and precision. This is aggregation by willingness to bear risk: a group with greater population or lower uncertainty responds more strongly to a given difference between valuation and price. In the limiting case $N_U = 0$, informed beliefs receive weight one.

Substitute the signal-update equation into (2.9):

$$p = \bar{F} + \omega\kappa(S - \bar{F}) + \frac{A}{D_N} z.$$

The clearing price depends on S. An uninformed investor who observes p therefore generally acquires information about F. Her original prior-only demand ceases to be optimal. Clearing is necessary, but belief consistency is still missing.

If z is known, and the signal coefficient is nonzero, an observer can invert this equation and recover S. If z is unobserved and random, the observer faces an identification problem: was the high price caused by favourable information, by exogenous buying, or by both? This is the opening for a noisy rational-expectations equilibrium.

2.6 Solving a noisy rational-expectations equilibrium

The equilibrium must close a loop: the price rule determines what uninformed investors learn; that learning determines their demand; and their demand must reproduce the conjectured price rule. A linear-normal environment makes this consistency problem tractable without removing the uncertainty created by noise trading.

For this section use the payoff-decomposition normalization

$$F = S + \epsilon, \qquad S \sim \mathcal{N}(0, \sigma_S^2), \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2), \quad z \sim \mathcal{N}(0, \sigma_z^2), \tag{2.10}$$

with mutual independence, $\sigma_S^2, \sigma_\epsilon^2 > 0$, and $\sigma_z^2 \geq 0$. This is a new parametrization: here S is the component of payoff that informed investors observe, whereas in the previous section S was a noisy measurement of payoff. The equations are not interchangeable without changing their variance definitions.

Let $N > 0$ be the mass of informed investors and $M \geq 0$ the mass of uninformed investors. All are competitive and have coefficient $A > 0$. Interpret these as population masses in a price-taking model, not as a claim that a finite collection of large investors would ignore its market power.

The informed investor already knows S. In this model price can reveal information about z, but that does not help predict the independent residual ϵ. Therefore

$$\mathbb{E}[F \mid S, p] = S, \qquad \mathrm{Var}(F \mid S, p) = \sigma_\epsilon^2, \qquad X_I = \frac{S - p}{A\sigma_\epsilon^2}.$$

The informed group's aggregate demand responds to expected excess payoff with sensitivity

$$a = \frac{N}{A\sigma_\epsilon^2}.$$

This coefficient measures units of aggregate risky demand per unit change in expected excess payoff. All investors see the clearing price, but uninformed investors do not directly observe S or z. They infer value from price using the equilibrium price function.

Conjecture an uninformed demand schedule $X_U = -Hp$. The coefficient H is still unknown: it must reflect both the direct effect of a higher purchase price and the information conveyed by that price. Market clearing becomes

$$a(S-p) - MHp + z = 0,$$

so that

$$p = \frac{aS+z}{L}, \qquad L = a + MH. \tag{2.11}$$

Here L is aggregate depth against an additive order-flow shock, holding equilibrium schedules fixed.

2.6.1 Extract the signal before matching coefficients

For any positive finite L, observing p is equivalent to observing

$$\Psi = \frac{L}{a}p = S + \frac{z}{a}. \tag{2.12}$$

The transformation does not reveal S exactly unless z is known or has zero variance. The signal-to-noise problem is now transparent. Set

$$D = \mathrm{Var}(\Psi) = \sigma_S^2 + \frac{\sigma_z^2}{a^2}, \tag{2.13}$$

$$k = \frac{\mathrm{Cov}(F,\Psi)}{\mathrm{Var}(\Psi)} = \frac{\sigma_S^2}{D}, \tag{2.14}$$

$$V = \mathrm{Var}(F \mid \Psi) = \sigma_\epsilon^2 + \sigma_S^2 - \frac{\sigma_S^4}{D} = \sigma_\epsilon^2 + \sigma_S^2(1-k). \tag{2.15}$$

Gaussian conditioning gives $\mathbb{E}[F \mid p] = k(L/a)p$ and $\mathrm{Var}(F \mid p) = V$. In particular, the coefficient on p is not obtained by dividing mechanically by its signal loading and ignoring its noise loading.

Uninformed optimal demand is consequently

$$X_U = \frac{k(L/a)p - p}{AV}.$$

Consistency with $X_U = -Hp$ requires

$$H = \frac{1 - kL/a}{AV} = \frac{1 - k - kMH/a}{AV}.$$

Solving this scalar fixed point gives

$$H = \frac{1-k}{AV + Mk/a}, \qquad L = a + MH, \tag{2.16}$$

which is nonnegative. It is sometimes convenient to write $p = b\Psi$, where

$$b = \frac{a}{L} = \frac{aAV + Mk}{aAV + M}. \tag{2.17}$$

All the coefficients on the right are determined by primitives. The loop is now closed: the price produces the posterior used to choose demand, and those demands clear at that same price. Substitution into the three conditions provides a direct verification.

The parameter restrictions also verify the conjecture's domain. Since $V \geq \sigma_\epsilon^2 > 0$, $a > 0$, and $0 < k \leq 1$, the denominator in (2.16) is strictly positive. Thus $H \geq 0$ and $L \geq a > 0$: the price-to-signal transformation used in the derivation is well defined at the solution. Here H is a demand-schedule slope and L is aggregate depth, whereas b and k are dimensionless learning and price-scaling coefficients.

Why does an uninformed investor sell when price is positive? A high price is favourable news, so her posterior mean is positive. But price also includes the influence of the better-informed investors and noise demand. In this equilibrium $k \leq b \leq 1$, implying $\mathbb{E}[F \mid p] = (k/b)p \leq p$ for $p \geq 0$. A positive signal in the price does not imply a positive expected excess payoff at that price.

Example 2.3 (A complete equilibrium calculation)**.** Take $A = 1$, $N = M = 1$, $\sigma_S^2 = 4$, $\sigma_\epsilon^2 = 1$, and $\sigma_z^2 = 4$. These are illustrative quantities, not estimated market parameters. Then

$$a = 1, \quad D = 8, \quad k = \frac{1}{2}, \quad V = 3, \quad H = \frac{1}{7}, \quad L = \frac{8}{7}, \quad b = \frac{7}{8}.$$

Thus $p = \frac{7}{8}(S + z)$, and the uninformed posterior is

$$\mathbb{E}[F \mid p] = \frac{4}{7}p, \qquad \mathrm{Var}(F \mid p) = 3.$$

For the realization $S = 2$, $z = 0$, price is 1.75. Informed demand is $2 - 1.75 = 0.25$; uninformed demand is $(1 - 1.75)/3 = -0.25$. The market clears. The uninformed investor learns good news—her expected payoff rises from zero to one—but still sells because the price exceeds her posterior valuation.

2.6.2 An order cannot serve as its own new information

The competitive equilibrium is now complete. Before considering information acquisition, a separate information-accounting exercise clarifies what changes when an investor can influence the observed price. Take M to be an integer at least one and examine one uninformed investor submitting X, with the other $M-1$ stipulated schedules fixed. This is not a reinterpretation of an atomless competitive investor as a strategic large trader. Define

$$L_{-} = a + (M-1)H.$$

The resulting residual clearing equation is

$$p = \frac{aS + z + X}{L_{-}}, \qquad \frac{L_{-}p - X}{a} = S + \frac{z}{a} = \Psi.$$

An investor who knows her own order should not count its mechanical contribution to price as an independent signal about value. The displayed residualization is an accounting identity conditional on that known order. It does not, by itself, establish that Ψ and a strategically chosen X are independent random variables. Nor does it imply that conditioning on p alone is equivalent to conditioning on Ψ off equilibrium. The relevant equivalence here concerns (p, X) and (Ψ, X) under the specified clearing rule.

If this investor is large enough to choose X while internalizing $dp/dX = 1/L_{-}$, her optimization is a different, strategic problem. Its first-order condition must include price impact and its information structure must be specified. The competitive demands derived above must not be relabelled a Nash equilibrium. Kyle's demand-function analysis makes this distinction explicit [27]. There is no model-free rule that a strategic market reveals exactly, or at most, one-half of all private information.

2.7 Information acquisition and market quality

If noise variance vanishes, then $k = 1$, $V = \sigma_\epsilon^2$, $H = 0$, and $p = S$. Price reveals the informed signal, not the independent residual payoff shock. Perfect signal revelation is therefore different from perfect foresight about F.

The fixed-population equilibrium so far takes N and M as given. If observing S is privately costly, an additional participation decision is needed: does the improvement in investment opportunities justify the information cost? If everyone can infer the same signal from price without paying, paying for it offers no informational advantage in this benchmark. Yet prices cannot transmit privately gathered information if no one gathers it. The information-acquisition tension is the central insight of Grossman and Stiglitz [19]. The noisy model explains how imperfect revelation can preserve the value of information; it does not determine the endogenous informed population until that acquisition problem is added.

2.7.1 Depth is not volatility

From (2.11), price impact of an additional unit of aggregate exogenous buying is

$$\frac{\partial p}{\partial z} = \frac{1}{L}.$$

Depth L is its reciprocal. If noise demand is written as $z = Z\tilde{x}$, then $\partial p/\partial \tilde{x} = Z/L$; the units and the quantity differentiated with respect to matter. Calling both derivatives "liquidity" without specifying the order unit invites a scaling error.

Unconditional price variance is

$$\mathrm{Var}(p) = \frac{a^2\sigma_S^2 + \sigma_z^2}{L^2} = b^2 D. \tag{2.18}$$

One part comes from information; another comes from noninformational demand. A volatile price may be responding promptly to important news. A quiet price may simply reveal little. Volatility alone does not resolve this distinction.

A natural informational-precision measure is

$$\mathrm{IE} = \frac{1}{\mathrm{Var}(F \mid p)} = \frac{1}{V}.$$

It measures uncertainty remaining *after* observing price, not the variability of price itself. In the numerical example, prior payoff variance is 5, posterior payoff variance is 3, and price variance is 6.125. Price is informative even though its unconditional variance exceeds payoff variance: it also responds to noise demand. The lower uncertainty after conditioning and the high unconditional variability answer different questions.

An increase in N, holding other primitives fixed, increases a and reduces the noise variance σ_z^2/a^2 in the price signal. Hence k rises and V falls. An increase in σ_z^2 has the opposite informational effect. By contrast, changing the uninformed population M alters the scale b but not Ψ's information content in this common-signal benchmark. This last result is model-specific: it relies on competitive schedules, the shared informed signal, and the stipulated independence assumptions.

2.8 A bridge to safe claims and financial innovation

A related question concerns security design: what claims will intermediaries supply when investors care intensely about safety? To see why a different preference model matters, consider an objective of the form

$$C_0 + C_1 + \vartheta \min_{\omega\in\Omega} C_2(\omega), \qquad \vartheta > 0,$$

where Ω is the set of states investors consider possible. This is a worst-case terminal-consumption criterion, not the CARA expected-utility problem solved above.

For example, if a bond pays R and a risky claim pays $Ry(\omega)$, a nonnegative portfolio (a, b) has worst-case terminal payoff $R[b + a \min_{\omega\in\Omega} y(\omega)]$. With purchase prices p_A, p_B and wealth W, its date-zero cost is $ap_A + bp_B$. Thus the expression

$$W - ap_A - bp_B + \vartheta R\big[b + a \min_{\omega\in\Omega} y(\omega)\big]$$

captures this valuation logic when interim consumption is suppressed. This is a preference-and-budget illustration, not a complete equilibrium: feasible holdings, borrowing rules, intermediary technology, and supplies must also be specified to solve the security-design problem. The liquidity-fragility lecture returns to the importance of perceived safety for financing. Moving to the worst-case criterion changes preferences and, potentially, beliefs about possible states; it is not an alternative solution of the normal–CARA equilibrium.

2.9 Exercises

Exercise 2.1 (The sign of compensation)**.** An investor has $A = 0.025$, $R_f = 1.02$, conditional payoff mean 105, payoff variance 64, and an initial endowment of one share. Find her demand at $p = 100$. What price would make a final holding of three shares optimal? Explain the direction of the price adjustment.

Exercise 2.2 (Precision and clearing)**.** Let $F \sim \mathcal{N}(50, 9)$ and $S = F + \eta$ with independent noise variance 16. For $S = 60$, compute posterior mean and variance. In the naive-clearing benchmark, set $N_I = 2$, $N_U = 3$, $A = 1$, and $z = 0$. Calculate ω and price. Is this price a rational-expectations equilibrium?

Exercise 2.3 (Noise and information)**.** Keep all parameters in Example 2.3 except increase σ_z^2 from 4 to 12. Calculate D, k, V, H, L, b. Compare informational precision and price variance with the original example.

Exercise 2.4 (Two uses of the word rational)**.** A trader correctly extracts $S + z/a$ from an equilibrium price but ignores the effect of her own large order on that price. Which requirement has she met, and which strategic feature has she omitted? Explain why subtracting her own order does not prove independence of her order and the residual signal.

2.10 Solutions and interpretation

Exercise 2.1. At $p = 100$, expected excess payoff is $105 - 1.02(100) = 3$ and $Av = 1.6$. Desired final holdings are $3/1.6 = 1.875$, so net demand is 0.875. To induce a final holding of three, price must be $(105 - 0.025 \cdot 3 \cdot 64)/1.02 = 98.2353$. Taking more risk requires a lower entry price at the same payoff distribution.

Exercise 2.2. The signal weight is $9/25 = 0.36$, giving posterior mean 53.6 and variance $144/25 = 5.76$. The informed weight in market clearing is

$$\omega = \frac{2/5.76}{2/5.76 + 3/9} = \frac{25}{49}.$$

Price is $50 + (25/49)(3.6) = 51.8367$. Since noise demand is known here, the price reveals the signal under the stipulated rule. The uninformed prior-only demand therefore fails belief consistency; the calculation clears naive schedules but is not the claimed rational-expectations equilibrium.

Exercise 2.3. Now $D = 16$, $k = 1/4$, and $V = 4$. Hence $H = 3/17$, $L = 20/17$, and $b = 17/20$. Informational precision falls from $1/3$ to $1/4$, while price variance rises from 6.125 to $(17/20)^2(16) = 11.56$. More noise makes price less informative and more variable in this comparison. This example does not establish that the two measures always move oppositely under every parameter change.

Exercise 2.4. She has used price information consistently with the stated price rule, but has not solved an optimization problem that internalizes her market power. Residualization removes a known mechanical contribution to price; it says nothing by itself about statistical independence. A chosen order can depend on the same underlying information that appears in the residual. The competitive and strategic problems require different first-order conditions.

Reading onward

For the symmetric-information-to-REE progression, see Chapter 2 of [9]. The next lecture turns from equilibrium demands to observed transaction prices. It asks how bid–ask trading and information updates leave different signatures in price changes, and what those signatures can actually identify.

Lecture 3

Transaction Prices: Spreads, Innovations, and Price Discovery

A sequence of transaction prices can move up and down even when no one has learned anything new about the asset. A buyer pays the ask; the next seller receives the bid. The resulting reversal is a feature of trading, not necessarily a reversal in fundamental value. At the same time, a trade can genuinely reveal information and permanently change valuation. Empirical microstructure begins by distinguishing these mechanisms.

This lecture develops the distinction through the Roll model and a simple information-sensitive extension. We derive their covariance restrictions rather than treating negative serial dependence as a sufficient diagnosis. We then build an innovations representation, use it for linear forecasting and filtering, and ask which structural quantities prices alone identify. The main lesson is that fitting a time-series model and recovering an economic mechanism are different accomplishments.

3.1 Three objects and an observation clock

Time t indexes trades. This is *event time*: the elapsed calendar time between trade $t-1$ and trade t need not be constant. Write p_t for the observed transaction price and m_t for the efficient or fundamental-value component. In these models m_t is a conditional valuation, not necessarily the asset's eventual cash payoff. A market can continue to learn about that payoff after trade t.

Define trade direction by

$$q_t = \begin{cases} +1, & \text{buyer-initiated trade,} \\ -1, & \text{seller-initiated trade.} \end{cases}$$

A buyer-initiated trade executes against a liquidity supplier's ask; a seller-initiated trade executes against its bid. The sign belongs to the initiating trader. It is not the sign of the price change: a buyer can execute below the previous transaction price if the underlying quotes have fallen enough.

A martingale satisfies $\mathbb{E}[m_t \mid \mathcal{F}_{t-1}] = m_{t-1}$ for a specified information set $\mathcal{F}_{t-1}$. Its next change is not forecastable from that information. A random walk with independent, zero-mean increments is a stronger specification. In particular, a martingale need not have constant conditional variance, normal increments, or independent increments.

Transaction prices need not be martingales even when the efficient component is one. Short-run predictability in p_t can reflect the spread. Therefore a rejection of a transaction-price random walk is not, by itself, evidence that available fundamental information has been left unexploited.

For the stationary price-change series $\Delta p_t = p_t - p_{t-1}$, define

$$\gamma_k = \mathrm{Cov}(\Delta p_t, \Delta p_{t-k}), \qquad \rho_k = \frac{\gamma_k}{\gamma_0}, \quad \gamma_0 > 0.$$

The first quantity is an autocovariance and has squared-price units. The second is an autocorrelation and is dimensionless. The distinction is necessary when a covariance is used to recover a monetary spread.

3.2 The basic Roll model

Start with a market in which order direction carries no information about value. This deliberately separates the cost of crossing the spread from genuine news. If the spread alone creates predictable reversals, those reversals should leave a recognisable signature in transaction-price changes.

Assume

$$m_t = m_{t-1} + u_t, \tag{3.1}$$

$$p_t = m_t + cq_t, \qquad c \geq 0. \tag{3.2}$$

The increments u_t are independent across time with mean zero and variance $\sigma_u^2 < \infty$. The signs q_t are independent across time, take $+1$ and -1 with probability one-half, and are independent of the entire public-news sequence $\{u_t\}$. The half-spread c is constant over the estimation window.

Thus $\mathbb{E}[q_t] = 0$ and $\mathrm{Var}(q_t) = 1$, with a two-point rather than normal distribution. Two distinct independence assumptions do the economic work: independence across trade signs removes order-flow persistence, while independence from news removes informed trading. Persistent order splitting violates the first; buying correlated with news violates the second.

Under (3.2), bid and ask are $m_t - c$ and $m_t + c$, giving full spread $s = 2c$. A liquidity demander pays the half-spread relative to m_t on either side of the market: buyers pay more and sellers receive less. What they pay for immediacy is distinct from a dealer's net profit, which also depends on inventory changes, information losses, operating costs, and the timing of offsetting trades.

3.2.1 Bid–ask bounce without news

To isolate the mechanism, temporarily set $u_t = 0$, so the efficient price is constant. Each transaction occurs at one of two prices. A transition from bid to ask changes price by s; a transition from ask to bid changes it by $-s$; consecutive transactions on the same side generate zero change.

With independent symmetric signs,

$$\Pr(\Delta p_t = s) = \frac{1}{4}, \quad \Pr(\Delta p_t = 0) = \frac{1}{2}, \quad \Pr(\Delta p_t = -s) = \frac{1}{4}.$$

Two consecutive price increases of s are impossible in this constant-value, constant-spread example. Once the first increase places the transaction at the ask, the next transaction can remain there or move to the bid; it cannot move one more spread above the ask. This does *not* mean a second buyer-initiated trade is impossible. Another buy simply leaves the transaction price unchanged.

Enumerating the eight equally likely triples (q_{t-1}, q_t, q_{t+1}) gives the joint distribution:

	Δp_t		
Δp_{t+1}	$-s$	0	$+s$
$-s$	0	1/8	1/8
0	1/8	1/4	1/8
$+s$	1/8	1/8	0

The only cells contributing a nonzero product of consecutive changes have opposite signs. Their combined probability is $1/4$, so the covariance is $-s^2/4 = -c^2$. The negative covariance is generated by the shared middle trade, even though trade signs themselves are independent.

3.2.2 Deriving the covariance restrictions

Return to stochastic public news. Subtract consecutive transaction prices:

$$\Delta p_t = u_t + c(q_t - q_{t-1}). \tag{3.3}$$

All three terms have zero mean. Independence therefore gives

$$\begin{aligned}
\gamma_0 &= \mathrm{Var}\big(u_t + cq_t - cq_{t-1}\big) && (3.4)\\
&= \sigma_u^2 + c^2\,\mathrm{Var}(q_t) + c^2\,\mathrm{Var}(q_{t-1}) && (3.5)\\
&= \sigma_u^2 + 2c^2. && (3.6)
\end{aligned}$$

For the first autocovariance, write the two increments beside each other:

$$\Delta p_t = u_t + cq_t - cq_{t-1}, \qquad \Delta p_{t-1} = u_{t-1} + cq_{t-1} - cq_{t-2}.$$

The only common independent shock is q_{t-1}. Its coefficients are $-c$ and c. Consequently

$$\gamma_1 = -c^2, \qquad \gamma_k = 0 \quad \text{for } k \geq 2. \tag{3.7}$$

The shared-shock calculation also explains the one-lag limit: once the two changes are separated by at least two trades, they no longer contain the same sign. The negative first covariance and zero higher covariances therefore come from the same timing assumption.

We obtain the Roll spread relation [31]:

$$s = 2\sqrt{-\gamma_1}, \qquad \sigma_u^2 = \gamma_0 + 2\gamma_1. \tag{3.8}$$

The minus sign belongs *inside* the square root. In the population model, $\gamma_1 \leq 0$ and $\gamma_0 + 2\gamma_1 \geq 0$. Its first autocorrelation lies between $-1/2$ and zero. Public news raises contemporaneous price-change variance without changing the negative covariance generated by the spread, making the normalized autocorrelation less negative.

Example 3.1 (Recovering a spread from moments)**.** Suppose population moments of the trade-time price changes are $\gamma_0 = 0.09$ and $\gamma_1 = -0.01$, measured in squared currency units. Then $c = 0.10$, the full spread is 0.20, and efficient-price innovation variance is $0.09 - 0.02 = 0.07$. The first autocorrelation is $-1/9$. These are a constructed population example; an empirical estimate would have sampling uncertainty and would require checking the model's restrictions.

If an empirical first covariance is positive, taking $\sqrt{|\widehat{\gamma}_1|}$ is not the Roll estimator justified by this model. A researcher may use a separately motivated constrained estimator, but must distinguish it from evidence that the population restriction holds. Finite samples, changing spreads, dependence in signs, information effects, or data errors can all affect the observed moment.

3.3 Adding information revealed by trades

The basic model assumes order direction contains no news about value. To allow permanent information effects, consider

$$m_t = m_{t-1} + u_t + \lambda q_t, \qquad \lambda \geq 0, \tag{3.9}$$

$$p_t = m_t + cq_t. \tag{3.10}$$

Retain independent symmetric signs and independent public shocks as stipulated structural assumptions. The increment $w_t = u_t + \lambda q_t$ combines public news with the valuation update caused by the order. It has mean zero and variance

$$\sigma_w^2 = \sigma_u^2 + \lambda^2.$$

The clock remains discrete trade time: w_t is a one-trade innovation, not a Brownian differential. A continuous-time model would require a separate convention for elapsed-time variance; in particular, Brownian increments over arbitrary intervals are not automatically unit-variance shocks.

3.3.1 The timing that determines the spread

A precise timeline prevents confusion about c and λ:

1. Immediately before date-t public news, the market's efficient valuation is m_{t-1}.
2. The public innovation u_t arrives. The pre-trade public valuation becomes $\mu_t = m_{t-1} + u_t$.
3. A buy or sell order arrives. Its direction changes efficient valuation to $m_t = \mu_t + \lambda q_t$.
4. The transaction occurs at $p_t = \mu_t + (\lambda + c)q_t$.

This is the timing clarified in Hasbrouck's author-hosted corrigenda [22]. In this model the two execution prices conditional on trade side are

$$p_t^{\text{ask}} = \mu_t + (\lambda + c), \qquad p_t^{\text{bid}} = \mu_t - (\lambda + c).$$

The spread relative to the same pre-trade public information is $2(\lambda + c)$. The permanent information component is λ per side; the residual trading-cost component is c per side. Yet $p_t - m_t = cq_t$ remains true because m_t is the *post-trade* valuation. A spread measured around pre-trade public value and a transaction's deviation from post-trade value are not the same object.

The extension is a reduced-form structural model. It stipulates a trade's information effect rather than deriving the probability of an informed trader from an underlying private-information economy. A separate adverse-selection model can provide such a foundation. We should not treat the parameter's label as a proof of its economic origin in every dataset.

3.3.2 Moments and observational equivalence

First differences satisfy

$$\Delta p_t = (c + \lambda)q_t - cq_{t-1} + u_t. \tag{3.11}$$

The covariance calculations now yield

$$\gamma_0 = (c+\lambda)^2 + c^2 + \sigma_u^2, \tag{3.12}$$

$$\gamma_1 = -c(c+\lambda), \tag{3.13}$$

$$\gamma_k = 0, \qquad k \geq 2. \tag{3.14}$$

In the first covariance, the shared sign q_{t-1} has coefficients $-c$ and $c+\lambda$. Higher lags share no shocks under the maintained independence assumptions.

Adding the first two moments in the appropriate combination gives

$$\gamma_0 + 2\gamma_1 = \lambda^2 + \sigma_u^2 = \sigma_w^2. \tag{3.15}$$

Thus price-change moments identify total permanent innovation variance in this model. They do not separately identify λ, c, and σ_u^2: two moment equations cannot generally recover three unrestricted unknowns.

Moreover, the naive Roll inversion becomes

$$2\sqrt{-\gamma_1} = 2\sqrt{c(c+\lambda)}.$$

For $c, \lambda > 0$, this lies above $2c$ but below the full information-sensitive spread $2(c+\lambda)$. A successful computation does not guarantee that the statistic has the same interpretation after the model changes.

3.4 Why long-horizon variation looks different

The preceding moments concern adjacent trades. To connect them with risk measured over a longer horizon, ask which price movements accumulate and which cancel when many transaction-price changes are added together. Permanent news accumulates; crossing from one side of a spread to the other mainly affects the endpoints.

Since m_t has independent increments,

$$\operatorname{Var}(m_t - m_{t-k}) = k\sigma_w^2.$$

For transaction prices, sum (3.11) over k trades, or equivalently use the covariance restrictions. Either route gives

$$\operatorname{Var}(p_t - p_{t-k}) = k\gamma_0 + 2(k-1)\gamma_1 \tag{3.16}$$

$$= k\sigma_w^2 + 2c(c+\lambda), \qquad k \geq 1. \tag{3.17}$$

The permanent component grows with the number of trades; the endpoint microstructure contribution is constant. Hence

$$\lim_{k\to\infty} \frac{\mathrm{Var}(p_t - p_{t-k})}{k} = \sigma_w^2.$$

The normalization by k is important. The unscaled variance difference does not vanish; it remains $2c(c + \lambda)$ in this model.

At a one-trade horizon, spread-related variation can dominate. Over many trades, its proportion of total variance diminishes if $\sigma_w^2 > 0$. This explains how noisy transaction prices at fine sampling intervals can coexist with more stable long-horizon measures of efficient-price variation. It does not guarantee that every real market obeys a constant-spread independent-increment model over long calendar periods.

The increments, not the integrated price levels, are covariance stationary here: their means are time-invariant and their covariances depend only on the lag. To estimate these population moments from one observed time series, we also need a law of large numbers for the relevant time averages. If the independent public shocks are also identically distributed over time, the finite-memory construction from those shocks and iid signs gives a stationary, ergodic increment process. Independence and constant second moments alone should not be mistaken for a statement that the entire distribution is time-invariant. The covariance derivations above do not require that stronger distributional assertion.

3.5 An innovations representation

A price-only observer cannot see the separate news and order-sign shocks in the structural equations. That observer can nevertheless ask which part of each new price change was unpredictable from previous prices. An innovations representation answers this *linear-prediction* question. It describes the same covariance structure as the structural representation, but its shocks have a statistical meaning rather than separately identifying public news and trading costs.

Write the scalar MA(1) representation as

$$\Delta p_t = \varepsilon_t + \theta\varepsilon_{t-1}, \qquad \mathrm{Var}(\varepsilon_t) = \sigma_\varepsilon^2. \tag{3.18}$$

Here ε_t is zero-mean white noise: innovations at different dates are uncorrelated. Under a general non-Gaussian structural model, white noise need not be independent or Gaussian. Our binary trade signs make this qualification substantive.

Moment matching gives

$$\gamma_0 = (1 + \theta^2)\sigma_\varepsilon^2, \qquad \gamma_1 = \theta\sigma_\varepsilon^2.$$

Assume $\gamma_0 > 2|\gamma_1|$, so the spectral density is strictly positive and the invertible representation exists. Eliminating θ yields a quadratic in innovation variance:

$$(\sigma_\varepsilon^2)^2 - \gamma_0\sigma_\varepsilon^2 + \gamma_1^2 = 0.$$

Choose the larger root:

$$\sigma_\varepsilon^2 = \frac{\gamma_0 + \sqrt{\gamma_0^2 - 4\gamma_1^2}}{2}, \qquad \theta = \frac{\gamma_1}{\sigma_\varepsilon^2}. \tag{3.19}$$

This selects $|\theta| < 1$; in the present model $-1 < \theta \le 0$. The smaller variance root corresponds to the noninvertible alternative when $\gamma_1 \neq 0$.

Why discuss invertibility rather than just fit either root? Under invertibility, current and past innovations can be recovered by a stable linear filter of current and past price changes. The resulting shocks are the one-step *linear* prediction errors relative to that history. Without this restriction, the same autocovariances can correspond to a moving-average representation whose shocks are not recoverable from observations in that way.

Wold's theorem supplies a moving-average representation for the nondeterministic part of a covariance-stationary series. The finite MA(1) here additionally follows from factorizing this specific degree-one covariance spectrum. It does not follow that every stationary process has independent shocks, nor that zero linear autocovariances eliminate every form of nonlinear dependence. The deterministic component and regularity conditions should not be discarded by assertion.

In the boundary case of pure bid–ask bounce with no permanent news, $\lambda = \sigma_u = 0$, we have $\gamma_0 = 2c^2$, $\gamma_1 = -c^2$, and $\theta = -1$. The stable inverse discussed next is not available at this unit-circle boundary. Treating $|\theta| < 1$ as automatic would conceal precisely the case used in the introductory bounce illustration.

3.5.1 The lag operator and stable inversion

Let $\mathsf{L}x_t = x_{t-1}$. Equation (3.18) is

$$\Delta p_t = (1 + \theta\mathsf{L})\varepsilon_t.$$

For $|\theta| < 1$, the geometric-series inverse gives

$$\begin{aligned} \varepsilon_t &= (1 + \theta\mathsf{L})^{-1}\Delta p_t & (3.20)\\ &= \Delta p_t - \theta\Delta p_{t-1} + \theta^2\Delta p_{t-2} - \theta^3\Delta p_{t-3} + \cdots . & (3.21) \end{aligned}$$

Equivalently,

$$\Delta p_t = \theta \Delta p_{t-1} - \theta^2 \Delta p_{t-2} + \theta^3 \Delta p_{t-3} - \cdots + \varepsilon_t.$$

The alternating algebraic coefficients come from the inverse of $1 + \theta \mathsf{L}$. Replacing this expression by an unqualified sum of positive powers loses the sign convention.

Finite datasets require initialization. A practical recursion starts from an initial innovation estimate and allows a warm-up period; the influence of its error decays geometrically when $|\theta| < 1$. Results very near the boundary can be sensitive to initialization, estimation error, and the length of the observed history.

3.6 Forecasting and filtering are not the same task

With recoverable innovations in hand, we can use a new transaction to answer two questions. Where is the next transaction price likely to be? And what does the current transaction tell us about the efficient price already prevailing? The first is forecasting, the second filtering. Their answers coincide in a particular sense below, but their targets are different.

Let Proj_t denote the best linear projection onto the closed linear span of information available from the price history through t, with constants included. Under the stationary infinite-past innovations representation,

$$\mathrm{Proj}_t(\Delta p_{t+1}) = \theta \varepsilon_t, \qquad \mathrm{Proj}_t(\Delta p_{t+h}) = 0 \quad (h \geq 2).$$

The one-step price forecast is $p_t + \theta \varepsilon_t$. These are statements about linear forecasts. If the entire relevant system were jointly Gaussian, they would also be conditional expectations. The binary-sign model generally does not satisfy that condition, so an exact Bayesian forecast can be nonlinear.

These are steady-state formulas using an infinite history of price changes, with the current observed price anchoring the level. A finite observation history and an informative prior for the initial efficient price generally introduce filtering transients. An initialization rule and its uncertainty must then accompany the recursive calculation.

Filtering asks a different question: what is the best estimate of the *current latent* efficient price m_t? From $m_t = p_t - c q_t$,

$$\widehat{m}_t^{\mathrm{lin}} = p_t - c\,\mathrm{Proj}_t(q_t).$$

Because current q_t is orthogonal to past price innovations, only its covariance with the current innovation enters the projection. From (3.11) and $\varepsilon_t = \Delta p_t - \theta \varepsilon_{t-1}$,

$$\mathrm{Cov}(q_t, \varepsilon_t) = c + \lambda.$$

The projection is therefore

$$\mathrm{Proj}_t(q_t) = \frac{c+\lambda}{\sigma_\varepsilon^2}\varepsilon_t,$$

and the efficient-price filter becomes

$$\widehat{m}_t^{\mathrm{lin}} = p_t - \frac{c(c+\lambda)}{\sigma_\varepsilon^2}\varepsilon_t = p_t + \theta\varepsilon_t. \tag{3.22}$$

The equality of the filtered efficient price and the next-transaction linear forecast is meaningful. Future independent trade signs and future public shocks have zero projection on today's history, so the next transaction price is forecast at today's filtered valuation.

The formula does not say that the full conditional expectation of a binary sign is linear. Indeed, a linear projection of q_t can fall outside $[-1, 1]$, while its exact conditional mean cannot. That is a useful diagnostic of the difference between a linear approximation and a probability calculation. If trade signs themselves are observed and c is known, the structural formula $m_t = p_t - cq_t$ recovers m_t directly; the filtering problem above is for an observer using prices alone.

3.6.1 Pricing error versus filtering error

The *pricing error* is the transaction's deviation from the structural efficient price:

$$s_t = p_t - m_t = cq_t, \qquad \mathrm{Var}(s_t) = c^2.$$

The *linear filtering error* is the estimation error $m_t - \widehat{m}_t^{\mathrm{lin}}$. Its variance is

$$\mathrm{Var}(m_t - \widehat{m}_t^{\mathrm{lin}}) = c^2\left(1 - \frac{(c+\lambda)^2}{\sigma_\varepsilon^2}\right) \tag{3.23}$$

$$= c^2 - \frac{\gamma_1^2}{\sigma_\varepsilon^2}. \tag{3.24}$$

The exact conditional-mean estimator, if available, has no greater mean-squared error than the best linear estimator. Its uncertainty should not be inferred by pretending the binary signs are normal.

A third quantity is the adjustment from transaction price to the linear filter:

$$p_t - \widehat{m}_t^{\mathrm{lin}} = -\theta\varepsilon_t, \qquad \mathrm{Var}(p_t - \widehat{m}_t^{\mathrm{lin}}) = \theta^2\sigma_\varepsilon^2.$$

Projection orthogonality decomposes the structural pricing-error variance into this explained part plus the filtering-error variance. Confusing these three quantities can produce an apparently sophisticated but incorrectly labelled estimate of market efficiency.

3.7 What do price-only second moments identify?

The innovations parameters can be recovered from two observed autocovariances. The structural model still contains three unknowns: residual trading cost, trade-related information impact, and public-news variance. Rather than select one decomposition arbitrarily, we now find the full range of trading costs consistent with those moments and the maintained sign restrictions. This is identification from second moments, not an assertion that the full non-Gaussian price distribution contains no additional information. Equal autocovariances need not imply equal higher-order distributions.

Suppose the population moments obey $\gamma_1 < 0$ and $\gamma_0 > 2|\gamma_1|$. Define $g = -\gamma_1 > 0$, $v = \gamma_0$, and $x = c^2$. The structural moment restrictions imply

$$c + \lambda = \frac{g}{c}, \qquad \sigma_u^2 = v - x - \frac{g^2}{x}.$$

Nonnegative public-news variance requires

$$x^2 - vx + g^2 \leq 0.$$

Let

$$x_- = \frac{v - \sqrt{v^2 - 4g^2}}{2}, \qquad x_+ = \frac{v + \sqrt{v^2 - 4g^2}}{2}.$$

Then $x \in [x_-, x_+]$. The maintained adverse-selection sign $\lambda \geq 0$ adds $g/c - c \geq 0$, or $x \leq g$. Since $x_- \leq g \leq x_+$, the identified set is

$$\boxed{c^2 \in [x_-, g]} \qquad \text{within this restricted generalized Roll model.} \tag{3.25}$$

The interval is sharp for these moment restrictions. For any $x \in [x_-, g]$, choose $c = \sqrt{x}$, $\lambda = g/\sqrt{x} - \sqrt{x}$, and $\sigma_u^2 = v - x - g^2/x$. These are admissible parameters and reproduce both observed moments, so no point inside the interval can be excluded using those restrictions alone.

The lower bound equals

$$x_- = \frac{\gamma_1^2}{\sigma_\varepsilon^2} = \theta^2 \sigma_\varepsilon^2.$$

Thus price-only second moments identify a lower bound on structural pricing-error variance, while the precise value remains undetermined by those moments. Equation (3.24) becomes $\operatorname{Var}(m_t - \widehat{m}_t^{\text{lin}}) = c^2 - x_-$. At the lower boundary the linear filtering error is zero; at the upper boundary $\lambda = 0$.

Both bounds draw their content from the maintained model. The finite upper bound g uses $c \geq 0$, $\lambda \geq 0$, independent signs, and nonnegative public-news variance. Other random-walk decompositions can admit different bounds, including no finite upper bound, because they allow different latent structures. If c is directly known instead, the structural pricing-error variance is simply c^2 and no price-only identification exercise is needed. The filtering-error variance remains a separate object.

Example 3.2 (The same price moments can hide different structures)**.** Take $c = 1$, $\lambda = 1$, and $\sigma_u^2 = 3$. Then

$$\gamma_0 = 8, \qquad \gamma_1 = -2, \qquad \sigma_w^2 = 4.$$

The pre-trade spread is 4, while the basic Roll statistic is $2\sqrt{2} \approx 2.8284$. The invertible innovations parameters are

$$\sigma_\varepsilon^2 = 4 + 2\sqrt{3} \approx 7.4641, \qquad \theta = \sqrt{3} - 2 \approx -0.2679.$$

If the latest innovation is 0.5, the next-transaction linear forecast and current efficient-price filter lie 0.13397 below the current transaction price.

The identified set for pricing-error variance is

$$c^2 \in [4 - 2\sqrt{3},\ 2] \approx [0.5359, 2].$$

At the stated structural parameters $c^2 = 1$, so linear filtering-error variance is $1-(4-2\sqrt{3}) = 2\sqrt{3} - 3 \approx 0.4641$. But the alternative structure $c = \sqrt{2}$, $\lambda = 0$, and $\sigma_u^2 = 4$ has exactly the same γ_0 and γ_1. The same price-only second moments therefore support different trading-cost and information decompositions.

Their full distributions need not coincide. If the public shocks in both examples are Gaussian, expanding the fourth power gives

$$\mathbb{E}[(\Delta p_t)^4] = 3\gamma_0^2 - 2\big[(c + \lambda)^4 + c^4\big].$$

The two structures then have fourth moments 158 and 176, respectively. Additional distributional restrictions and higher moments can therefore supply information that the covariance-based identified set deliberately does not use.

Observed trade signs can supply additional identifying moments. For example, under the maintained model,

$$\mathrm{Cov}(\Delta p_t, q_t) = c + \lambda, \qquad \mathrm{Cov}(\Delta p_t, q_{t-1}) = -c.$$

These moments distinguish c and λ when the sign assumptions hold. In real data, signing error and sign persistence can invalidate this direct inversion. Additional variables help only when their timing, measurement, and model implications are also credible.

3.8 General ARMA representations and the long-run multiplier

Real transaction-price changes can exhibit dependence beyond the single lag allowed here. A richer time-series model can accommodate that dependence while preserving the distinction between short-run variation and the cumulative effect of an innovation. Consider a stable, invertible ARMA model for zero-mean price changes:

$$\phi(\mathsf{L})\Delta p_t = \Theta(\mathsf{L})\varepsilon_t,$$

where

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_r z^r, \qquad \Theta(z) = 1 + \theta_1 z + \cdots + \theta_s z^s.$$

After cancelling any common factors, assume $\phi(z) \neq 0$ and $\Theta(z) \neq 0$ for $|z| \leq 1$. The first condition gives a causal, stable autoregressive part; the second gives invertibility. The resulting stable moving-average form is

$$\Delta p_t = \psi(\mathsf{L})\varepsilon_t, \qquad \psi(z) = \frac{\Theta(z)}{\phi(z)}.$$

The cumulative response of the price level to a one-unit innovation is the long-run multiplier

$$\psi(1) = \frac{\Theta(1)}{\phi(1)}.$$

There is no general identity $\Theta(1) = \phi(1)^{-1}$: the two polynomials describe separate parts of the fitted process. Under the usual absolute-summability conditions, long-horizon variance per observation is

$$\lim_{k\to\infty} \frac{1}{k} \operatorname{Var}(p_t - p_{t-k}) = \left[\frac{\Theta(1)}{\phi(1)}\right]^2 \sigma_\varepsilon^2. \tag{3.26}$$

For an ARMA(1,1), the multiplier is $(1+\theta)/(1-\phi)$, and its square has denominator $(1-\phi)^2 = 1 - 2\phi + \phi^2$. The squared term carries a plus sign.

For the MA(1) derived above, (3.26) reduces to

$$(1+\theta)^2\sigma_\varepsilon^2 = \gamma_0 + 2\gamma_1 = \sigma_w^2.$$

This equality is a useful cross-check between statistical and structural representations. More generally, an ARMA long-run variance is a feature of the observed process. Calling it a fundamental-news variance requires an economic model that justifies the random-walk interpretation.

3.9 Using the model empirically

An empirical implementation should begin by fixing what one observation means and which price is observed. Mixing event-time and calendar-time observations changes the interpretation of lagged moments. Using a midquote instead of a transaction price changes the measurement equation. Treating a stale price as a new trade can create a different dependence pattern again.

A useful sequence is to examine the sign and scale of the first autocovariance, inspect higher-lag covariances, assess whether a roughly stable local window is plausible, and then compare any fitted parameters with actual quote or trade information. None of these checks proves the model. Together they can reveal whether a simple spread interpretation is plausible or contradicted by obvious features of the data.

Persistence in trade signs is especially important. When signs are serially correlated, the clean overlap argument leading to $\gamma_k = 0$ for $k \geq 2$ no longer applies. A string of buys need not mean a sequence of independent information shocks; it may be a single large order split into pieces. Inventory adjustment can also have its own dynamics rather than appearing as the constant c used here. Those mechanisms motivate richer specifications, not the automatic reuse of the simple inversion formulas.

The distinction between estimation and identification should survive into the interpretation of results. An accurately estimated MA(1) parameter is evidence about linear price predictability. It is not, on its own, an estimate of the fraction of informed traders, dealer profits, or the true cost of executing a large order. A model becomes useful when its claims are matched to the data it actually needs.

3.10 Exercises

Exercise 3.1 (A diagnostic covariance)**.** In the basic Roll model, let $c = 0.05$ and $\sigma_u^2 = 0.02$. Calculate $\gamma_0, \gamma_1, \rho_1$, and the full spread. Explain why the efficient price can be a martingale while the transaction-price change has negative first autocorrelation.

Exercise 3.2 (Information and the spread)**.** For the parameters in Example 3.2, compute $\mathrm{Var}(p_t - p_{t-k})$ at $k = 1, 10, 100$. Compare each value divided by k with permanent innovation variance. Why does the spread-related variance not accumulate linearly with the horizon?

Exercise 3.3 (An innovations calculation)**.** An MA(1) fit produces $\theta = -0.4$ and $\sigma_\varepsilon^2 = 0.25$. Find its first two autocovariances and its long-run variance. If $p_t = 50$ and $\varepsilon_t = 0.10$, give the next-price linear forecast. Must this be the exact conditional expected price in a model with binary trade signs?

Exercise 3.4 (Recovering an identified set)**.** Suppose $\gamma_0 = 5$ and $\gamma_1 = -2$. Under the restricted generalized Roll model, find the admissible interval for c^2. Construct two different parameter triples (c, λ, σ_u^2) with those same moments. What do the data identify about permanent innovation variance?

Exercise 3.5 (A persistent-sign challenge)**.** Keep the basic pricing equation $p_t = m_t + cq_t$. Now allow serial dependence in trade signs. Assume their stationary, zero-mean process has $\mathrm{Cov}(q_t, q_{t-k}) = \rho^k$ for $k \geq 0$, with $|\rho| < 1$. Public news remains independent of signs. Derive γ_1 and γ_k for $k \geq 1$. Does the usual Roll square-root formula still recover $2c$?

3.11 Solutions and interpretation

Exercise 3.1. The moments are $\gamma_0 = 0.02+2(0.05)^2 = 0.025$ and $\gamma_1 = -(0.05)^2 = -0.0025$. Thus $\rho_1 = -0.1$ and spread is 0.10. The reversal comes from the shared transaction-side component, not from predictability in independent fundamental news.

Exercise 3.2. Here $\sigma_w^2 = 4$ and $2c(c + \lambda) = 4$, so horizon variances are 8, 44, and 404. Dividing by the number of trades gives 8, 4.4, and 4.04, tending to 4. Intermediate spread terms cancel when transaction-price changes are summed; what remains is an endpoint effect plus its covariance with endpoint information.

Exercise 3.3. The moments are $\gamma_0 = (1 + 0.16)(0.25) = 0.29$ and $\gamma_1 = -0.4(0.25) = -0.10$. Long-run variance is $(0.6)^2(0.25) = 0.09$, also $0.29 - 0.20$. The next-price linear forecast is $50 - 0.4(0.10) = 49.96$. Exact conditional expectation need not coincide with this projection in a non-Gaussian system.

Exercise 3.4. The lower root is $(5-\sqrt{25-16})/2 = 1$, and the upper admissible value under $\lambda \geq 0$ is $g = 2$. Thus $c^2 \in [1, 2]$. The boundary triples are $(1, 1, 0)$ and $(\sqrt{2}, 0, 1)$. Both give $\gamma_0 = 5$ and $\gamma_1 = -2$. Permanent innovation variance is identified as $5 - 4 = 1$, though its split between λ^2 and σ_u^2 differs.

Exercise 3.5. For $k \geq 1$, independent public news contributes no serial covariance, and

$$\begin{aligned}\gamma_k &= c^2\,\mathrm{Cov}(q_t - q_{t-1}, q_{t-k} - q_{t-k-1}) \\ &= c^2\big(2\rho^k - \rho^{k+1} - \rho^{k-1}\big) \\ &= -c^2(1-\rho)^2\rho^{k-1}.\end{aligned}$$

In particular $\gamma_1 = -c^2(1-\rho)^2$. The usual inversion returns $2c(1-\rho)$, not $2c$, since $\rho < 1$. Positive sign persistence therefore reduces this particular implied-spread statistic relative to the true spread, and produces nonzero higher-lag covariances. The exercise illustrates why one assumption about order flow can change both estimation and interpretation.

Lecture 4

Strategic Trading and the Release of Private Information

An investor who knows more than the market faces an unusual constraint: using information destroys part of its value. A large purchase earns money when the asset is underpriced, but also tells other traders that the asset may be worth more. The question is therefore not simply whether to buy. It is how much to buy, how quickly, and how to preserve profitable opportunities for later.

Kyle's model makes this tension explicit. An informed investor chooses a quantity while anticipating its effect on the execution price. Competitive market makers infer value from total order flow, without seeing which orders are informed. Liquidity demand gives the investor room to trade without being identified perfectly. Prices remain rational conditional on public information even though the insider earns positive expected profits.[26]

This lecture develops that equilibrium in stages. We first solve one auction, then allow imperfect information and competition among insiders. We next use dynamic programming to understand how an insider allocates information across auctions. The final sections connect the mechanism to disagreement, disclosure, funding constraints, and more general sequential games.

4.1 Information as a trading cost

Trading costs can reflect operating costs, adverse selection, and the cost of bearing inventory. These mechanisms should not be conflated. The basic Kyle market has no processing cost, no risk-aversion premium for its market makers, and no explicit bid–ask quotation mechanism. Its upward-sloping pricing rule isolates the informational component of trading costs. A positive price-impact coefficient is not literally a quoted spread; it describes how execution prices respond to order imbalance.

To see the distinction from inventory risk, suppose a dealer evaluates a normally distributed terminal asset payoff with mean m and variance s^2. With constant absolute risk aversion $\gamma > 0$, the certainty equivalent of holding q shares, apart from cash, is

$$qm - \frac{\gamma}{2}q^2 s^2.$$

The marginal valuation is $m - \gamma q s^2$. An already-long dealer values another share less because it increases exposure. This particular inventory-risk term vanishes when $\gamma = 0$. It does not follow that risk-neutral dealers can never incur inventory-related costs: financing charges, position limits, or liquidation costs can matter under other assumptions. Nor does risk neutrality imply that an uninformed trader cannot trade. At a price equal to its conditional expectation, an unconstrained risk-neutral market maker is willing to supply the quantity required at zero conditional expected profit.

4.2 One auction: strategy and inference must agree

4.2.1 Primitives, signs, and the order of events

Let the terminal payoff per share be

$$V \sim \mathcal{N}(p_0, \sigma_V^2), \qquad \sigma_V > 0.$$

For now, one risk-neutral insider observes V perfectly. Independent liquidity orders satisfy $Z \sim \mathcal{N}(0, \sigma_Z^2)$, with $\sigma_Z > 0$. Positive quantities mean purchases; negative quantities mean sales. The market maker sees only

$$Y = X + Z.$$

It sees a *signed aggregate imbalance*, not the decomposition into insider and liquidity orders. Anonymity here is about trader identity and information, not about erasing the buy–sell sign of total demand.

The timing is essential:

1. The insider observes V and chooses a market order X. It does not observe the current realization of Z.
2. Liquidity traders submit Z simultaneously.
3. Market makers observe Y, set $P(Y)$, and take the opposite position $-Y$.
4. The asset pays V, so the insider earns $\Pi_I = X(V - P)$.

All units execute at the auction's clearing price. Thus a buyer pays XP, not the integral of a marginal price schedule. This convention produces the factor of two in the insider's first-order condition.

Definition 4.1 (Linear Kyle equilibrium)**.** A linear equilibrium consists of an insider strategy and a pricing rule such that: (i) the strategy maximizes conditional expected profit given the pricing rule; and (ii) market makers price competitively,

$$P(Y) = \mathbb{E}[V \mid Y],$$

using the distribution of order flow induced by that strategy.

The market maker's inference is not an externally chosen statistical regression. Its regression coefficient depends on the insider's equilibrium strategy, which itself depends on that coefficient.

4.2.2 The insider's best response

Conjecture an affine price $P = \mu + \lambda Y$, with $\lambda > 0$. Conditional on $V = v$, the chosen x is fixed and the only execution uncertainty comes from Z. Independence and $\mathbb{E}[Z] = 0$ give

$$\mathbb{E}[x(V - P) \mid V = v] = x(v - \mu) - \lambda x^2.$$

Differentiating with respect to the insider's own order,

$$v - \mu - 2\lambda x = 0, \qquad \frac{\partial^2 \mathbb{E}[\Pi_I \mid v]}{\partial x^2} = -2\lambda < 0.$$

The optimum is therefore

$$x(v) = \frac{v - \mu}{2\lambda}. \tag{4.1}$$

More favorable information increases demand. Greater price impact restrains demand. The insider is risk neutral, but does not demand infinitely many shares: the restraint is strategic, not a dislike of risk.

4.2.3 The market maker's inference

Suppose the insider uses $X = a + bV$. Joint normality implies

$$\begin{aligned}
\mathbb{E}[V \mid Y] &= p_0 + \frac{\mathrm{Cov}(V, Y)}{\mathrm{Var}(Y)}\big(Y - \mathbb{E}[Y]\big) \\
&= p_0 + \frac{b\sigma_V^2}{b^2\sigma_V^2 + \sigma_Z^2}\big(Y - a - bp_0\big).
\end{aligned}$$

The subtraction of the mean of order flow matters when $p_0 \neq 0$. Matching this rule to the insider's best response yields $b = 1/(2\lambda)$, $a = -\mu/(2\lambda)$, and

$$\mu = p_0 - \lambda(a + bp_0) = p_0 - \frac{p_0 - \mu}{2}.$$

Hence $\mu = p_0$, and the equilibrium strategy can be centered as $X = b(V - p_0)$. The remaining consistency condition is

$$\lambda = \frac{b\sigma_V^2}{b^2\sigma_V^2 + \sigma_Z^2}, \qquad 2\lambda b = 1.$$

Combining them gives $b^2\sigma_V^2 = \sigma_Z^2$. Selecting the positive root, which is required by the concavity condition, establishes the result.

Proposition 4.1 (One-auction equilibrium)**.** *The unique equilibrium within the affine-linear class is*

$$X = \frac{\sigma_Z}{\sigma_V}(V - p_0), \qquad P = p_0 + \frac{\sigma_V}{2\sigma_Z}Y. \tag{4.2}$$

The trading intensity is $b = \sigma_Z/\sigma_V$ and the price-impact coefficient is $\lambda = \sigma_V/(2\sigma_Z)$.

The qualification "within the affine-linear class" identifies what the calculation proves. Guessing linear functions and verifying them is not, by itself, a proof that arbitrary non-linear equilibria cannot exist.

4.2.4 Depth, profits, and learning

Define market depth as $1/\lambda$. More liquidity-trading dispersion increases depth because a given imbalance becomes less diagnostic of information. The insider responds by trading more aggressively. Greater uncertainty about the payoff increases price impact because adverse selection is potentially more severe. These are equilibrium comparative statics: both the pricing rule and insider behavior adjust.

At the optimum,

$$\mathbb{E}[\Pi_I \mid V] = \frac{(V - p_0)^2}{4\lambda}, \qquad \mathbb{E}[\Pi_I] = \frac{\sigma_V\sigma_Z}{2}.$$

The liquidity traders' trading profit and trading cost are different signed objects:

$$\Pi_Z = Z(V - P), \qquad C_Z = Z(P - V) = -\Pi_Z.$$

Since Z is independent of V and X,

$$\mathbb{E}[\Pi_Z] = -\lambda\sigma_Z^2, \qquad \mathbb{E}[C_Z] = \lambda\sigma_Z^2 = \frac{\sigma_V\sigma_Z}{2}. \tag{4.3}$$

Market makers earn $\Pi_M = Y(P - V)$, with $\mathbb{E}[\Pi_M \mid Y] = 0$. Trading profits sum to zero in every realization:

$$\Pi_I + \Pi_Z + \Pi_M = 0.$$

This is a trading-profit accounting identity, not a complete welfare calculation. Liquidity traders may receive hedging or consumption benefits that are deliberately outside this model.

Finally,

$$\mathrm{Var}(P) = \frac{\sigma_V^2}{2}, \qquad \mathrm{Var}(V \mid P) = \frac{\sigma_V^2}{2}.$$

Price variation and residual uncertainty are distinct. The first measures variation in the market's estimate across states; the second measures how much uncertainty remains after observing that estimate. A more informative price can vary more across states while leaving less residual uncertainty.

Example 4.1 (A quantity decision with units)**.** Let $p_0 = 100$, $\sigma_V = 10$ currency units per share, and $\sigma_Z = 1{,}000$ shares. Then $b = 100$ shares per unit of price and $\lambda = 0.005$ units of price per additional share of imbalance. Here one unit of price means one currency unit per share. An insider observing $V = 112$ buys 1,200 shares. Its expected execution price, before observing liquidity demand, is 106, and conditional expected profit is $1{,}200(112 - 106) = 7{,}200$. Unconditional expected insider profit is 5,000. A realization can differ substantially because the insider cannot condition its order on the simultaneous liquidity order.

4.3 Imperfect information and the meaning of "half"

Perfect knowledge is not necessary. Suppose instead that the insider observes

$$S = V + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma_\eta^2),$$

with V, η, and Z independent apart from the definition of S. If observation involves two independent errors, $\eta = \epsilon + u$, then $\sigma_\eta^2 = \sigma_\epsilon^2 + \sigma_u^2$; neither error can simply be omitted. Write

$$\sigma_S^2 = \sigma_V^2 + \sigma_\eta^2, \qquad \phi = \frac{\sigma_V^2}{\sigma_S^2}.$$

Bayesian updating gives

$$M := \mathbb{E}[V \mid S] = p_0 + \phi(S - p_0), \qquad \mathrm{Var}(M) = \phi\sigma_V^2. \tag{4.4}$$

Here ϕ measures how strongly the signal changes the insider's valuation. The insider trades on M, not on the unobserved realization of V.

For a conjectured price $P = p_0 + \lambda Y$,

$$\mathbb{E}[X(V - P) \mid S] = X(M - p_0) - \lambda X^2.$$

Thus $X = \phi(S - p_0)/(2\lambda)$. If we write $X = b_S(S - p_0)$, the market maker's coefficient is

$$\lambda = \frac{b_S\sigma_V^2}{b_S^2\sigma_S^2 + \sigma_Z^2}.$$

Notice that noise-order variance is added *outside* the b_S^2 term: liquidity orders have not been multiplied by the insider's signal coefficient. Solving with $b_S = \phi/(2\lambda)$ gives

$$b_S = \frac{\sigma_Z}{\sigma_S}, \qquad \lambda = \frac{\sigma_V^2}{2\sigma_Z\sigma_S} = \frac{\sigma_V\sqrt{\phi}}{2\sigma_Z}, \qquad \mathbb{E}[\Pi_I] = \lambda\sigma_Z^2. \tag{4.5}$$

The insider's information advantage has standard deviation $\sqrt{\mathrm{Var}(M)} = \sigma_V\sqrt{\phi}$. The perfect-information formulas apply after replacing payoff uncertainty by uncertainty in this conditional valuation.

Using the Gaussian conditional-variance formula,

$$\begin{aligned}\mathrm{Var}(V \mid P) &= \sigma_V^2 - \frac{b_S^2\sigma_V^4}{b_S^2\sigma_S^2 + \sigma_Z^2} = \sigma_V^2\left(1 - \frac{\phi}{2}\right), \\ \mathrm{Var}(P) &= \frac{\phi\sigma_V^2}{2}.\end{aligned} \tag{4.6}$$

The insider's signal reduces uncertainty by $\phi\sigma_V^2$; the auction price reduces it by half that amount. This is the precise sense in which one auction releases half the insider's information: half of the *variance reduction*, not half of the signal's numerical value or half of Shannon information.

The decomposition makes the economics transparent:

$$\underbrace{\mathrm{Var}(V \mid P)}_{\text{public uncertainty}} = \underbrace{\sigma_V^2(1-\phi)}_{\text{uncertainty the insider also faces}} + \underbrace{\frac{\phi\sigma_V^2}{2}}_{\text{private valuation not yet revealed}}.$$

More noise trading increases both insider order dispersion and liquidity-order dispersion in the same proportion. Consequently, σ_Z changes depth and profits but not this information ratio in the risk-neutral one-auction equilibrium. More order flow need not mean a more informative price.

Example 4.2 (Separating signal quality from market depth)**.** Keep $\sigma_V = 10$ and $\sigma_Z = 1{,}000$, but let $\sigma_\eta = 10$. Then $\phi = 1/2$, $b_S \simeq 70.71$, and $\lambda \simeq 0.003536$. The insider's signal leaves conditional variance 50; the price leaves variance 75. The price has revealed half of the signal's 50-unit variance reduction, not half of the original 100-unit uncertainty. Expected insider profit is approximately 3,535.53.

4.4 Competition among insiders

Increasing the number of insiders is not the same as increasing the number of trading dates. Consider a *single* auction with $K \geq 1$ risk-neutral insiders, all observing the same V. Each maximizes its own profit, taking the other insiders' order strategies as given. If $P = p_0 + \lambda Y$, with $Y = \sum_{j=1}^{K} X_j + Z$, insider i solves

$$\max_{x_i}\ x_i\left(v - p_0 - \lambda x_i - \lambda\sum_{j\neq i} x_j\right).$$

The first-order condition is

$$2\lambda x_i = v - p_0 - \lambda \sum_{j \neq i} x_j.$$

In a symmetric equilibrium $X_i = b_K(V - p_0)$, so

$$b_K = \frac{1}{\lambda(K+1)}.$$

The market maker learns from aggregate informed demand $Kb_K(V - p_0)$, implying

$$\lambda = \frac{K b_K \sigma_V^2}{K^2 b_K^2 \sigma_V^2 + \sigma_Z^2}.$$

Combining the two equations gives

$$b_K = \frac{\sigma_Z}{\sigma_V \sqrt{K}}, \qquad \lambda_K = \frac{\sigma_V \sqrt{K}}{(K+1)\sigma_Z}. \tag{4.7}$$

Each insider trades less as K grows, but their aggregate intensity $Kb_K = \sigma_Z \sqrt{K}/\sigma_V$ increases. An individual insider internalizes the price effect on its own position, not the reduction in the other insiders' profits. Competition therefore erodes their joint monopoly power.

The informational and profit consequences are

$$\mathrm{Var}(V \mid P) = \frac{\sigma_V^2}{K+1}, \qquad \mathrm{Var}(P) = \frac{K}{K+1}\sigma_V^2, \tag{4.8}$$

$$\mathbb{E}[\Pi_i] = \frac{\sigma_V \sigma_Z}{(K+1)\sqrt{K}}, \qquad \mathbb{E}\left[\sum_i \Pi_i\right] = \frac{\sigma_V \sigma_Z \sqrt{K}}{K+1} = \mathbb{E}[C_Z]. \tag{4.9}$$

For $K \geq 1$, the function $\sqrt{K}/(K+1)$ is nonincreasing, and strictly decreases when moving between distinct integer values above one. Thus more identically informed risk-neutral competitors make prices more informative, lower price impact, and reduce aggregate liquidity-trader losses in this model. They do not increase those losses. Meanwhile, $\mathrm{Var}(P)$ increases because prices track fundamentals more closely. Higher price variance alone is not evidence of worse market quality.

4.4.1 A common noisy signal

The same argument extends when every insider observes the *same* noisy signal S. Define M as in (4.4) and $\sigma_M = \sigma_V \sqrt{\phi}$. Then

$$X_i = \frac{\sigma_Z}{\sigma_M \sqrt{K}}(M - p_0) = \frac{\sigma_Z}{\sigma_S \sqrt{K}}(S - p_0), \qquad \lambda_K = \frac{\sigma_M \sqrt{K}}{(K+1)\sigma_Z}.$$

Furthermore,

$$\operatorname{Var}(P) = \frac{K}{K+1}\phi\sigma_V^2, \quad \operatorname{Var}(V \mid P) = \sigma_V^2 - \frac{K}{K+1}\phi\sigma_V^2, \quad \mathbb{E}[\Pi_i] = \frac{\sigma_M \sigma_Z}{(K+1)\sqrt{K}}.$$

With $\sigma_V^2 = 1$ and signal-noise variance σ_η^2, the price variance becomes

$$\operatorname{Var}(P) = \left[(1+\sigma_\eta^2)\left(1+\frac{1}{K}\right)\right]^{-1}.$$

These formulas require common information. Independent private signals introduce inference about other insiders' information and cannot be inserted into the common-signal formulas unchanged. Risk aversion changes the best responses as well. Subrahmanyam's extension finds nonmonotonic liquidity relationships under risk aversion; that is a different equilibrium, not a universal hump-shaped law for every specification.[33]

Example 4.3 (Four insiders do not share monopoly profits)**.** For $\sigma_V = 10$ and $\sigma_Z = 1{,}000$, one perfectly informed insider earns 5,000 on average and leaves residual variance 50. With four such insiders, each earns 1,000, total informed profit falls to 4,000, and residual variance falls to 20. Price impact falls from 0.005 to 0.004. Dividing the monopoly profit by four would miss the endogenous change in their aggregate trading behavior.

4.5 Dynamic programming: the value of waiting

To trade repeatedly, the insider must compare profit today with the profit that today's trade leaves available tomorrow. Dynamic programming expresses this comparison using a state variable that summarizes the information relevant for future choices.

In a generic finite-horizon problem, let s_t be the state before the date-t action, a_t the action, and ε_t a shock arriving after that action. Suppose

$$s_{t+1} = f_t(s_t, a_t, \varepsilon_t),$$

and let $r_t(s_t, a_t, \varepsilon_t)$ be the current reward. A policy chooses actions using information available when they are chosen. The value function is the maximal expected reward from that date onward:

$$J_t(s) = \sup_{\text{admissible policies}} \mathbb{E}\left[\sum_{j=t}^{N} r_j \,\middle|\, s_t = s\right].$$

With no terminal reward after date N,

$$J_{N+1}(s) = 0, \qquad J_t(s) = \sup_a \mathbb{E}\left[r_t(s,a,\varepsilon_t) + J_{t+1}\big(f_t(s,a,\varepsilon_t)\big)\right]. \tag{4.10}$$

The second term is not tomorrow's realized reward. It is the value of all remaining optimized rewards, evaluated at tomorrow's state.

Backward induction begins at N, where there is no continuation value. Solving that problem identifies both an optimal final action and J_N. These determine the date-$N-1$ problem, and the calculation proceeds backward. Existence and uniqueness of a control do not follow merely from writing a Bellman equation or inverting a transition function. They require conditions such as a well-defined feasible set and appropriate concavity.

In linear–quadratic settings, one can conjecture a quadratic value function and verify that optimization preserves its form. This is a verification method, not a shortcut around optimality conditions. Random shocks usually produce an additive constant as well as the quadratic term. That constant may not affect the optimal control, but it matters for the level of expected profit.

4.6 Several auctions with one long-lived signal

Return to a single insider who knows the fixed payoff V. Auctions occur at $n = 1, \dots, N$. Let $\mathcal{F}_n^Y$ contain public order-flow history through auction n. Independent liquidity orders satisfy $Z_n \sim \mathcal{N}(0, q_n)$, where $q_n > 0$. If time intervals have length Δt_n and noise variance rate is σ_Z^2, then $q_n = \sigma_Z^2 \Delta t_n$.

Consider deterministic equilibrium coefficients in

$$X_n = b_n(V - P_{n-1}), \qquad P_n = P_{n-1} + \lambda_n(X_n + Z_n), \qquad P_n = \mathbb{E}[V \mid \mathcal{F}_n^Y].$$

Before auction n, the insider knows the residual mispricing

$$m_n = V - P_{n-1}.$$

It chooses X_n before seeing Z_n. The next state is $m_{n+1} = m_n - \lambda_n(X_n + Z_n)$, and its profit from this auction is $X_n(V - P_n)$. The continuation state uses the *post-trade* price; the current value function uses the *pre-trade* price.

4.6.1 The insider's recursion

Conjecture

$$J_n(m) = A_n m^2 + B_n, \qquad A_{N+1} = B_{N+1} = 0.$$

The public conditional-variance path and coefficient sequence are fixed when evaluating a deviation against the conjectured pricing strategy. Conditional on the insider's current information, its Bellman objective is

$$J_n(m) = \max_x \{mx - \lambda_n x^2 + A_{n+1}(m - \lambda_n x)^2 \\ + A_{n+1}\lambda_n^2 q_n + B_{n+1}\}.$$

The variance term appears because

$$\mathbb{E}\left[(m - \lambda_n x - \lambda_n Z_n)^2\right] = (m - \lambda_n x)^2 + \lambda_n^2 q_n.$$

Expanding the terms that contain x gives

$$m(1 - 2A_{n+1}\lambda_n)x - \lambda_n(1 - A_{n+1}\lambda_n)x^2.$$

If $\lambda_n > 0$ and $A_{n+1}\lambda_n < 1$, the objective is strictly concave. Its maximizing order is

$$X_n = b_n m_n, \qquad b_n = \frac{1 - 2A_{n+1}\lambda_n}{2\lambda_n(1 - A_{n+1}\lambda_n)}. \tag{4.11}$$

The positive-trading equilibrium branch has $A_{n+1}\lambda_n < 1/2$. Substituting the optimum into the objective gives

$$A_n = A_{n+1} + \frac{(1 - 2A_{n+1}\lambda_n)^2}{4\lambda_n(1 - A_{n+1}\lambda_n)} = \frac{1}{4\lambda_n(1 - A_{n+1}\lambda_n)}, \tag{4.12}$$

$$B_n = B_{n+1} + A_{n+1}\lambda_n^2 q_n. \tag{4.13}$$

The algebra verifies the quadratic conjecture. At the final auction, $A_{N+1} = 0$ and $b_N = 1/(2\lambda_N)$, recovering the static best response. Earlier, the continuation term reduces trading intensity relative to $1/(2\lambda_n)$ at a given λ_n: revealing less now preserves future opportunities.

4.6.2 The market maker's recursion

Define residual public uncertainty

$$\Sigma_{n-1} = \mathrm{Var}(V \mid \mathcal{F}_{n-1}^Y).$$

Conditional Gaussian projection gives

$$\lambda_n = \frac{b_n \Sigma_{n-1}}{b_n^2 \Sigma_{n-1} + q_n}, \qquad \Sigma_n = (1 - \lambda_n b_n)\Sigma_{n-1}, \qquad \Sigma_0 = \sigma_V^2. \tag{4.14}$$

The equilibrium couples the insider's backward-induction calculation from the terminal condition with the market maker's forward evolution from the prior variance. Equations

(4.11)–(4.14) must hold together. Solving the insider's Bellman problem for arbitrary price coefficients does not yet solve the market equilibrium.

Example 4.4 (Two auctions and gradual information release)**.** Normalize $\Sigma_0 = 1$ and set $q_1 = q_2 = 1$. Let $r = \lambda_1 b_1$ be the fraction of initial variance revealed at auction one. The pricing equations imply

$$b_1 = \sqrt{\frac{r}{1-r}}, \quad \lambda_1 = \sqrt{r(1-r)}, \quad \Sigma_1 = 1 - r.$$

The final-auction solution is

$$b_2 = \frac{1}{\sqrt{1-r}}, \quad \lambda_2 = \frac{\sqrt{1-r}}{2}, \quad A_2 = \frac{1}{2\sqrt{1-r}}.$$

Substituting into the first-auction best response gives

$$r = \frac{1-\sqrt{r}}{2-\sqrt{r}}.$$

Writing $u = \sqrt{r}$, solve $u^3 - 2u^2 - u + 1 = 0$ on the admissible branch. This yields $r \simeq 0.307979$, and hence

n	b_n	λ_n	Σ_n	A_n
1	0.667115	0.461658	0.692021	0.749496
2	1.202099	0.415939	0.346011	0.601049

The first auction reveals about 30.8% of initial variance, whereas the final auction reveals half the variance then remaining. It would be incorrect to halve initial variance at every date. Also, $B_1 \simeq 0.128100$, so ex ante total insider profit is $A_1\Sigma_0 + B_1 \simeq 0.877597$. The additive constant is economically measurable, not disposable notation.

4.6.3 Public prices are martingales

Because $P_n = \mathbb{E}[V \mid \mathcal{F}_n^Y]$, the tower property gives

$$\mathbb{E}[P_n \mid \mathcal{F}_{n-1}^Y] = P_{n-1}.$$

Thus the expected price change is zero given public history, although it need not be zero given the insider's information. The model can have gradual learning without mechanically generating publicly predictable returns. To explain momentum, reversals, or risk premia, one must specify the additional mechanism and the information set relative to which returns are predictable.

4.7 Belief aggregation and return predictability

The martingale result raises a natural question: what extra mechanism could produce predictable public returns? One possibility is disagreement about information quality. Kyle, Obizhaeva, and Wang study competitive traders who use Bayesian updating but disagree about signal precision. Their model distinguishes long-horizon valuations from positions taken in anticipation of future disagreement, and shows that this short-horizon speculation can dampen price movements and generate momentum.[28] This is not simply the strategic monopolist model with more dates: both the competitive trading structure and the belief assumptions differ.

The signal normalization in the study notes can be understood independently. In their zero-net-supply sketch, let $V \sim \mathcal{N}(0, 1/\tau_V)$ with $\tau_V > 0$ and define the standardized payoff $W = \sqrt{\tau_V} V$, so $\mathrm{Var}(W) = 1$. At dates $t = 1, 2$, a normalized signal with $\tau_i > 0$ can be written

$$I_{i,t} = \sqrt{\tau_i}\, W + \epsilon_{i,t}, \qquad \epsilon_{i,t} \sim \mathcal{N}(0, 1).$$

The coefficient τ_i is a signal-to-noise precision relative to the standardized payoff. Dividing by $\sqrt{\tau_i \tau_V}$ would produce a direct signal of V with error variance $1/(\tau_i \tau_V)$. The square roots therefore keep units and relative precision consistent; they are not an arbitrary mixture of variances.

If an observer receives a collection of such signals with independent errors and agrees on their precisions, direct completion of the Gaussian square gives

$$\mathbb{E}[W \mid I_1, \ldots, I_K] = \frac{\sum_{j=1}^{K} \sqrt{\tau_j} I_j}{1 + \sum_{j=1}^{K} \tau_j}, \qquad \mathrm{Var}(W \mid I_1, \ldots, I_K) = \frac{1}{1 + \sum_{j=1}^{K} \tau_j}.$$

For $K \geq 2$, an average signal of other traders is

$$I_{-i,t} = \frac{1}{K-1} \sum_{j \neq i} I_{j,t}.$$

Whether this average is observed directly or inferred from prices must be specified. Also, an unweighted average need not retain all information when signal loadings differ.

Disagreement changes the weights traders assign to these observations. A trader can expect a future resale price to differ from its own expected terminal payoff because it expects other traders to interpret future news differently. But the normalization above alone proves no momentum result. Return predictions require the equilibrium pricing and trading rules, the disagreement structure, and an explicit observer's probability measure. Bayesian updating under heterogeneous beliefs does not supply the common conditional-expectation martingale used in the preceding section.

4.8 Trading now, disclosing afterward

In the baseline sequential auction, the market sees total order flow but not the insider's individual orders. What changes if the insider must disclose an order after it has executed? Timing makes disclosure consequential even though it is retrospective: it can reveal information before the *next* trade.

Consider a deterministic strategy $X_1 = b_1(V - p_0)$ with known $b_1 > 0$. A later public report of X_1 reveals

$$V = p_0 + X_1/b_1.$$

The simultaneous liquidity order no longer protects the insider once its own order is separately identified. Applying the nondisclosure strategy unchanged would therefore destroy the continuation profit that justified gradual trading.

Huddart, Hughes, and Levine solve a dynamic disclosure model in which the insider responds by dissimulating: its trading behavior need not be a deterministic monotone reading of its information. Relative to their nondisclosure benchmark, information reaches prices faster and insider profits are lower.[24] With competitive risk-neutral makers, lower total informed rents also imply lower aggregate expected liquidity-trader losses by the trading-profit accounting identity. This need not mean that every individual trade or date has a lower price-impact coefficient. The important lesson is strategic adjustment. Disclosure changes the information released by an action and therefore changes the action itself. An isolated trade in the direction opposite to a trader's ultimate valuation need not imply that its information is poor.

Remark 4.1 (Institutional distinctions). The model's "insider" means an informationally advantaged trader, not necessarily a particular legal category. In the United States, beneficial-ownership reporting above the 5% threshold is distinct from Section 16 transaction reporting for directors, officers, and holders above 10%. Most covered transactions are reported on Form 4 within two business days. These different disclosure regimes should not be collapsed into a rule that every holder above 5% reports every trade identically.[34]

4.9 Funding capacity and the production of information

So far information has arrived without cost and the insider can finance any optimal position. The study notes extend the discussion to a different setting: a competitive rational-expectations economy in which specialists choose how much information to acquire and face portfolio constraints. The central issue is a feedback loop between the incentive to learn and the capacity to trade on what is learned. Glebkin, Gondhi, and Kuong formalize this mechanism.[15]

There are three dates. At date zero, specialists choose information precision. At date one, agents observe their information and trade. At date two, the risky asset pays

$$F = V + \Theta,$$

where V is learnable and Θ is an independent unlearnable component. The risk-free return is zero and risky-asset supply is one. Specialist i observes $S_i = V + \epsilon_i$, with error precision $\tau_{\epsilon,i}$, and an endowment loading $e_i = z + u_i$. Its nontraded endowment is $e_i\Theta$, so the same investor has both informational and hedging motives.

Writing $\mathcal{I}_i = \{P, S_i, e_i\}$, terminal wealth is

$$W_i = W_{0,i} + x_i(F - P) + e_i\Theta.$$

The specialist maximizes $\mathbb{E}[-\exp(-\gamma W_i) \mid \mathcal{I}_i]$ subject to admissible positions $\ell(P) \leq x_i \leq h(P)$. The leading minus sign in CARA utility is essential: without it, maximizing the expression would not represent increasing risk-averse preferences.

For a conditional Gaussian benchmark, the position-dependent part of the certainty equivalent is

$$x_i\big(\mathbb{E}[F \mid \mathcal{I}_i] - P\big) - \frac{\gamma}{2}\left[x_i^2 \operatorname{Var}(F \mid \mathcal{I}_i) + 2x_i e_i \operatorname{Cov}(F, \Theta \mid \mathcal{I}_i)\right].$$

Hence the unconstrained demand is

$$x_i^{\text{unc}} = \frac{\mathbb{E}[F \mid \mathcal{I}_i] - P - \gamma e_i \operatorname{Cov}(F, \Theta \mid \mathcal{I}_i)}{\gamma \operatorname{Var}(F \mid \mathcal{I}_i)}.$$

With interval constraints and this concave objective, demand is the constrained projection

$$x_i = \min\{h(P), \max\{\ell(P), x_i^{\text{unc}}\}\}.$$

Market clearing then requires

$$\int_0^1 x_i(P, S_i, e_i)\, di + x_m(P) = 1,$$

where x_m is aggregate nonspecialist demand. This is an equilibrium condition, not a claim that the price remains affine under constraints. The Gaussian certainty-equivalent calculation must be used only when the relevant conditional distribution justifies it.

Why can an investor “choose precision”? Precision summarizes the output of research effort: more analysis, better data, or closer monitoring can reduce signal error. Let $\bar{\tau}_\epsilon > 0$ be baseline precision and let incremental information cost $C(\tau_{\epsilon,i} - \bar{\tau}_\epsilon)$. Define $CE_{0,i}^{\text{gross}}(\tau)$ as the ex ante certainty equivalent of the optimized trading opportunity before deducting the deterministic information cost. An interior choice satisfies

$$\frac{\partial CE_{0,i}^{\text{gross}}}{\partial \tau_{\epsilon,i}} = C'(\tau_{\epsilon,i} - \bar{\tau}_{\epsilon}). \tag{4.15}$$

The equality is between marginal *value* and marginal *cost*, not between cost and precision itself. It is an intensive-margin optimality condition; a separate free-entry interpretation would require an entry decision. At the lower boundary, marginal value may be below marginal cost, so equality need not hold.

Tighter funding constraints limit the positions through which information earns a return. This can reduce the incentive to acquire information, weakening price informativeness. Less informative prices leave financiers more exposed to uncertainty about collateral value, potentially raising required margins and tightening constraints further. The initiating shock is a loss of funding capacity or wealth; a change in the asset price is not automatically a direct increase in the technology's research cost. This information channel complements, rather than replaces, the inventory and funding mechanisms considered elsewhere in market microstructure.

4.10 Discrete games, continuous limits, and inconspicuous orders

The Gaussian calculations give unusually transparent formulas. A more general treatment must also specify behavior at histories that are unlikely or never reached. An extensive-form game records the information received by the insider, the simultaneous insider and noise orders, the aggregate observation available to the market maker, and the subsequent pricing decision. A sequential equilibrium includes strategies and beliefs consistent with sequential optimization and the relevant limiting Bayesian-consistency requirement.

Kühn and Lorenz prove existence for discrete Kyle games with finite action and information structures and finite-support distributions of payoff and noise demand. Mixed strategies may be needed. They also obtain an equilibrium with a continuum of actions through approximation by finite-action games.[25] A continuum of possible order *sizes* is not the same as continuous *time*. Nor does the finite-support theorem assert existence for every arbitrary unbounded distribution without additional conditions.

4.10.1 What inconspicuousness means

In certain continuous-time models, an equilibrium can be constructed so that cumulative total order flow $Y = X + Z$ has the same public law as the exogenous noise process Z:

$$\mathcal{L}\big((Y_t)_{0\leq t\leq T}\big) = \mathcal{L}\big((Z_t)_{0\leq t\leq T}\big).$$

This is a statement about the distribution of the whole observable process, not a statement that the insider submits zero orders or that $Y_t = Z_t$ path by path. It is also stronger than an informal claim that orders provide camouflage.

The one-auction equilibrium demonstrates why the distinction matters. There, X and Z are independent and each has variance σ_Z^2, giving

$$\mathrm{Var}(Y) = 2\sigma_Z^2 \neq \mathrm{Var}(Z).$$

Thus the one-auction Kyle equilibrium is not inconspicuous in the same-law sense. In continuous time, the insider responds to the history of prices and noise-induced movements; cumulative informed and noise orders need not be independent.

For intuition, consider a Gaussian continuous-time benchmark over a horizon $T > 0$. Let $V \sim \mathcal{N}(p_0, \sigma_V^2)$, and let $Z_t = \sigma_Z B_t$, where B is a standard Brownian motion independent of V and $\sigma_V, \sigma_Z > 0$. Start with $X_0 = Y_0 = 0$. The candidate equilibrium is

$$P_t = p_0 + \lambda Y_t, \qquad \lambda = \frac{\sigma_V}{\sigma_Z\sqrt{T}}, \qquad dX_t = \frac{V - P_t}{\lambda(T-t)}\,dt, \quad 0 \le t < T. \tag{4.16}$$

The insider trades toward its valuation, with a response coefficient that increases as the horizon approaches. To check the pricing rule, set $M_t = V - P_t$. Substituting the candidate strategy gives

$$dM_t = -\frac{M_t}{T-t}\,dt - \lambda\sigma_Z\,dB_t, \qquad M_t = (T-t)\left[\frac{V-p_0}{T} - \lambda\sigma_Z \int_0^t \frac{dB_s}{T-s}\right].$$

Independence and the Itô isometry then give $\mathrm{Var}(M_t) = \sigma_V^2(T-t)/T$ and $\mathrm{Cov}(M_t, P_s) = 0$ for every $0 \le s \le t < T$. Joint Gaussianity makes the residual independent of the entire public price history, not merely uncorrelated with the current price. Since price and total-flow histories determine each other, $P_t = \mathbb{E}[V \mid \mathcal{F}_t^Y]$ follows. Total flow therefore has zero public conditional drift and quadratic variation $\sigma_Z^2 t$; under the stated filtering and admissibility conditions it is Brownian in its public filtration. These calculations verify inference consistency; optimality of an admissible insider strategy is a separate equilibrium requirement.

The public residual variance is consequently

$$\Sigma_t = \sigma_V^2\left(1 - \frac{t}{T}\right),$$

so information is fully incorporated at the terminal limit. The trading intensity in (4.16) is a rate, not the discrete order coefficient b_n. Likewise, σ_Z is now a noise-flow standard deviation per square root of time. Confusing rates, increments, and cumulative quantities produces erroneous comparisons between the discrete and continuous models. Kyle's original work establishes the continuous-trading limit for the Gaussian benchmark.[26]

4.10.2 Terminal monotone coupling

Another useful ingredient in continuous-time constructions is a monotone relation between terminal order flow and value. Suppose a desired terminal total-flow distribution has continuous cumulative distribution function F_Y, and the payoff has distribution function F_V. Then

$$U = F_Y(Y_T) \sim \text{Uniform}(0,1), \qquad V = F_V^{-1}(U),$$

where the generalized quantile is

$$F_V^{-1}(u) = \inf\{v \in \mathbb{R} : F_V(v) \geq u\}, \qquad 0 < u < 1.$$

Thus a monotone terminal map is

$$V = g(Y_T), \qquad g(y) = F_V^{-1}\big(F_Y(y)\big). \tag{4.17}$$

Large total purchases are paired with high terminal values. For example, if Y_T is standard normal and V is uniform on $[0,1]$, the map is $g(y) = \Phi(y)$, where Φ is the standard normal distribution function.

Continuity of F_Y matters: when Y_T has atoms, $F_Y(Y_T)$ is generally not uniform. Randomized distributional transforms or a more general coupling may be required, and a deterministic map cannot always produce an arbitrary payoff distribution from an atomic variable. Moreover, a terminal coupling is not a complete equilibrium proof. One must construct an admissible insider strategy that generates it and verify optimality and rational pricing at every earlier time.

Finally, market makers price at conditional expected value because competition and risk neutrality impose their break-even condition. They do not need a monotone terminal map to justify that rule. The coupling helps construct particular equilibria; it is not the economic reason for conditional-expectation pricing.

4.11 Exercises

Exercise 4.1 (Profit signs and inventory risk)**.** An insider buys $x = 100$ shares at $P = 48$ and the payoff is $V = 52$. Compute its profit and the profit of a counterparty taking the opposite position. Next derive a CARA dealer's marginal valuation from a normal payoff's certainty equivalent. Explain exactly which inventory effect disappears under risk neutrality.

Exercise 4.2 (Two layers of signal error)**.** Let $V \sim \mathcal{N}(50, 36)$ and let the insider observe $S = V+\epsilon+u$, where the independent error variances are 9 and 19. Liquidity-order variance is 400. Find the posterior valuation, trading intensity on $S-50$, price impact, expected insider profit, and public residual variance. Interpret the last quantity using the half-information result.

Exercise 4.3 (Competition and market quality)**.** Take perfect information, $\sigma_V = 6$, and $\sigma_Z = 20$. Compare $K = 1$ and $K = 9$ for price impact, each insider's profit, aggregate liquidity cost, and price variance. Explain why price variance and liquidity cost move in opposite directions. Would the same formulas apply if the nine insiders had independent noisy signals?

Exercise 4.4 (The continuation-value penalty)**.** At a given auction suppose the conjectured impact is $\lambda = 0.5$, next-period value is $J_{n+1}(m) = 0.4m^2 + 0.1$, and noise variance is $q = 1$. Find the current trading intensity and the two current value-function coefficients. Check concavity. Explain why this is a verified best response but not yet a verified market equilibrium.

Exercise 4.5 (Disclosure and public predictability)**.** Suppose the insider uses $X_1 = b(V - p_0)$ and b is publicly known and nonzero. Explain what is learned when X_1 is separately disclosed after the auction. Separately show, using the tower property, why gradual learning in a nondisclosure Kyle model does not imply that past public price changes predict the next price change.

Exercise 4.6 (A coupling is not a trading strategy)**.** Let $Y_T \sim \mathcal{N}(0, 4)$ and let the desired payoff be uniform on $[80, 120]$. Construct a monotone map from Y_T to V. State two additional things that must be checked before calling this construction a Kyle equilibrium. Explain why the probability integral transform fails unchanged if Y_T takes only two values.

4.12 Solutions and discussion

1. Profit signs and inventory risk. The insider earns $100(52-48) = 400$; the opposite position earns -400. For a dealer holding q shares, the certainty equivalent is $qm - \gamma q^2 s^2/2$, so the marginal valuation is $m - \gamma q s^2$. At $\gamma = 0$ the risk-aversion inventory premium vanishes. This calculation says nothing about financing charges or binding position constraints, which were absent from its assumptions.

2. Two layers of signal error. Total signal variance is $36 + 9 + 19 = 64$, so $\phi = 36/64 = 0.5625$ and the posterior valuation is $M = 50 + 0.5625(S-50)$. Since $\sigma_Z = 20$ and $\sigma_S = 8$, the order coefficient is $b_S = 20/8 = 2.5$. Impact is $\lambda = 36/(2 \cdot 20 \cdot 8) = 0.1125$, and expected profit is $\lambda\sigma_Z^2 = 45$. Public residual variance is $36(1 - 0.5625/2) = 25.875$. The signal alone removes 20.25 units of variance, whereas the price removes 10.125. Omitting either observation-error variance would overstate information quality.

3. Competition and market quality. For one insider, $\lambda_1 = 0.15$, insider profit and aggregate liquidity cost are both 60, and $\mathrm{Var}(P) = 18$. For nine insiders, $\lambda_9 = 18/200 = 0.09$, each profit is $120/(10 \cdot 3) = 4$, aggregate cost is 36, and $\mathrm{Var}(P) = 32.4$. Competition reduces information rents but makes prices more responsive to actual payoff differences. Independent noisy signals require a different inference and strategic calculation; they are not nine copies of a common observation.

4. The continuation-value penalty. Here $A_{n+1}\lambda = 0.2$. Formula (4.11) gives $b = (1 - 0.4)/(1 \cdot 0.8) = 0.75$. The current coefficients are $A_n = 1/(4 \cdot 0.5 \cdot 0.8) = 0.625$ and $B_n = 0.1 + 0.4(0.5)^2 = 0.2$. The second derivative is $-2\lambda(1 - A_{n+1}\lambda) = -0.8 < 0$. A myopic trader would use $b = 1$, so the continuation value reduces current trading. Market equilibrium still requires the price-impact coefficient to equal the Bayesian projection coefficient for the induced order flow and prior residual variance.

5. Disclosure and predictability. Disclosure reveals $V = p_0 + X_1/b$, eliminating private information under this deterministic strategy. A model with repeated disclosure must therefore reconsider equilibrium behavior, including possible randomization. Without disclosure, the public price is $P_n = \mathbb{E}[V \mid \mathcal{F}_n^Y]$. Taking its expectation conditional on $\mathcal{F}_{n-1}^Y$ gives P_{n-1}, so $\mathbb{E}[P_n - P_{n-1} \mid \mathcal{F}_{n-1}^Y] = 0$. Slow variance reduction and predictable public returns are not equivalent statements.

6. Coupling and equilibrium. Since $Y_T/2$ is standard normal, use $V = 80 + 40\Phi(Y_T/2)$. The map is increasing and has the desired uniform marginal. One must also produce an admissible trading strategy consistent with the insider's information and prove that it is optimal against prices satisfying conditional-expectation pricing. If Y_T has two possible values, a deterministic transformation has at most two values as well; it cannot create a continuous uniform distribution. Extra randomization can construct a coupling, but must itself be compatible with the trading model.

The organizing lesson

Private information affects prices through feasible behavior. Strategic restraint preserves information rents; competition makes insiders reveal more. Repeated trading gives information a continuation value. Disclosure changes what actions reveal, while funding limits change the return to learning. Every equilibrium must reconcile the trader's optimal behavior with the market's inference from that behavior.

Lecture 5

The probability and timing of limit-order execution

A resting limit order offers a price and waits for somebody else to trade. The investor gains control over the worst acceptable execution price but gives up control over the execution time. That exchange is the starting point for an economic theory of passive trading. A favorable quoted price is useful only if the order executes; a purchase that occurs just before bad news may be attractive in price but costly in value.

There are two natural ways to study the waiting time. A first-passage model asks when a price first reaches the order's limit. Survival analysis starts instead from recorded order episodes and asks when the matching system actually fills them. We develop the price-touch benchmark first, then introduce the risk sets, censoring and cancellation needed to study execution itself.

5.1 Define the event before modeling the time

Fix an order submitted at time zero with quantity $Q_0 > 0$ and limit price L. A resting buy can execute at L or a lower price; a resting sell can execute at L or a higher price. A still-active order is a standing instruction to trade under the venue's rules. A cancellation request does not retrospectively invalidate an execution that the matching engine has already processed. Let Q_t^{fill} be the cumulative executed quantity. Distinct event times are

$$
\begin{aligned}
T_1 &= \inf\{t \geq 0 : Q_t^{\text{fill}} > 0\}, && \text{time to first fill,}\\
T_Q &= \inf\{t \geq 0 : Q_t^{\text{fill}} = Q_0\}, && \text{time to completion,}\\
T_C &= \inf\{t \geq 0 : \text{the remaining order is cancelled}\}, && \text{time to cancellation.}
\end{aligned}
$$

An order may receive a first fill but never complete. A price amendment may create a new priority position or a new order episode, depending on the venue and the data definition. An expiry or the end of an observation window is yet another event. A model of T_1 cannot be presented as a model of T_Q without additional information about the remaining quantity.

Lo, MacKinlay and Zhang [30] examine actual limit-order execution times using survival analysis. Their distinction between observed executions and hypothetical executions is central here. Historical sample percentages of day orders, or differences between buyers and sellers, are features of a particular dataset and trading environment; they are not defining properties of all limit orders.

5.2 A diffusion benchmark for touching a price

Temporarily set aside the queue and suppose that execution opportunities are summarized by a reference price reaching a barrier. This simplification separates a tractable price-path question from the matching details that we will restore afterward.

Suppose a reference price follows geometric Brownian motion,

$$\frac{dP_t}{P_t} = \alpha\, dt + \sigma\, dW_t, \qquad P_0 > 0, \quad \sigma > 0, \tag{5.1}$$

with constant parameters. Here α is the proportional drift per unit time and σ is volatility per square root of time. Log prices convert the multiplicative process into an additive one. Itô's formula gives

$$X_t := \log(P_t/P_0) = mt + \sigma W_t, \qquad m = \alpha - \frac{\sigma^2}{2}.$$

Therefore $X_t \sim N(mt, \sigma^2 t)$: both its mean displacement and its variance accumulate with elapsed time.

For a buy limit below the initial reference price, set $L < P_0$ and $a = \log(P_0/L) > 0$. Define the downward hitting time

$$\tau_- = \inf\{t \geq 0 : X_t \leq -a\}.$$

Its distribution function is

$$F_-(t) = \Pr(\tau_- \leq t) = \Phi\left(\frac{-a - mt}{\sigma\sqrt{t}}\right) + e^{-2ma/\sigma^2}\Phi\left(\frac{-a + mt}{\sigma\sqrt{t}}\right), \quad t > 0, \tag{5.2}$$

where Φ is the standard normal distribution function. Both the sign of the barrier and the sign of the drift matter. The event concerns the running minimum of the entire path, not just P_t at its endpoint.

For completeness, a sell barrier $U > P_0$ has log-distance $b = \log(U/P_0) > 0$ and upward hitting time $\tau_+ = \inf\{t : X_t \geq b\}$. Reflecting the process gives

$$F_+(t) = \Phi\left(\frac{mt - b}{\sigma\sqrt{t}}\right) + e^{2mb/\sigma^2}\Phi\left(\frac{-mt - b}{\sigma\sqrt{t}}\right). \tag{5.3}$$

For a sell barrier the relevant path statistic is the maximum, not the minimum.

5.2.1 Where the hitting formula comes from

When $m = 0$, the reflection principle pairs paths that hit $-a$ and finish above $-a$ with reflected paths finishing below $-a$. Hence

$$\Pr\left(\min_{0\le s\le t} \sigma W_s \le -a\right) = 2\Phi\left(-\frac{a}{\sigma\sqrt{t}}\right).$$

For nonzero drift, use the transition density of Brownian motion killed when it reaches $-a$. For $x > -a$, its density is

$$g(t,x) = \frac{1}{\sigma\sqrt{t}}\left[\phi\left(\frac{x-mt}{\sigma\sqrt{t}}\right) - e^{-2ma/\sigma^2}\phi\left(\frac{x+2a-mt}{\sigma\sqrt{t}}\right)\right],$$

where ϕ is the standard normal density. The second term is the reflected image, weighted for the drift. Integrating over the surviving region gives

$$S_-(t) = \Phi\left(\frac{a+mt}{\sigma\sqrt{t}}\right) - e^{-2ma/\sigma^2}\Phi\left(\frac{-a+mt}{\sigma\sqrt{t}}\right).$$

Taking $1-S_-(t)$ produces equation (5.2). This argument also shows why simply subtracting a deterministic drift from the running minimum does not work: the minimizing time is itself random.

Differentiating the hitting CDF gives

$$f_-(t) = \frac{a}{\sigma\sqrt{2\pi t^3}}\exp\left[-\frac{(a+mt)^2}{2\sigma^2 t}\right]. \tag{5.4}$$

If $m \le 0$, the downward barrier is eventually hit with probability one. If $m > 0$, the eventual hitting probability is $e^{-2ma/\sigma^2} < 1$. There is then positive probability mass at $\tau_- = \infty$; a finite-time density need not integrate to one over $(0,\infty)$. With zero drift, eventual hitting is certain but the unconditional mean hitting time is infinite. Probability one does not mean a short or even integrable waiting time.

Example 5.1 (Touching is more likely than finishing below)**.** Let $P_0 = 100$, $L = 99$, $\sigma = 0.02$ per square-root day, $t = 1$ day and log-drift $m = 0$. The log-distance is $a = \log(100/99) \approx 0.0100503$. Equation (5.2) gives

$$F_-(1) = 2\Phi(-0.502517) \approx 0.6153.$$

The probability of finishing the day below 99 is only $\Phi(-0.502517) \approx 0.3077$. Paths that hit 99 and then recover account for the difference. With the same volatility and zero log-drift, a limit of 98 has one-day touch probability approximately 0.3124, using its exact log-distance $\log(100/98)$.

Figure 5.1 extends the example across horizons. Waiting longer raises the probability of reaching either fixed barrier; moving the barrier farther away reduces the probability at a given horizon.

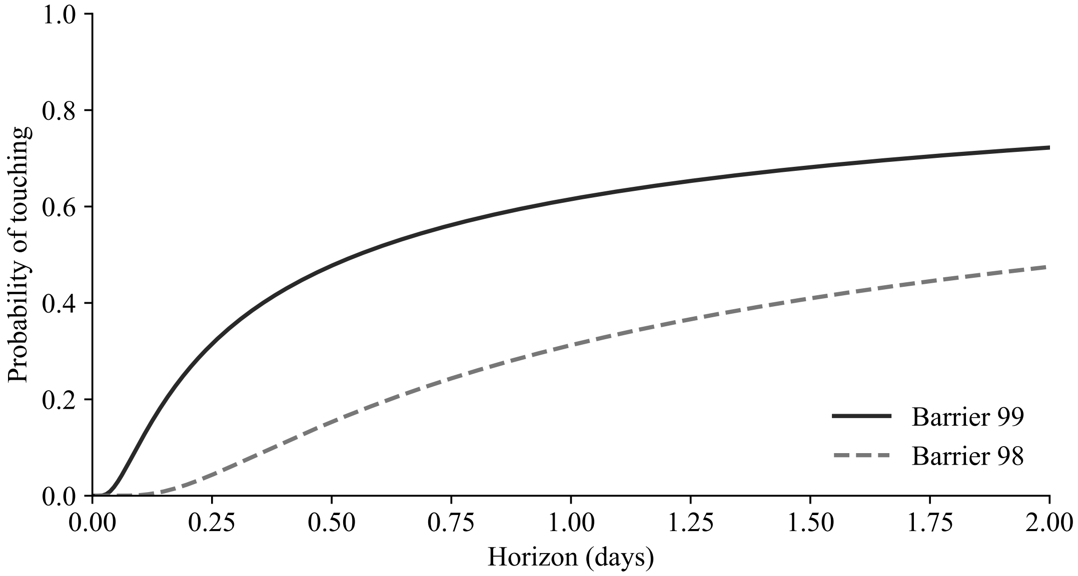

Figure 5.1: Hypothetical downward barrier-touch probabilities for $P_0 = 100$, zero log-drift and volatility $\sigma = 0.02$ per square-root day. The lower barrier at 98 is less likely to be reached than the barrier at 99. These are price-path probabilities, not queue-aware limit-order fill probabilities.

5.2.2 From a price touch to a fill

A reference midpoint may touch a buy limit without any seller trading against the order. Even if a transaction occurs at the order's price, other orders may have priority. A large order may receive a partial fill, and an order cancelled before the touch is no longer eligible. Hidden liquidity, routing and venue fragmentation further separate a consolidated price path from the matching history of one particular order.

The diffusion calculation is therefore a *hypothetical barrier-touch probability*. It becomes an execution probability under a rule such as immediate complete execution at the first touch, with no queue or cancellation. For other matching rules the relation must be derived. In particular, a touch probability is not a universal upper or lower bound when the reference series can be a midpoint, a transaction price or a quote from another venue. This is the point at which actual order histories become indispensable.

5.3 Survival functions and hazards

The empirical question changes once we observe actual orders: among orders still waiting at a given age, how quickly do fills arrive? Let T now denote the actual event time of interest, such as first fill. Its CDF and survival function are

$$F(t) = \Pr(T \le t), \qquad S(t) = \Pr(T > t) = 1 - F(t).$$

If the finite-time distribution has a density f, then $f(t) = F'(t) = -S'(t)$. A density is not the probability of an event occurring at an exact instant: that probability is zero for a continuous distribution. Instead, $f(t)\,dt$ approximates the probability of an event in a short interval around t.

The hazard is

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{\Pr(t < T \le t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{S(t)}, \tag{5.5}$$

where $S(t) > 0$. It has units of inverse time and can exceed one. It measures an instantaneous conditional rate, not an unconditional probability.

For continuous waiting times with no immediate-event probability mass, $S(0) = 1$. Because $S'(t) = -h(t)S(t)$,

$$\frac{d}{dt} \log S(t) = -h(t), \qquad H(t) = \int_0^t h(u)\,du, \qquad S(t) = e^{-H(t)}.$$

The cumulative hazard $H(t)$ is dimensionless. For an order that has already waited s units of time, the probability of a fill in the next u units is

$$\Pr(T \le s + u \mid T > s) = 1 - \frac{S(s+u)}{S(s)}. \tag{5.6}$$

The past waiting time changes the relevant population: we are examining the orders still unfilled, not the original submission cohort.

Example 5.2 (A constant hazard)**.** If $h(t) = 0.05$ per minute, then $S(t) = e^{-0.05t}$. The probability of filling within ten minutes is $1 - e^{-0.5} \approx 0.3935$. Given survival for twenty minutes, the next-ten-minute fill probability is still 0.3935. This memorylessness is a property of the exponential model. A limit order with a changing queue position or time-varying covariates need not have it.

5.4 Right censoring and the information in an unfilled order

An order need not fill during the period for which it is observed. Such an episode still tells us that the waiting time exceeded its observed duration. Survival methods retain this partial information.

Suppose observation ends at time C. We observe $Y = \min(T, C)$ and event indicator $\delta = \mathbf{1}\{T \leq C\}$. Under independent censoring, or an appropriate conditional version given observed covariates, the event-time likelihood contribution is

$$L_i = f(Y_i \mid x_i)^{\delta_i} S(Y_i \mid x_i)^{1-\delta_i}. \tag{5.7}$$

A filled order contributes its event density. An order still unfilled when observation stops contributes the probability of surviving that long. Dropping unfilled orders throws away precisely the evidence about long waiting times and can severely bias the estimated execution distribution.

For nonparametric estimation, let $t_1 < \cdots < t_k$ be distinct observed event times, d_j the number of events at t_j, and n_j the number at risk immediately before t_j. The Kaplan–Meier estimator is

$$\widehat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right). \tag{5.8}$$

Each factor estimates conditional survival through one event time. The product follows the chain rule: survival to a later time requires survival through every earlier risk set. A censored order remains in the risk set until its censoring time and then leaves without causing a downward step.

Consider five submitted orders: fills at minutes 2, 4 and 7, and censoring at minutes 3 and 8. At minute 2 all five are at risk, so survival falls to $4/5$. The censoring at minute 3 leaves three at risk immediately before minute 4. Survival then becomes $(4/5)(2/3) = 8/15$. Two orders remain at risk immediately before minute 7, so survival falls to $4/15$. The final censoring at minute 8 causes no further drop. The data do not demonstrate that the remaining probability mass fills at minute 8 or at any later finite time.

Figure 5.2 shows the different effects of a fill and a censored observation: a fill lowers estimated survival, while censoring changes the risk set without a downward step.

5.4.1 Cancellation as a competing event

The end of a data file can be administrative censoring. A trader's decision to cancel because the market moved is economically different. If cancellation depends on the same state

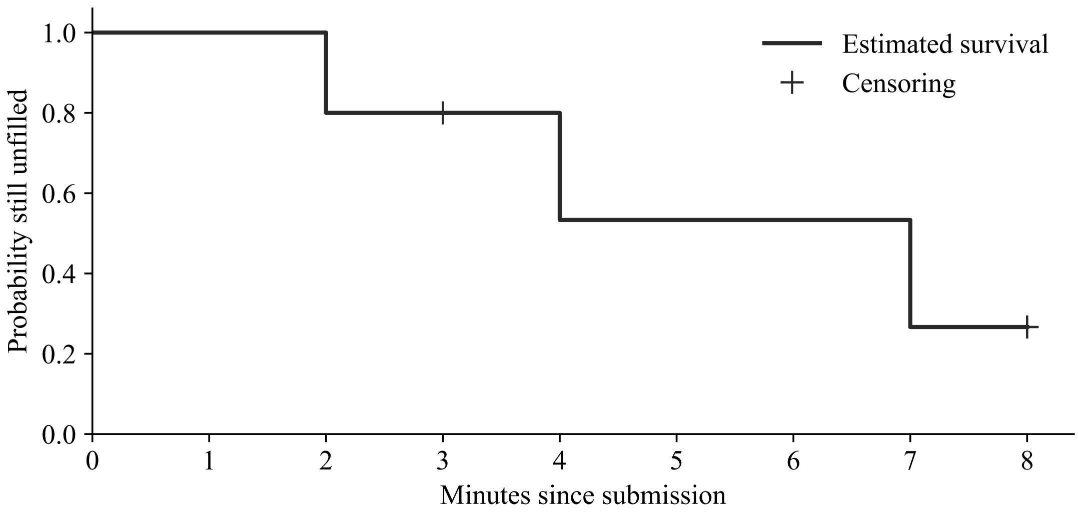

Figure 5.2: Kaplan–Meier survival for the five-order example. Fills at minutes 2, 4 and 7 create downward steps. Plus signs mark censoring at minutes 3 and 8; these observations remain informative about survival up to their censoring times. The curve is not evidence of what happens beyond the observation window.

variables that predict execution, treating it as independent censoring is not automatically valid.

For the probability of a *first fill before cancellation*, model two cause-specific hazards, $h_F(t)$ for first fill and $h_C(t)$ for cancellation, among orders experiencing neither event. Completion after partial fills requires further order states rather than this two-event model. Then

$$S_{\text{active}}(t) = \exp\left[-\int_0^t \{h_F(u) + h_C(u)\}\,du\right]$$

and the cumulative incidence of a fill is

$$F_F(t) = \int_0^t S_{\text{active}}(u) h_F(u)\,du. \tag{5.9}$$

For constant rates $h_F = 0.05$ and $h_C = 0.10$ per minute, the probability of filling before cancellation and within ten minutes is

$$\frac{0.05}{0.15}(1 - e^{-1.5}) \approx 0.2590.$$

The value $1 - e^{-0.5} \approx 0.3935$ answers the different question generated by a fill-only exponential model. Cause-specific hazards can be estimated from the observed active risk sets; interpreting the removal of cancellation as a policy counterfactual requires assumptions about how fill behavior would change under that intervention.

5.5 Accelerated failure time models

Risk sets describe which orders remain exposed; a regression model can then relate their waiting times to characteristics such as price distance and volatility. An accelerated failure time (AFT) specification does this by stretching or compressing a baseline clock. Write

$$T = e^{x'\beta} T_0, \qquad \text{or} \qquad \log T = x'\beta + \varepsilon,$$

where the baseline distribution of $T_0 = e^{\varepsilon}$ must also be specified or estimated. Conditional on x,

$$S(t \mid x) = S_0(te^{-x'\beta}), \qquad h(t \mid x) = e^{-x'\beta} h_0(te^{-x'\beta}).$$

A positive coefficient lengthens waiting times under this sign convention. Increasing a covariate by one unit multiplies every corresponding conditional quantile by e^{β_j}. For example, $\beta_j = \log 2$ doubles the median waiting time, as well as other finite quantiles. An AFT coefficient is not generally a proportional-hazards coefficient.

The ordinary log-time regression uses positive, finite latent event times. Right censoring can hide such a time without making it infinite. If the model instead assigns positive probability to never filling, as the upward-drift barrier example can, that probability must be represented separately; dropping those observations does not turn the model into a standard AFT regression.

Economically useful covariates include the limit's distance from an executable quote, spread, volatility, time of day and queue-related variables available at submission. A time-varying covariate must be constructed using information available before the event whose rate it explains. End-of-order realized price movements cannot be inserted into a forecast at submission without introducing future information.

Finally, a predicted fill probability is only one part of a trading decision. A buy order may be more likely to fill when the asset's value is falling. Expected gains depend jointly on execution, post-fill value, fees, inventory and the alternative use of time. Survival analysis describes the clock; an execution policy also needs the economics of the event that stops it.

5.6 Exercises

Exercise 5.1 (A barrier and a terminal event)**.** For zero log-drift and $a/(\sigma\sqrt{t}) = 1$, compare the probabilities of touching the lower barrier and ending below it. Explain the difference without algebra. What happens to the touch probability as $t \downarrow 0$ for fixed $a > 0$?

Exercise 5.2 (Conditional waiting)**.** Suppose $S(t) = \exp[-(t/10)^2]$, with t in minutes. Find $h(t)$ and the probability of a fill by minute 10 conditional on being unfilled at minute 5. Compare it with the unconditional first-five-minute fill probability.

Exercise 5.3 (Risk sets)**.** Six orders have fills at minutes 1, 3, 3 and 6, and censoring at minutes 2 and 5. Compute the Kaplan–Meier survival estimate immediately after each fill time. State how tied fills enter the formula.

Exercise 5.4 (Competing clocks)**.** If the constant fill rate is 0.08 and the cancellation rate 0.12 per minute, find the eventual probability of filling first and the expected time until either event. Why is the latter not the unconditional mean time to fill?

Exercise 5.5 (Interpreting acceleration)**.** An AFT model has coefficient −0.4 on a one-unit increase in price aggressiveness. By what factor does the median waiting time change? Explain why this coefficient alone does not establish the profit effect of a more aggressive order.

5.7 Solutions and discussion

1.The touch probability is $2\Phi(-1) \approx 0.3173$, while the terminal probability is $\Phi(-1) \approx 0.1587$. Some paths cross and recover. As the horizon tends to zero, both probabilities tend to zero for a barrier at strictly positive log-distance.

2.The hazard is $h(t) = 2t/100$. Conditional survival from minute 5 to 10 is $e^{-1}/e^{-0.25} = e^{-0.75}$, so the conditional fill probability is approximately 0.5276. The first-five-minute probability is $1 - e^{-0.25} \approx 0.2212$. The rising hazard makes age relevant.

3.At minute 1 the risk set is six, giving $\widehat{S} = 5/6$. The minute-2 censoring leaves four at risk before minute 3; two fills at that time give $\widehat{S} = (5/6)(1 - 2/4) = 5/12$. The minute-5 censoring leaves one at risk before the final fill at minute 6, so the estimate falls to zero. A tie is handled with $d_j = 2$, not by forgetting that both events share the same pre-event risk set.

4.The eventual fill-first probability is $0.08/(0.08 + 0.12) = 0.4$. The mean time to the first event is $1/0.20 = 5$ minutes. For 60% of orders the first event is cancellation; those observations are not successful fills after five minutes. An unconditional fill time requires a convention for orders that never fill.

5.The median is multiplied by $e^{-0.4} \approx 0.6703$, a reduction of roughly 33%. Greater aggressiveness can speed execution while worsening the price paid or the selection of states in which a trade occurs. Neither the AFT coefficient nor the median by itself determines expected profit.

Lecture 6

Funding Liquidity and Fragile Markets

An investor may recognise that an asset is underpriced and still be unable to buy it. Knowledge of value is not the same as the capacity to finance a position. This distinction becomes especially important when the investors who normally absorb other people's trades have already suffered losses. Attractive buying opportunities then appear precisely when potential buyers have the least usable capital.

This lecture develops the funding mechanism in the original liquidity-fragility notes, following the framework of Brunnermeier and Pedersen [6]. We begin with the accounting and timing that make the mechanism work, solve the customers' and speculators' decisions, and then examine how margins respond to prices. A one-asset calculation separates two feedback channels: losses on existing positions and increases in the capital required for new positions. The final section connects the model to the broader liquidity literature.

The key calculation is the return earned by one additional margin dollar. Once that return is understood, the feedback between a dealer's balance sheet and the price needed to clear the market can be derived rather than assumed.

6.1 Two meanings of liquidity

Definition 6.1 (Market liquidity and funding liquidity)**.** *Market liquidity* is the ability to trade a specified quantity promptly at a small trading cost. Its dimensions include the spread, price impact, depth, and the speed with which the market recovers after a trade. *Funding liquidity* is the ability to obtain and retain the financing needed to support a position, including the ability to meet collateral requirements as they change.

Neither concept is simply trading volume. A distressed market can have heavy selling and high volume while offering very poor execution prices. Nor is funding liquidity simply the interest rate charged on a loan: a trader may be willing to pay interest but unable to supply the required collateral.

Suppose one share costs $p = 100$, and a lender will advance 90 against it. The investor must supply $m^+ = 10$ of equity per share. Here m^+ is a *dollar margin*: its units are currency per share. The corresponding proportional haircut is $h = m^+/p = 10\%$. With equity of 1,000 and no other constraint, the investor can hold 100 shares, worth 10,000. If the margin rises to 20 while equity remains unchanged, the maximum position falls to 50 shares.

The deposited margin is not, by itself, an execution fee or a permanent loss. It is capital committed to supporting a position. Its economic cost arises because the same capital cannot simultaneously finance another attractive trade. Short positions also require collateral: the proceeds of a short sale are not assumed to be freely reusable equity.

The three participants. Customers want to change their exposure to risk; speculators temporarily take the other side; financiers supply collateralised credit to the speculators. These are roles, not immutable institutional categories. A dealer can provide market liquidity in one transaction and demand it in another. The model isolates the particular chain in which customers need immediacy, speculators warehouse the resulting exposure, and financiers determine how much warehousing is possible.

6.2 The economy: dates, information, and balance sheets

There are J risky securities and dates $t = 0, 1, 2, 3$. Security j pays v_3^j at date 3. Interest is normalised to zero, and the net supply of traded risky claims is zero. Let

$$v_t^j = \mathbb{E}_t[v_3^j], \qquad v_{t+1}^j = v_t^j + \sigma_{t+1}^j \varepsilon_{t+1}^j, \qquad \sigma_{t+1}^j = \bar{\sigma}^j + \theta^j |v_t^j - v_{t-1}^j|. \tag{6.1}$$

The innovations ε_{t+1}^j are independent standard normal variables across dates and securities; $\bar{\sigma}^j > 0$ and $\theta^j \geq 0$. An initial volatility is specified at date 0. Customers and speculators observe current fundamental values; the financiers' information is specified separately below. Thus σ_{t+1}^j, measured in currency per share, is known at date t to an observer of fundamentals. Large fundamental innovations increase the next period's conditional volatility. This is an ARCH-type specification for the *standard deviation*; it should not be confused with the usual linear recursion for a conditional variance.

6.2.1 Customers arrive before their hedging needs can all be matched

Customers are indexed by $k = 0, 1, 2$. Customer k anticipates an endowment of $z^{j,k}$ shares of security j at date 3. Endowments offset:

$$\sum_{k=0}^{2} z^{j,k} = 0 \qquad \text{for every } j. \tag{6.2}$$

A positive endowment creates a natural desire to sell a claim on the security and remove exposure to its payoff; a negative endowment creates the opposite hedging need.

With probability $1 - a$, all customers can trade from date 0. With probability a, their access is sequential: customer 0 enters at date 0, customer 1 at date 1, and customer 2 at date 2. Once a customer arrives, that customer remains able to trade. Before arrival, the position in traded claims is constrained to zero. This is a restriction on market participation, not an assertion that the customer has no information about the future endowment.

Under sequential arrival, define the cumulative endowment of participating customers as

$$Z_t^j = \sum_{k=0}^{t} z^{j,k}, \qquad t = 0, 1, 2.$$

Although $Z_2^j = 0$, generally Z_0^j and Z_1^j are not zero. The eventual buyer may simply not yet be available when the early seller needs to trade. That timing mismatch is the demand for intermediation.

6.2.2 Positions, trades, and wealth are different objects

Let $y_t^{j,k}$ denote customer k's position in traded claims *after* trading at date t. A position is not an order: the net trade is the change in the position. The customer's total exposure to the next price change is $y_t^{j,k} + z^{j,k}$, because the anticipated endowment also changes in value.

Let B_t^k be cash after trading and mark the endowment at the same price as the traded claim. Economic wealth is

$$W_t^k = B_t^k + \sum_j p_t^j (y_t^{j,k} + z^{j,k}). \tag{6.3}$$

Self-financing trades change cash and traded holdings by equal and opposite amounts. Consequently,

$$W_{t+1}^k = W_t^k + \sum_j (p_{t+1}^j - p_t^j)(y_t^{j,k} + z^{j,k}), \qquad p_3^j = v_3^j. \tag{6.4}$$

The position indexed by t, not $t + 1$, earns the price change from t to $t + 1$. Initial cash and initial economic wealth must also be distinguished: with zero traded shares, $W_0^k = B_0^k + \sum_j p_0^j z^{j,k}$.

For example, if economic wealth is 400, the traded position is zero, the anticipated endowment is two shares, and the price rises by 1.50, economic wealth becomes 403. The extra three units are a revaluation of the endowment, not cash received from selling two shares. Customer preferences are

$$U(W_3^k) = -\exp(-\gamma W_3^k), \qquad \gamma > 0,$$

where γ has units inverse to currency.

6.2.3 Speculators' funding constraint

Speculators are competitive and risk-neutral. Write their signed positions as

$$x_t^j = x_t^{j+} - x_t^{j-}, \qquad x_t^{j+} = \max(x_t^j, 0), \quad x_t^{j-} = \max(-x_t^j, 0).$$

Their equity evolves according to

$$W_{t+1} = W_t + \sum_j (p_{t+1}^j - p_t^j)x_t^j + \eta_{t+1}, \tag{6.5}$$

where η_{t+1} is an independent wealth shock from other activities. We take its conditional mean to be zero when deriving expected trading profits. With positive equity, feasible positions satisfy

$$\sum_j \left(m_t^{j+}x_t^{j+} + m_t^{j-}x_t^{j-}\right) \leq W_t. \tag{6.6}$$

Every term in this inequality is measured in currency. In this stripped-down specification, margins are added security by security; no portfolio-netting benefit is assumed.

If a price fall reduces W_t, or a lender raises m_t, an existing position may no longer be financeable. The speculator must obtain additional equity or reduce the position. A margin shortfall is *not automatically balance-sheet insolvency*: a trader with positive net worth can still fail a collateral requirement. When $W_t \leq 0$, we instead enter a separate default regime in which the trader cannot initiate new positions. A nonnegative margin bill cannot satisfy an inequality with a negative right-hand side.

To represent the consequences of default, let the continuation payoff there be $\chi_t W_t$, with $\chi_t \geq 0$. Setting $\chi_t = 0$ gives limited liability; setting $\chi_t = 1$ counts a negative dollar of terminal wealth as a dollar loss. The date-1 linear optimisation below uses $\chi_2 = 1$. With limited liability at date 2, its objective would instead involve a positive-part payoff, and the same linear programme could not simply be reused.

6.2.4 Financiers and equilibrium

A financier's information is denoted $\mathcal{F}_t^F$. Informed financiers observe fundamentals and liquidity shocks. Uninformed financiers observe the price history and infer the underlying risks from it. Both may act rationally given their information. Information-poor margin setting need not be a behavioural mistake.

An equilibrium consists of prices, positions, and margins such that customers maximise expected utility subject to market access; speculators optimise subject to funding and default rules; financiers apply their specified risk rule; and traded positions clear:

$$x_t^j + \sum_{k=0}^{2} y_t^{j,k} = 0 \qquad \text{for every } j, t. \tag{6.7}$$

The endowments do not appear as an additional net supply in this equation. They are the source of hedging demands and already sum to zero.

6.3 Solve the customers' problem backwards

6.3.1 Date 2: why complete participation restores the benchmark price

At date 2 there is only one remaining payoff innovation, so conditional terminal wealth is normal. Maximising expected exponential utility is equivalent to maximising its certainty equivalent. If $q^{j,k} = y_2^{j,k} + z^{j,k}$ is total exposure, the customer maximises

$$W_2^k + \sum_j q^{j,k}(v_2^j - p_2^j) - \frac{\gamma}{2}\sum_j (q^{j,k})^2(\sigma_3^j)^2. \tag{6.8}$$

The first-order condition gives

$$y_2^{j,k} = \frac{v_2^j - p_2^j}{\gamma(\sigma_3^j)^2} - z^{j,k}. \tag{6.9}$$

This is a speculation demand minus an endowment hedge. At $p_2^j = v_2^j$, customer k chooses $y_2^{j,k} = -z^{j,k}$, eliminating total risky exposure. By (6.2), these trades clear among customers, and speculators can hold zero positions.

The offsetting-endowment assumption is essential. Without it, full participation would not by itself imply that prices equal conditional mean payoffs: somebody would have to bear aggregate risk. Here, any price below the fundamental creates positive aggregate customer demand as well as a desire by speculators to buy; a price above the fundamental reverses both signs. Hence the clearing benchmark is

$$p_2 = v_2, \qquad V_2^k(W_2^k) = -e^{-\gamma W_2^k}, \qquad J_2(W_2) = W_2. \tag{6.10}$$

If all customers arrive at date 0, the same offsetting hedge can be implemented immediately, and prices equal fundamentals at each trading date.

6.3.2 Date 1: the price concession for early customers

Under sequential arrival, only customers 0 and 1 are present at date 1. They know that at date 2 they can eliminate their remaining risk. Their relevant holding-period uncertainty is therefore $p_2 - p_1 = v_2 - p_1$, with variance $(\sigma_2^j)^2$, not the date-3 variance $(\sigma_3^j)^2$. Repeating the certainty-equivalent calculation yields

$$y_1^{j,k} = \frac{v_1^j - p_1^j}{\gamma(\sigma_2^j)^2} - z^{j,k}, \qquad k = 0, 1, \tag{6.11}$$

and

$$V_1^k = -\exp\left[-\gamma\left\{W_1^k + \sum_j \frac{(v_1^j - p_1^j)^2}{2\gamma(\sigma_2^j)^2}\right\}\right]. \tag{6.12}$$

To see the last term, substitute the optimum $q^* = (v - p)/(\gamma\sigma^2)$ into $q(v - p) - \gamma q^2\sigma^2/2$. The resulting gain is $(v - p)^2/(2\gamma\sigma^2)$.

For a security experiencing selling pressure, set

$$\ell = v_1 - p_1 \geq 0, \qquad Z = Z_1 > 0, \qquad b = \frac{2}{\gamma\sigma_2^2} > 0.$$

Market clearing then requires the speculators to hold

$$x_1 = Z - b\ell. \tag{6.13}$$

At the fundamental price, speculators must absorb Z shares. At a discount, the customers are willing to retain more of their own exposure, so the required dealer position falls. With no speculator capacity, the customer-only discount is $\ell = Z/b$. The model therefore explains both why capital improves liquidity and why a finite price concession can eventually clear a market with no active dealer.

6.3.3 Date 0: tomorrow's trading opportunities matter today

Customer 0 chooses y_0 to maximise $\mathbb{E}_0[V_1^0]$, with W_1^0 generated by (6.4). The squared mispricing term in (6.12) is now random. One cannot obtain the date-0 optimum merely by replacing date-1 variables with date-0 variables in (6.11): future liquidity affects the continuation value. This is why backward induction is more than a notational device. It identifies which future market opportunities the current decision must anticipate.

6.4 The value of a margin dollar

At date 1, speculators expect to close their positions at $p_2 = v_2$. Given prices and positive margins, define expected profits per share as

$$g_j^+ = v_1^j - p_1^j, \qquad g_j^- = p_1^j - v_1^j.$$

The date-1 problem is

$$J_1(W_1) = W_1 + \max_{x^{j+},x^{j-}\geq 0} \sum_j (g_j^+ x^{j+} + g_j^- x^{j-}) \quad \text{subject to (6.6).} \tag{6.14}$$

An independent mean-zero wealth shock does not change this expected-profit objective. The trader is competitive, so prices and margin schedules are taken as given when choosing a position.

Proposition 6.1 (Return on scarce capital)**.** *Suppose $W_1 > 0$, relevant margins are strictly positive, and the date-2 continuation payoff is linear in wealth. Define*

$$\lambda = \max\left\{0, \max_j \frac{v_1^j - p_1^j}{m_1^{j+}}, \max_j \frac{p_1^j - v_1^j}{m_1^{j-}}\right\}, \qquad \phi_1 = 1 + \lambda. \tag{6.15}$$

Then $J_1(W_1) = \phi_1 W_1$. If $\lambda > 0$, all available equity is used and every actively held direction earns λ units of expected profit per unit of margin.

Proof. For every feasible portfolio, $g_j^{\pm} \leq \lambda m_1^{j\pm}$, and therefore total expected trading profit is at most λW_1. Investing all equity in any direction attaining the maximum achieves this bound. Equivalently, attach the multiplier λ to the margin constraint. The optimality inequalities are $g_j^{\pm} - \lambda m_1^{j\pm} \leq 0$, with equality for a positive holding. The extra unit of initial equity is retained as wealth and earns λ additional profit, giving marginal continuation value $1 + \lambda$. □

The economic comparison is profit per *margin dollar*, not profit per share and not profit divided by the entire asset price. If one trade earns 4 per share and requires margin 20, while another earns 3 and requires margin 10, the second uses scarce capital more effectively: its ratio is 0.30 rather than 0.20.

At the optimum, the multiplier $\lambda = \phi_1 - 1$ is the shadow value of relaxing the capital constraint in units of extra trading profit. The total marginal value of initial wealth is ϕ_1. This difference of one matters: if $\lambda = 0.30$ and capital is 1,000, expected final wealth is 1,300, not 300. It is an expected value, not a risk-free contractual return.

6.4.1 From individual optimisation to equilibrium illiquidity

Define signed price dislocation and its magnitude by

$$\Lambda_t^j = p_t^j - v_t^j, \qquad \ell_t^j = |\Lambda_t^j|. \tag{6.16}$$

The magnitude is the model's measure of *market illiquidity*, not of liquidity. It is a price-level deviation in currency per share, not a bid–ask spread or an estimated price-impact slope.

For an actively intermediated security, the optimality condition becomes

$$\ell_1^j = m_1^{j,s}(\phi_1 - 1), \tag{6.17}$$

where $s = +$ for a long position absorbing selling pressure and $s = -$ for a short position absorbing buying pressure. For an inactive direction, the corresponding expected profit satisfies an inequality, $g_j^s \leq m_1^{j,s}(\phi_1 - 1)$, not necessarily equality. A speculator need not hold every mispriced asset.

Equation (6.17) is a joint equilibrium restriction. It does not say that ϕ_1 is an externally chosen number or that a margin increase leaves it unchanged. Customer demand, available equity, prices, and financiers' margin schedules must all be solved together. If a positive expected profit could be financed with zero margin and no other position limit, the price-taking linear programme would be unbounded. Such a proposed price cannot be a finite equilibrium of this benchmark.

6.5 How financiers set margins

6.5.1 The tail loss differs for long and short positions

A lender uses margin to cover a one-period adverse price move. For a small tail probability $0 < \pi < 1/2$, the risk requirements are

$$\Pr(-\Delta p_{t+1}^j > m_t^{j+} \mid \mathcal{F}_t^F) \leq \pi, \tag{6.18}$$

$$\Pr(\Delta p_{t+1}^j > m_t^{j-} \mid \mathcal{F}_t^F) \leq \pi. \tag{6.19}$$

A long position loses when price falls; a short position loses when it rises. Margins are chosen as the smallest nonnegative amounts satisfying these inequalities. Equality holds for an interior, continuously distributed loss quantile; a zero lower bound can make the inequality strict. These are collateral-loss rules for financing positions, not an assurance that all portfolio losses are bounded by margin.

6.5.2 Informed financiers: a discount provides a cushion

At date 1, $p_2 = v_2$. An informed financier therefore knows that

$$\Delta p_2^j = \sigma_2^j \varepsilon_2^j - \Lambda_1^j.$$

Writing Φ for the standard normal distribution function and $q_\pi = \Phi^{-1}(1-\pi) > 0$, the long-position tail probability is

$$\Pr(-\Delta p_2^j > m_1^{j+} \mid \mathcal{F}_1^F) = 1 - \Phi\left(\frac{m_1^{j+} - \Lambda_1^j}{\sigma_2^j}\right). \tag{6.20}$$

Solving the quantile conditions gives

$$m_1^{j+} = \max\{q_\pi \sigma_2^j + \Lambda_1^j, 0\}, \qquad m_1^{j-} = \max\{q_\pi \sigma_2^j - \Lambda_1^j, 0\}. \tag{6.21}$$

If the security is underpriced by ℓ, then $\Lambda = -\ell$, and the long margin is $\max\{q_\pi\sigma_2 - \ell, 0\}$. Expected price convergence offsets part of the adverse fundamental innovation. A larger discount consequently lowers the margin required to buy the undervalued asset. For an overpriced asset, the same cushioning logic lowers the short margin.

Example 6.1 (A correctly signed margin)**.** Let $v_1 = 100$, $p_1 = 97$, $\sigma_2 = 4$, and $\pi = 1\%$, so $q_\pi \simeq 2.326$. The long margin is $2.326(4) - 3 \simeq 6.30$; the short margin is $2.326(4) + 3 \simeq 12.30$. A further temporary discount to 96 lowers the long margin to about 5.30, holding fundamental volatility fixed. Buying below value is cushioned; shorting below value is not.

Call margins *stabilising* when greater illiquidity reduces the margin on the position that supplies liquidity. This conclusion requires knowledge of fundamental value and of sufficiently prompt convergence. It does not mean an informed lender must lower collateral after every negative return: genuine fundamental news can raise σ_2, and delayed convergence can leave substantial liquidation risk.

6.5.3 Uninformed financiers: price volatility can tighten financing

With only prices, the lender faces a filtering problem. The observed return combines fundamental news and changes in dislocation:

$$\Delta p_1^j = \Delta v_1^j + \Delta\Lambda_1^j.$$

If sequential arrival is believed to be very unlikely, the lender tends to attribute observed price changes to fundamentals. In the limiting benchmark $a \to 0$, the margin schedule approaches

$$m_1^j = q_\pi\left(\bar{\sigma}^j + \theta^j |\Delta p_1^j|\right), \tag{6.22}$$

for both long and short positions. This is a limiting result, not the exact filtering solution for every a. With $\theta^j > 0$, a liquidity-driven fall that makes the observed price move more negative raises perceived future volatility and hence margin. If it instead offsets an earlier positive move, its effect need not have the same sign.

Call margins *destabilising* on a selling-pressure branch when $dm^+/d\ell > 0$. The same increase in expected convergence profit that attracts a liquidity provider then also increases the capital required to enter the trade. The financier's individual effort to limit risk can reduce the market's aggregate ability to absorb sales.

6.6 Two spirals and a measure of amplification

The verbal sequence is simple: losses or tighter margins reduce financeable positions; customers must accept larger price concessions; those price changes can create further losses and higher margins. To identify when this sequence actually amplifies a shock, we now solve a local one-security illustration. This calculation is a teaching reduction, not a claim that the full multi-security equilibrium always has a closed form.

6.6.1 A clearing equation that contains both channels

Retain $x = Z - b\ell$ from (6.13). Consider a price decline due purely to a liquidity discount, holding the fundamental fixed. If the speculator carries $x_0 \geq 0$ shares into the repricing, equity becomes

$$W = w - x_0\ell,$$

where w collects initial equity, any independently specified wealth shock, and the value at the no-dislocation reference price. The old position x_0 determines the mark-to-market loss; the new position x determines the required margin. Confusing these positions would erase the loss channel.

On a branch where $W > 0$, $0 < x < Z$, and the constraint binds,

$$m(\ell)(Z - b\ell) + x_0\ell = w. \tag{6.23}$$

Let $m' = dm/d\ell$. Implicit differentiation gives

$$\frac{d\ell}{dw} = \frac{1}{m'x - bm + x_0} = -\frac{1}{bm}\frac{1}{1-G}, \qquad G = \frac{m'x + x_0}{bm}. \tag{6.24}$$

Here G is dimensionless. If there is neither an existing-position loss nor a margin response, $G = 0$, and a one-unit capital loss raises the discount by $1/(bm)$. With $0 < G < 1$, the increase is larger by the factor $1/(1-G)$.

- **Loss spiral:** $x_0 > 0$ makes a larger discount reduce equity on existing holdings. Even a fixed margin can then amplify the shock.
- **Margin spiral:** $m' > 0$ makes each remaining share consume more equity. This can amplify the shock even when the initial position is zero.

- **Cushioning:** $m' < 0$ offsets some of the loss feedback. It does not eliminate a sufficiently strong loss spiral automatically.

There is a precise local adjustment interpretation. Given a trial discount, the financeable position is $(w - x_0\ell)/m(\ell)$. The discount needed to clear the residual customer supply is

$$T(\ell) = \frac{1}{b}\left[Z - \frac{w - x_0\ell}{m(\ell)}\right].$$

At a fixed point, $T'(\ell) = G$. Under this particular iterative adjustment rule, local stability requires $|G| < 1$. On the destabilising branch with $G \geq 0$, the critical value is $G = 1$. As that value is approached from below, the comparative-static response in (6.24) becomes large. A statement about local iterative stability is not, by itself, a proof of global uniqueness or a description of real-time trading dynamics.

6.6.2 Fragility is more than a large percentage loss

An investor with little equity suffers a larger percentage loss from a fixed currency shock. That observation is useful intuition but is not a complete theory of market fragility. The stronger result comes from the feedback between the price that clears customers' trades and the capital available at that same price.

If the left-hand side of (6.23) is non-monotone, there can be multiple feasible clearing discounts. A turning point occurs when $bm = m'x + x_0$. A branch can disappear as w changes, forcing the equilibrium to move to a different branch or a boundary regime. Establishing a particular discontinuity requires checking those other equilibria and the admissible domain; a large derivative alone does not prove that a crash occurs. Conversely, with ample capital, speculators can absorb the full imbalance at $\ell = 0$, and a small change in wealth may have no effect on price at all.

Example 6.2 (An equity shock with both feedback channels)**.** Let fundamental value be 100 per share, customer pressure $Z = 100$ shares, and $b = 5$ shares per unit of price discount. Suppose the dealer initially holds $x_0 = 20$ shares, reference equity is $w = 800$, and the illustrative margin schedule is

$$m(\ell) = 10 + 0.2\ell.$$

The funding equation is

$$(10 + 0.2\ell)(100 - 5\ell) + 20\ell = 800, \quad \text{or} \quad \ell^2 + 10\ell - 200 = 0.$$

Its economically relevant root is $\ell = 10$. Thus price is 90, the dealer holds 50 shares, required margin is 12 per share, and equity after repricing is 600. The identity $50(12) = 600$

checks financing and clearing simultaneously. Expected convergence profit is $50(10) = 500$, so $\lambda = 10/12 = 5/6$, $\phi_1 = 11/6$, and expected date-2 wealth is $600 + 500 = 1{,}100$.

At this equilibrium,

$$G = \frac{0.2(50) + 20}{5(12)} = \frac{1}{2}, \qquad \frac{d\ell}{dw} = -\frac{1}{30}.$$

A capital loss of 30 therefore predicts an increase of about one in the discount. Solving exactly at $w = 770$ gives

$$\ell = \frac{-10 + \sqrt{1020}}{2} \simeq 10.969,$$

so price falls to about 89.031, the position falls to 45.156 shares, margin rises to 12.194, and equity falls to 550.626. The new equity loss, about 49.374, exceeds the external shock of 30 because the existing 20 shares also lose value. These numbers are a deliberately specified scenario, not an empirical calibration.

6.7 Commonality, flight to quality, and precautionary capital

6.7.1 One balance sheet links many securities

For two actively intermediated assets sharing the same capital pool,

$$\ell^j = m^j \lambda, \qquad \ell^k = m^k \lambda.$$

A rise in the common shadow cost λ increases both dislocations when margins are held fixed. Shocks originating in one market can therefore affect another even when their fundamental payoff innovations are independent. The connection is the intermediary's balance sheet, not necessarily cash-flow correlation.

This is a mechanism for *commonality in liquidity*. Chordia, Roll, and Subrahmanyam [7] document co-movement in measures such as spreads and depth. Their finding is an empirical fact to explain, not unique identification of the funding mechanism: correlated customer flows and information shocks can also create commonality.

If $m^H > m^L$, then on active branches

$$\ell^H - \ell^L = (m^H - m^L)\lambda.$$

When capital becomes scarce, high-margin securities experience a larger deterioration in liquidity. Since volatile securities often require larger margins, this produces a flight toward less capital-intensive liquidity provision. In this mechanism, lower-volatility assets generally have *lower*, not higher, margin requirements. "Flight to quality" describes a relative reallocation and a widening liquidity differential; it does not imply that every safer asset's price must rise in absolute terms.

6.7.2 Funding risk matters before today's constraint binds

At date 0 the speculator chooses a portfolio to maximise

$$\mathbb{E}_0[J_1(W_1, S_1)]\,, \tag{6.25}$$

where S_1 denotes the future state determining prices, margins, and trading opportunities. If wealth stays positive, this becomes $\mathbb{E}_0[\phi_1 W_1]$. One unit of wealth is particularly valuable in a state with a high ϕ_1, because it can finance unusually profitable intermediation at that date.

For an interior, marginal date-0 position with a slack current funding constraint, positive future wealth, and price-taking continuation opportunities, the first-order condition is

$$\mathbb{E}_0[\phi_1(p_1^j - p_0^j)] = 0, \quad \text{hence} \quad p_0^j = \mathbb{E}_0[p_1^j] + \frac{\mathrm{Cov}_0(\phi_1, p_1^j)}{\mathbb{E}_0[\phi_1]}. \tag{6.26}$$

A security that pays well when capital is scarce has a funding-hedging benefit. A position that loses precisely when margins rise has the opposite property. A trader can therefore retain a capital buffer even when the current margin inequality is slack. Risk-neutral terminal preferences do not imply indifference to the state in which financing capacity is available. If future default is possible or today's constraint binds, the appropriate continuation derivative or multiplier must be included rather than applying (6.26) unchanged.

6.8 Locating the model in the liquidity literature

The literature overview accompanying the original notes covers several related but distinct questions. Keeping their objects separate prevents a measure designed for one purpose from being given an interpretation it does not support.

6.8.1 Liquidity levels and expected returns

Amihud's measure [4] averages the ratio of an absolute daily return to daily currency trading volume:

$$\mathrm{ILLIQ}_{i,T} = \frac{1}{D_{i,T}} \sum_{d \in T} \frac{|r_{i,d}|}{\mathrm{DVOL}_{i,d}}. \tag{6.27}$$

Here $D_{i,T}$ is the number of usable positive-volume days in the measurement period. The statistic is a coarse price-response proxy, not a structural estimate of the causal effect of an extra trade. Larger expected trading costs can be associated with higher required future returns. An unexpected deterioration in liquidity can simultaneously lower today's price. There is no contradiction: a lower entry price is one way to deliver a higher expected subsequent return, holding expected cash flows fixed.

Acharya and Pedersen [2] distinguish the expected level of trading costs from risk associated with their fluctuations. Let r_i be an asset return and c_i its liquidation cost expressed as a fraction of initial price. In the one-period notation, their liquidity-adjusted CAPM can be written

$$\mathbb{E}[r_i - c_i] - r_f = \beta_i^{\text{net}}\big(\mathbb{E}[r_M - c_M] - r_f\big), \qquad \beta_i^{\text{net}} = \frac{\operatorname{Cov}(r_i - c_i, r_M - c_M)}{\operatorname{Var}(r_M - c_M)}. \tag{6.28}$$

Expanding the numerator yields four covariance terms:

$$\operatorname{Cov}(r_i, r_M) + \operatorname{Cov}(c_i, c_M) - \operatorname{Cov}(r_i, c_M) - \operatorname{Cov}(c_i, r_M).$$

Beyond ordinary market-return exposure, the three liquidity channels concern costs rising together, poor asset returns when market liquidity deteriorates, and high asset costs when market returns are low. Expected cost compensation and covariance-based risk compensation are conceptually different.

6.8.2 Information, price impact, and participation

Kyle's strategic insider model explains how an informed trader trades against a price-impact schedule, spreading information revelation through order flow. Glosten–Milgrom explains adverse-selection spreads in sequential trade. Neither mechanism needs the funding constraint developed here; in a richer market they can operate alongside it.

Grossman–Stiglitz concerns the incentive to acquire costly information: a fully revealing price can eliminate the private reward that supports information production. It is not the no-trade theorem. Nor does the label "competitive rational expectations" alone imply multiple equilibria. Coordination-based multiplicity requires additional strategic complementarities or participation mechanisms, rather than merely incomplete information.

Foster and Viswanathan's interday model [13] studies how information asymmetry and the timing of liquidity trading interact. Hasbrouck [20] estimates the dynamic interaction of trades and quote revisions; the lasting response to an order-flow innovation is the relevant information-effect object. A regression on contemporaneous raw volume alone does not make the same distinction. Glosten and Harris [16] separate persistent information effects from transitory spread components. Their contribution is not adequately described as an AR(1) model for the price level.

6.8.3 From sequential-trade models to estimated trading intensity

Easley, Kiefer, O'Hara, and Paperman [11] estimate a probability-of-informed-trading model from buy and sell arrival counts. Its latent parameters describe information events and the intensity of trading. In the basic notation,

$$\mathrm{PIN} = \frac{\alpha\mu}{\alpha\mu + \epsilon_b + \epsilon_s},$$

where α is the probability of an information event, μ the informed-arrival intensity conditional on an event, and ϵ_b, ϵ_s the uninformed buy and sell intensities. This is a model-implied fraction, not an observed label identifying which individual trader has private information.

The volume-synchronised measure VPIN [12] instead uses estimated buy–sell imbalance in equal-volume buckets. It is a proxy for flow toxicity, sensitive to trade classification and bucket construction; it is not interchangeable with a structural PIN estimate. More generally, econometric models of sequential trade must account for irregular arrival times, changing activity, and latent trading regimes. Those issues connect information models to actual transaction data without turning every imbalance into evidence of informed trading.

6.8.4 Execution risk, forced sales, and constrained arbitrage

Optimal execution asks how to trade a given position when speed increases market impact but delay leaves price risk. In the standard constant-parameter, zero-drift Almgren–Chriss benchmark [3], a continuous-time illustration minimises

$$\int_0^T \left[\eta \dot{q}(t)^2 + \rho\sigma^2 q(t)^2\right] dt, \qquad q(0) = Q, \quad q(T) = 0,$$

where q is remaining inventory, $Q > 0$, $T > 0$, $\eta > 0$ measures temporary impact, $\sigma > 0$ is price volatility, and $\rho > 0$ weights risk. The first-order equation is $\ddot{q} = \kappa^2 q$, with $\kappa^2 = \rho\sigma^2/\eta$, giving

$$q(t) = Q\frac{\sinh(\kappa(T-t))}{\sinh(\kappa T)}.$$

The selling rate is front-loaded and decreasing, not generically U-shaped. In the zero-risk limit $\kappa \to 0$, the formula tends to $q(t) = Q(1 - t/T)$, so selling is uniform. Intraday liquidity patterns or other constraints can change the schedule. Hisata and Yamai [23] connect liquidation strategies and market impact to liquidity-adjusted value at risk. Execution-cost risk and a financier's collateral quantile are related risk-management problems, but they are not the same constraint.

Forced liquidation adds a strategic issue: other traders may anticipate the sale. Brunnermeier and Pedersen's predatory-trading model [5] studies how trading around another participant's need to liquidate can worsen price pressure. The dynamics of constrained arbitrage also include recovery: after capital losses widen discrepancies, higher subsequent expected profits can gradually rebuild arbitrage capital, as in Gromb and Vayanos [18]. A funding model should not be reduced to the claim that arbitrageurs always sell early or always stabilise markets.

6.8.5 System-wide and over-the-counter measurement

Liquidity is multidimensional in bonds as well as equities. A spread measures the cost of crossing quotes; depth concerns quantity available; price-impact statistics concern the response to trading; turnover measures activity. These can move differently. Sparse over-the-counter transactions make stale prices, trade matching, corrections, and repeated reports especially consequential. Dick-Nielsen's study of TRACE [10] highlights why the transaction record must be cleaned before a bond-liquidity statistic is interpreted. A corporate-bond yield spread, moreover, is not a pure liquidity spread: it can also contain credit and other risk components.

A unified reading strategy. For every model or estimate, ask: whose trading need creates the imbalance, who is expected to absorb it, what limits that participant, what is the precise liquidity measure, and over what horizon is it defined? The answers distinguish adverse selection from funding scarcity, and distinguish both from mechanical execution costs. They also explain why a single crisis can activate several mechanisms at once.

6.9 Exercises

Exercise 6.1 (Endowment hedging and clearing)**.** At date 2, three customers have endowments 12, −5, −7 shares. Suppose $\gamma = 0.02$, $\sigma_3 = 5$, and $v_2 = 100$. Find their positions at $p_2 = 100$. What is aggregate customer demand if $p_2 = 99$, and why cannot that price clear with a profit-maximising risk-neutral speculator? Finally, explain what changes if the last endowment is −6 instead of −7.

Exercise 6.2 (Capital allocation)**.** Two underpriced assets have expected convergence gains 2 and 3 per share and long margins 5 and 10. A price-taking speculator has equity 500. Find the optimal allocation, λ, ϕ_1, and expected date-2 wealth under linear terminal payoffs. Explain why the inactive asset need not satisfy $\ell = m\lambda$ with equality. Are these fixed prices necessarily general-equilibrium prices?

Exercise 6.3 (Tail probabilities and direction)**.** Let $v_1 = 50$, $p_1 = 52$, and $\sigma_2 = 3$. With $\pi = 5\%$, use $q_\pi = 1.645$ to calculate informed long and short margins. If the price rises to 53 without a change in fundamental value or volatility, which liquidity-supplying position becomes cheaper to fund? Explain the role of the zero floor if mispricing becomes very large.

Exercise 6.4 (Disentangling the two feedback channels)**.** In Example 6.2, recalculate the equilibrium with (i) no existing position, $x_0 = 0$, keeping the variable margin; (ii) constant margin 10, retaining $x_0 = 20$; and (iii) neither feedback channel. Keep $Z = 100$, $b = 5$, and $w = 800$. Compare the discounts with the original discount 10. Why is a comparison of equilibrium discounts not the same calculation as holding the equilibrium fixed and decomposing G?

Exercise 6.5 (A liquidity statistic is not a mechanism)**.** A researcher finds that average absolute daily returns divided by trading volume rise during a market decline. Does this establish a margin spiral? Identify two additional kinds of observations that would help evaluate that mechanism. Explain why high expected future returns and negative contemporaneous returns can both accompany an increase in illiquidity.

6.10 Solutions and discussion

Exercise 6.1. At the fundamental price the positions are $-12, 5, 7$, and each total exposure is zero. Since $\gamma\sigma_3^2 = 0.5$, a price of 99 adds two shares of speculation demand per customer. Aggregate customer demand is therefore six shares. A speculator also wants to buy an underpriced asset, not supply the negative six-share position needed to clear. With endowments $12, -5, -6$, aggregate endowment is one share, and the exact-hedging argument at $p = v$ no longer clears. The risk-bearing allocation must be solved afresh; the full-participation conclusion relied on zero aggregate endowment.

Exercise 6.2. The profit-per-margin ratios are $2/5 = 0.4$ and $3/10 = 0.3$. The speculator buys 100 shares of the first asset and none of the second, giving $\lambda = 0.4$, $\phi_1 = 1.4$, and expected wealth 700. The second asset satisfies $3 < 10(0.4) = 4$. Its return per unit of scarce capital is insufficient to attract this trader. These are individual demands at given prices; customer supply and margin setting must also clear before the prices can be called an equilibrium.

Exercise 6.3. Signed dislocation is $\Lambda = 2$. Long margin is $1.645(3) + 2 = 6.935$, and short margin is $1.645(3) - 2 = 2.935$. At price 53, short margin falls to 1.935 while long margin rises to 7.935. Shorting is the liquidity-supplying direction for an overpriced security. The nonnegative floor prevents a negative collateral requirement. A zero-margin profitable direction would make the unconstrained-size, risk-neutral optimisation unbounded; a finite equilibrium must preclude that configuration or include another binding restriction.

Exercise 6.4. With no old position, the equation is $(10 + 0.2\ell)(100 - 5\ell) = 800$, giving

$$\ell = \frac{-30 + \sqrt{1700}}{2} \simeq 5.616.$$

With only the loss channel, $10(100 - 5\ell) + 20\ell = 800$, so $\ell = 20/3 \simeq 6.667$. With neither channel, $10(100 - 5\ell) = 800$, so $\ell = 4$. Both channels together yield 10. Removing a channel changes the equilibrium position and margin as well as the discount. By contrast, the decomposition $G = m'x/(bm) + x_0/(bm)$ measures local feedback contributions at one specified equilibrium. These are complementary comparisons, not interchangeable decompositions.

Exercise 6.5. No. The statistic can rise because of fundamental volatility, lower volume, adverse selection, or combinations of these, and is not an observed collateral constraint. Useful evidence would include the timing and magnitude of actual margin changes, and intermediaries' equity, existing positions, or forced position reductions. To support a causal amplification interpretation, one would also need to distinguish the initiating shock from the endogenous response. When illiquidity is expected to persist, a lower current price can compensate future holders through a higher subsequent expected return. The contemporaneous price loss and the forward-looking premium refer to different horizons.

The central lesson. Liquidity is partly a property of the institutions able to take the other side of a trade. A security's fundamental value can remain unchanged while its transaction price moves sharply because the marginal liquidity provider has lost equity or faces tighter collateral terms. The equation $\ell = m(\phi - 1)$ records the allocation of scarce capital; the funding-clearing equation explains how that scarcity and the price concession can reinforce one another.

Supplement A

Mathematical supplement: fractional kernels and rough volatility

How should a shock from yesterday affect volatility today, and how should a shock from a moment ago differ? A convolution makes these questions precise by assigning weights to past inputs. Fractional kernels provide one family of weights; exponential kernels provide another. Their difference explains why a model can be easy to update recursively yet still struggle to reproduce behavior across many short time scales.

We begin by distinguishing dependence from path roughness, then build fractional integrals and their stochastic counterparts. The final step shows how a mixture of exponential factors can approximate a fractional kernel while allowing a finite-state numerical update. The aim is to understand the memory mechanism before adding the rest of an option-pricing model.

A.1 Brownian motion, fractional Brownian motion and regularity

A standard Brownian motion W starts at zero, has continuous paths, and has independent Gaussian increments satisfying $W_t - W_s \sim N(0, t-s)$ for $t > s$. It is both Markov and a martingale with respect to its natural filtration. These are separate properties. The Markov property concerns how the conditional distribution of the future depends on the past; the martingale property concerns only the conditional mean. Neither is a definition of continuity or smoothness.

A standard fractional Brownian motion B^H, with $0 < H < 1$, is a centered continuous Gaussian process with covariance

$$\mathbb{E}[B_t^H B_s^H] = \frac{1}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right).$$

Its increments have variance $|t-s|^{2H}$. When $H = 1/2$ this is Brownian motion. For other values of H, increments are generally dependent, and the process is generally neither Markov nor a martingale in its own natural filtration.

The parameter H controls the scaling of local fluctuations. Sample paths are almost surely locally Hölder continuous of every order strictly below H, and are not differentiable for any $H \in (0,1)$. A smaller H means lower local regularity. It does *not* mean that the process jumps: fractional Brownian motion has continuous paths even when H is very small. A model with discontinuous jumps requires another construction.

For equal-length adjacent increments,

$$\operatorname{Corr}(B^H_{t+h} - B^H_t,\ B^H_{t+2h} - B^H_{t+h}) = 2^{2H-1} - 1.$$

The correlation is negative for $H < 1/2$, zero at $H = 1/2$, and positive for $H > 1/2$. Thus local roughness and the familiar positive long-range dependence of fractional Gaussian noise are not synonyms. The latter arises for $H > 1/2$; the rough-volatility regime usually refers to $H < 1/2$.

Gatheral, Jaisson and Rosenbaum [14] motivate rough-volatility models using the short-scale behavior of estimated log-volatility, with a small Hurst exponent in their analysis. Such an empirical estimate describes an asset sample, estimator and observation scale. It is not a universal constant that can be assigned to every financial time series, and volatility-measurement error must be considered when interpreting short-scale variation.

A.2 Convolution: weighting the past

The covariance of fractional Brownian motion describes a process directly. A kernel offers a complementary construction: specify how much influence each past input retains, and integrate those influences to obtain the present state.

For suitable functions f, g on $[0, \infty)$, their causal convolution is

$$(f * g)(t) = \int_0^t f(t-s)g(s)\,ds.$$

The variable s is the date of an input; $t - s$ is its age. If $g(s)$ records the input, then $f(t - s)$ specifies its weight at the current date. For an exponential kernel $f(u) = e^{-\kappa u}$ with $\kappa > 0$, older inputs receive exponentially smaller weights. In the rough-volatility regime considered below, power-kernel weights also decay with age, but at a power-law rate. A general power kernel need not decay: its exponent determines whether weights fall, remain constant, or grow.

To generalize an ordinary integral while preserving consistent normalization across orders, we use the gamma function. For $\alpha > 0$, it is defined by

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}e^{-u}\,du.$$

It extends the factorial identity through $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. In particular, $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$. Define

$$K_\alpha(t) = \frac{t^{\alpha-1}}{\Gamma(\alpha)}, \qquad t > 0.$$

The Riemann–Liouville fractional integral of order α is

$$(I^\alpha f)(t) = (K_\alpha * f)(t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s)\, ds. \tag{A.1}$$

The operator acts on the entire input function f: changing the input changes the weighted history. At $\alpha = 1$, equation (A.1) reduces to the ordinary integral of f. At $\alpha = 1/2$, the weight is $(t-s)^{-1/2}/\sqrt{\pi}$, singular but integrable at $s = t$.

For the constant input $f(s) = 1$,

$$(I^\alpha 1)(t) = \frac{t^\alpha}{\Gamma(\alpha+1)}.$$

Thus $I^{1/2}1 = 2\sqrt{t/\pi}$, while $I^1 1 = t$. More generally, for $\beta > -1$,

$$I^\alpha(t^\beta) = \frac{\Gamma(\beta+1)}{\Gamma(\alpha+\beta+1)} t^{\alpha+\beta}.$$

To derive this expression, substitute $s = tu$ in the integral. The remaining factor is the beta integral

$$B(x,y) = \int_0^1 u^{x-1}(1-u)^{y-1}\, du = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \qquad x, y > 0.$$

Taking $x = \beta + 1$ and $y = \alpha$ gives the stated coefficient after division by $\Gamma(\alpha)$. For positive orders α, ρ and suitable integrable inputs, the same convolution identity gives $I^\alpha I^\rho = I^{\alpha+\rho}$. Applying two half-integrals therefore produces one ordinary integral. This composition rule explains what a fractional order means operationally.

A.3 Stochastic convolution and the Hurst exponent

To make the weighted history random, replace the deterministic input by Brownian shocks. The resulting Gaussian process is

$$Y_t = \int_0^t K_\alpha(t-s)\, dW_s, \qquad \alpha = H + \frac{1}{2}.$$

Here the integration is with respect to Brownian motion, rather than ds. For a fixed t, the Itô integral exists in mean square when $\int_0^t K_\alpha(u)^2 du < \infty$, requiring $\alpha > 1/2$, or $H > 0$. Its variance follows from the Itô isometry:

$$\mathrm{Var}(Y_t) = \frac{t^{2\alpha-1}}{(2\alpha-1)\Gamma(\alpha)^2} = \frac{t^{2H}}{2H\Gamma(H+1/2)^2}.$$

This one-sided Volterra process has the roughness scaling associated with H. It is not, in general, the standard fractional Brownian motion defined above: starting the convolution at zero changes the covariance and destroys stationary increments. The kernel representation must therefore be named precisely.

In the rough regime $0 < H < 1/2$, one has $1/2 < \alpha < 1$. The kernel is large near zero and decays as a power of elapsed time. A volatility model may place such a stochastic convolution inside a positive transformation for volatility or within a carefully specified variance equation. The kernel alone does not ensure positivity of a variance process.

A.4 Why mixtures of exponentials are useful

The power kernel weights the full past, whereas an exponential kernel can be updated from a single current state. A mixture of exponentials connects the desired memory pattern to that computational advantage.

For $0 < \alpha < 1$, a gamma-integral calculation gives

$$K_\alpha(t) = \frac{1}{\Gamma(\alpha)\Gamma(1-\alpha)} \int_0^\infty e^{-xt} x^{-\alpha}\, dx.$$

The power kernel is therefore a continuous mixture of exponential kernels. A quadrature approximation replaces this integral by

$$K_\alpha(t) \approx \sum_{j=1}^{N} w_j e^{-\kappa_j t}, \qquad w_j > 0, \quad \kappa_j > 0.$$

For the Gaussian convolution, define factors

$$Y_t^{(j)} = \int_0^t e^{-\kappa_j (t-s)}\, dW_s, \qquad dY_t^{(j)} = -\kappa_j Y_t^{(j)}\, dt + dW_t.$$

Then $Y_t \approx \sum_j w_j Y_t^{(j)}$. The vector of factors is Markov, so a numerical procedure can update a finite state rather than retain the entire path. The factors here share the same Brownian driver; treating them as independent would change the covariance of the approximation.

Abi Jaber and El Euch [1] develop multifactor approximations for rough-volatility models. The elementary derivation above explains the computational idea. A fixed finite sum of exponentials cannot reproduce the singular kernel exactly at zero. Indeed, its weighted factor sum is an Itô process with Brownian coefficient $\sum_j w_j$, so its arbitrarily small-scale regularity is Brownian rather than exactly that of the target rough process. The approximation targets a specified range of resolved time scales. The number and placement of factors must be evaluated over that range, and a good kernel fit does not by itself establish a good option-price approximation without the surrounding model and an error analysis.

A.5 Exercises and solutions

Exercise A.1 (Rough does not mean discontinuous)**.** Compute the correlation of adjacent equal-length increments for $H = 0.1, 0.5$ and 0.75. Does any of these values imply jumps?

Solution. The correlations are approximately -0.4257, 0 and 0.4142. None implies jumps; all three fractional Brownian motions are continuous. The change concerns increment dependence and local regularity.

Exercise A.2 (A fractional integral)**.** Calculate $(I^{1/2}f)(t)$ for $f(t) = t$. Verify the ordinary-integral answer by applying another half-integral.

Solution. The power rule gives $(I^{1/2}f)(t) = 4t^{3/2}/(3\sqrt{\pi})$. Applying the rule again gives $t^2/2$, equal to $\int_0^t s\,ds$.

Exercise A.3 (Keeping the driver fixed)**.** For two exponential factors driven by the same Brownian motion, show that

$$\mathrm{Cov}(Y_t^{(i)}, Y_t^{(j)}) = \frac{1 - e^{-(\kappa_i+\kappa_j)t}}{\kappa_i + \kappa_j}.$$

What would change with independent drivers?

Solution. Apply the cross version of the Itô isometry to the two integrands. With independent drivers the covariance would be zero for $i \neq j$, eliminating the cross terms from the variance of the weighted sum and changing the intended model.

Guide to the original notes and further reading

Nine documents underpin this volume: six principal notes set the lecture sequence; the shorter Kyle note is integrated into Lecture 4, the liquidity presentation supplies a literature map, and the rough-volatility note forms Supplement A. Counts below refer to the originals, not these expanded lectures.

Original document	Pages	Principal destination
Trading Mechanisms	11	Lecture 1: trading rules, adverse selection and book dynamics.
Rational Expectations	12	Lecture 2: portfolio choice, information and noisy equilibrium.
Competitive Equilibrium Models	11	Lecture 3: transaction prices, time series and information.
Kyle Games	14	Lecture 4: informed trading and strategic disclosure.
Sequential Kyle Games	2	Lecture 4: overlapping single-auction derivation.
Limit Orders	4	Lecture 5: execution times and survival analysis.
Liquidity Fragility	8	Lecture 6: funding constraints, margins and feedback.
Liquidity	8 slides	Lecture 6: liquidity literature and measurement.
Rough Volatility Models	2	Supplement A: regularity and fractional kernels.

Reading paths

For the relationship between information and prices, begin with de Jong and Rindi, Grossman–Stiglitz, and Kyle. Hasbrouck develops the empirical price-process perspective; Chapters 3, 4 and 8 of [21] correspond to the Roll, univariate time-series and generalized-Roll material in Lecture 3. For execution, Lo, MacKinlay and Zhang provide the central connection to observed order lifetimes. Brunnermeier and Pedersen supply the funding-liquidity framework; the other liquidity papers place its mechanism alongside trading-cost risk, adverse selection and constrained arbitrage. Gatheral, Jaisson and Rosenbaum, followed by Abi Jaber and El Euch, connect the mathematical supplement to rough-volatility research.

Some source fragments were exploratory rather than complete models. The financial-innovation discussion therefore preserves its economic question without inventing an equilibrium. Elsewhere, assumptions are made explicit and algebraic or terminological slips corrected. References credit original contributions; worked examples and explanatory calculations belong to this exposition.

Bibliography

[1] Abi Jaber, E., and O. El Euch (2019). Multifactor approximation of rough volatility models. *SIAM Journal on Financial Mathematics*, 10(2), 309–349. doi:10.1137/18M1170236.

[2] Acharya, V. V., and L. H. Pedersen (2005). Asset pricing with liquidity risk. *Journal of Financial Economics*, 77(2), 375–410. doi:10.1016/j.jfineco.2004.06.007.

[3] Almgren, R., and N. Chriss (2001). Optimal execution of portfolio transactions. *Journal of Risk*, 3(2), 5–39. doi:10.21314/JOR.2001.041.

[4] Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, 5(1), 31–56. doi:10.1016/S1386-4181(01)00024-6.

[5] Brunnermeier, M. K., and L. H. Pedersen (2005). Predatory trading. *Journal of Finance*, 60(4), 1825–1863. doi:10.1111/j.1540-6261.2005.00781.x.

[6] Brunnermeier, M. K., and L. H. Pedersen (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, 22(6), 2201–2238. doi:10.1093/rfs/hhn098.

[7] Chordia, T., R. Roll, and A. Subrahmanyam (2000). Commonality in liquidity. *Journal of Financial Economics*, 56(1), 3–28. doi:10.1016/S0304-405X(99)00057-4.

[8] Cont, R., S. Stoikov, and R. Talreja (2010). A stochastic model for order book dynamics. *Operations Research*, 58(3), 549–563. doi:10.1287/opre.1090.0780.

[9] de Jong, F., and B. Rindi (2009). *The Microstructure of Financial Markets*. Cambridge University Press. Chapter 2, "Financial market equilibrium," 32–51. doi:10.1017/CBO9780511818547.004.

[10] Dick-Nielsen, J. (2009). Liquidity biases in TRACE. *Journal of Fixed Income*, 19(2), 43–55. doi:10.3905/jfi.2009.19.2.043.

[11] Easley, D., N. M. Kiefer, M. O'Hara, and J. B. Paperman (1996). Liquidity, information, and infrequently traded stocks. *Journal of Finance*, 51(4), 1405–1436. doi:10.1111/j.1540-6261.1996.tb04074.x.

[12] Easley, D., M. M. López de Prado, and M. O'Hara (2012). Flow toxicity and liquidity in a high-frequency world. *Review of Financial Studies*, 25(5), 1457–1493. doi:10.1093/rfs/hhs053.

[13] Foster, F. D., and S. Viswanathan (1990). A theory of the interday variations in volume, variance, and trading costs in securities markets. *Review of Financial Studies*, 3(4), 593–624. doi:10.1093/rfs/3.4.593.

[14] Gatheral, J., T. Jaisson, and M. Rosenbaum (2018). Volatility is rough. *Quantitative Finance*, 18(6), 933–949. doi:10.1080/14697688.2017.1393551.

[15] Glebkin, S., N. Gondhi, and J. C.-F. Kuong (2021). Funding constraints and informational efficiency. *Review of Financial Studies*, 34(9), 4269–4322. doi:10.1093/rfs/hhaa124.

[16] Glosten, L. R., and L. E. Harris (1988). Estimating the components of the bid/ask spread. *Journal of Financial Economics*, 21(1), 123–142. doi:10.1016/0304-405X(88)90034-7.

[17] Glosten, L. R., and P. R. Milgrom (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1), 71–100. doi:10.1016/0304-405X(85)90044-3.

[18] Gromb, D., and D. Vayanos (2018). The dynamics of financially constrained arbitrage. *Journal of Finance*, 73(4), 1713–1750. doi:10.1111/jofi.12689.

[19] Grossman, S. J., and J. E. Stiglitz (1980). On the impossibility of informationally efficient markets. *American Economic Review*, 70(3), 393–408. American Economic Association, full text.

[20] Hasbrouck, J. (1991). Measuring the information content of stock trades. *Journal of Finance*, 46(1), 179–207. doi:10.1111/j.1540-6261.1991.tb03749.x.

[21] Hasbrouck, J. (2007). *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press. doi:10.1093/oso/9780195301649.001.0001.

[22] Hasbrouck, J. (2007). *Empirical Market Microstructure: Corrigenda and Addenda*. Author's corrections, updated 18 June. Author's corrections, full text.

[23] Hisata, Y., and Y. Yamai (2000). Research toward the practical application of liquidity risk evaluation methods. *Monetary and Economic Studies*, 18(2), 83–127. Bank of Japan, full text.

[24] Huddart, S., J. S. Hughes, and C. B. Levine (2001). Public disclosure and dissimulation of insider trades. *Econometrica*, 69(3), 665–681. doi:10.1111/1468-0262.00209.

[25] Kühn, C., and C. Lorenz (2025). Insider trading in discrete time Kyle games. *Mathematics and Financial Economics*, 19(1), 39–66. Published online in 2024. doi:10.1007/s11579-024-00376-w.

[26] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6), 1315–1335. doi:10.2307/1913210.

[27] Kyle, A. S. (1989). Informed speculation with imperfect competition. *Review of Economic Studies*, 56(3), 317–355. doi:10.2307/2297551.

[28] Kyle, A. S., A. A. Obizhaeva, and Y. Wang (2023). Beliefs aggregation and return predictability. *Journal of Finance*, 78(1), 427–486. doi:10.1111/jofi.13195.

[29] Lee, C. M. C., and M. J. Ready (1991). Inferring trade direction from intraday data. *Journal of Finance*, 46(2), 733–746. doi:10.1111/j.1540-6261.1991.tb02683.x.

[30] Lo, A. W., A. C. MacKinlay, and J. Zhang (2002). Econometric models of limit-order executions. *Journal of Financial Economics*, 65(1), 31–71. doi:10.1016/S0304-405X(02)00134-4.

[31] Roll, R. (1984). A simple implicit measure of the effective bid–ask spread in an efficient market. *Journal of Finance*, 39(4), 1127–1139. doi:10.1111/j.1540-6261.1984.tb03897.x.

[32] Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3), 425–442. doi:10.1111/j.1540-6261.1964.tb02865.x.

[33] Subrahmanyam, A. (1991). Risk aversion, market liquidity, and price efficiency. *Review of Financial Studies*, 4(3), 417–441. doi:10.1093/rfs/4.3.417.

[34] U.S. Securities and Exchange Commission (2024). Officers, directors and 10% shareholders. Small-business guidance, updated 6 June. SEC guidance; Investor bulletin: Insider transactions and Forms 3, 4, and 5. Accessed 9 September 2026.