Mechanics of Dynamic Markets

Aryan Ayyar

December 29, 2025

Contents

Chapter 1

Introduction

This book develops a rigorous yet intuitive framework for understanding sequential Kyle games, a class of models in financial market microstructure that explain how asymmetric information affects price formation over time. We begin by reviewing the foundational Kyle (1985) model, and then progressively introduce extensions involving multi-period settings, partially informed traders, information leakage, stochastic signals, and Bayesian learning by market makers.

1.1 Overview

In his book "Elements of Pure Economics", Leon Walsrus established the conceptual framework of a general equilibrium. In so far as the investor is concerned, market prices play two important roles, namely allocation of scarce resources and being vehicles of information. It is today well known that economics equilibrium is a system-wide phenomenon and is not isolated to individual markets.Arrow and Debreu (1954) provided the first conceptual proof of the existence of a general equilibrium. Later, as it were to be, Debreu(1959) werre to present the classic Arrow-Debrew framework with not just unparalleled mathematical rigour but with clarity and generality.

What was revolutionary was the 1980 paper by Sanford Grossman and Joseph Stiglitz "On the Impossibility of Informationally Efficient Markets" . This challenged the fundamental assumption of costless, symmetric information in the Arrow-Debreu general equilibrium and showed how incorporating the cost of information may lead to profound paradoxes. Formally, they proved a fundamental impossibility theorem which states that "perfectly informationally efficient markets are impossible if information is costly to acquire". Hence, they introduced the concept of a Rational Expectations Equilibrium. A Rational Expectations Equilibrium

is a state in which, once all market participants have observed the equilibrium price p^*, no one has an incentive to revise their portfolio choice. In this equilibrium, all agents agree that p^* is optimal given their information set and that further adjustments would not improve expected payoff.This directly contrasts with a Walrasian equilibrium: a price decline not only clears markets (the Walrasian effect) but also reduces perceived fundamental value (the REE effect).[1]

1.2 Financial Market Equilibrium

To illustrate this paradox, let's consider a representative agent endowed with I shares of a risky asset and I_f units of a risk-free asset. The risk-free asset yields a gross return $1 + r_f$, while the risky asset pays a random payoff F at time T. If the agent demands X units of the risky asset at price p, then initial wealth at $t = 0$ is

$$W_0 \;=\; I\,p \;+\; I_f. \tag{1.1}$$

At time T, terminal wealth is

$$W_t \;=\; (X + I)\,F \;+\; (I_f - Xp)\,(1 + r_f). \tag{1}$$

The agent maximizes expected utility $U(W_t)$ of terminal wealth, with $U'(W_t) = \frac{\mathrm{d}U}{\mathrm{d}W_t}$. The first-order condition for optimal demand X is

$$\mathbb{E}\big[U'(w)\,(F - p\,(1 + r_f))\big] \;=\; 0.$$

Using $\mathbb{E}[AB] = \mathbb{E}[A]\,\mathbb{E}[B] + \mathrm{Cov}(A, B)$ and Stein's lemma yields

$$\mathbb{E}\big[U'(w)\big]\,\mathbb{E}\big[F - p\,(1 + r_f)\big] + \mathbb{E}\big[U''(w)\big]\,(I + X)\,\mathrm{Var}(F) \;=\; 0. \tag{2}$$

Rearranging (2) gives the equilibrium price:

$$p \;=\; \frac{1}{1 + r_f}\left(\frac{\mathbb{E}\big[U''(w)\big]\,(I + X)\,\mathrm{Var}(F)}{\mathbb{E}\big[U'(w)\big]} + \mathbb{E}[F]\right). \tag{3}$$

Under CARA utility $U(W_t) = -e^{-A\,W_t}$, one has $U'(W_t) = -A\,U(W_t)$ and $U''(W_t) =$

[1] Kyle (1989) introduced imperfect competition among informed traders, demonstrating that prices reveal at most half of their private information, so even risk-neutral informed agents trade less aggressively and an REE exists.

$A^2U(W_t)$, so (3) simplifies to

$$p \;=\; \frac{1}{1+r_f}\Big(A\,(I+X)\,\mathrm{Var}(F)+\mathbb{E}[F]\Big). \tag{4}$$

Thus the risky asset's price equals the discounted expected payoff plus a risk premium proportional to position size and risk aversion. Consequently, the expected gross return satisfies

$$\mathbb{E}[r] = r_f + \frac{A\,(I+X)\,\mathrm{Var}(F)}{p}. \tag{5}$$

1.2.1 Capital Asset Pricing Model

In the case of multiple risky assets, we can now derive the CAPM smoothly (see Sharpe, 1964). Starting from (3), a slightly modified and generalized version reads

$$p_i \;=\; \frac{1}{1+r_f}\left[\mathbb{E}[F_i] + \frac{\mathbb{E}[U''(w)]}{\mathbb{E}[U'(w)]}\;\mathrm{Cov}(w,F_i)\right]. \tag{6}$$

Hence the expected gross return on asset i satisfies

$$\mathbb{E}[R_i] \;=\; r_f \;-\; \frac{\mathbb{E}[U''(w)]}{\mathbb{E}[U'(w)]}\;\mathrm{Cov}(w,r_i). \tag{7}$$

Consider the market portfolio M, whose price is $p_M = \sum_i p_i X_i$ and whose return is R_M. Define the value-weights

$$w_i = \frac{p_i X_i}{p_M}.$$

Weight-averaging (7) gives

$$\sum_i w_i\,\mathbb{E}[R_i] = \sum_i w_i\Big(r_f - \tfrac{\mathbb{E}[U''(w)]}{\mathbb{E}[U'(w)]}\;\mathrm{Cov}(w,r_i)\Big) \tag{8}$$

$$\mathbb{E}[R_M] - r_f = -\frac{\mathbb{E}[U''(w)]}{\mathbb{E}[U'(w)]}\sum_i w_i\;\mathrm{Cov}(w,r_i) = -\frac{\mathbb{E}[U''(w)]}{\mathbb{E}[U'(w)]}\;\mathrm{Cov}(w,R_M). \tag{9}$$

Under the CARA specification $U(w) = -e^{-Aw}$, one shows that $\frac{\mathbb{E}[U''(w)]}{\mathbb{E}[U'(w)]} = A$, and noting that $\mathrm{Cov}(w,R_M) = p_M^{-1}\,\mathrm{Var}(R_M)\,p_M$ yields, after substitution into (7),

$$\mathbb{E}[R_i] - r_f \;=\; \frac{\mathrm{Cov}(R_i,R_M)}{\mathrm{Var}(R_M)}\;[\mathbb{E}[R_M] - r_f]. \tag{10}$$

This is precicely the Sharpe–Lintner Capital Asset Pricing Model.

[2]

Proposition 1.1 (Representative-agent pricing identity)**.** *Let there be a risk-free asset with gross return $R_f = 1 + r_f > 0$ and a single risky asset with payoff F at date T. A representative agent with strictly increasing, twice continuously differentiable utility U over terminal wealth W_T holds I initial units of the risky asset, chooses demand X, and faces price p at $t = 0$. If F is integrable and $\mathrm{Var}(F) < \infty$, then any competitive equilibrium price p satisfies*

$$p \;=\; \frac{1}{R_f}\left(\mathbb{E}[F] \;+\; \frac{\mathbb{E}[U''(W_T)]}{\mathbb{E}[U'(W_T)]}\,(I+X)\;\mathrm{Var}(F)\right),$$

where expectations are taken under the objective probability measure and $W_T = (X+I)F + (W_0 - pX)R_f$.

Proof. The first-order condition is $\mathbb{E}\big[U'(W_T)\,(F - pR_f)\big] = 0$. Using $\mathbb{E}[AB] = \mathbb{E}[A]\mathbb{E}[B] + \mathrm{Cov}(A,B)$ and $\mathrm{Cov}\big(U'(W_T), F\big) = \mathbb{E}[U''(W_T)](I+X)\,\mathrm{Var}(F)$ by the law of iterated expectations and linearity of W_T in F, one gets

$$\mathbb{E}[U'(W_T)]\,(\mathbb{E}[F] - pR_f) \;+\; \mathbb{E}[U''(W_T)](I+X)\,\mathrm{Var}(F) \;=\; 0,$$

which rearranges to the stated identity. □

Proposition 1.2 (CARA–Normal specialization and risk premium)**.** *Under the conditions of the previous proposition, suppose $U(w) = -\exp(-Aw)$ with $A > 0$ and F is independent of W_0 with variance $\mathrm{Var}(F)$. Then $\mathbb{E}[U''(W_T)]/\mathbb{E}[U'(W_T)] = -A$ and the equilibrium price satisfies*

$$p \;=\; \frac{1}{R_f}\Big(\mathbb{E}[F] \;-\; A\,(I+X)\;\mathrm{Var}(F)\Big),$$

so the expected gross return on the risky asset obeys

$$\mathbb{E}[R] \;=\; \mathbb{E}\left[\frac{F}{p}\right] \;=\; R_f \;+\; \frac{A\,(I+X)\;\mathrm{Var}(F)}{p},$$

which identifies a positive risk premium proportional to risk aversion, position size, and payoff variance.

[2] We must remember that this insight of Arrow and Debreu was under clearly specified and relatively general conditions. In my opinion, this is arguably the most foundational paper of the concept. This was supported by Lckinzie (1954)'s independent and nearly coherent work with Arrow and Debreu who reinforced the possibility of a coherent competitive equilibrium system by specifically focusing on the Gale-Nikaido-Debreu Lemma. Yes, one may agree that there were significant limitations to the Arrow-Debrew framework, but it can be shown that the aggregate excess demand function can behave almost arbitrarily (see, Sonnenschein (1972, 1973), Mantel (1974), Debreu (1974)). One must also note that the Gale–Nikaido–Debreu Lemma is a mathematical tool used to prove the existence of a competitve equilibrium. Specifically, it provides the conditions under which a system of inequalities has a deterministic solution.

Proof. For CARA, $U'(w) = Ae^{-Aw}$ and $U''(w) = -A^2e^{-Aw}$, hence $\mathbb{E}[U''(W_T)]/\mathbb{E}[U'(W_T)] = -A$. Substitute into the pricing identity and divide by p to obtain the return expression. □

1.3 Asymmetric Information

A classical platform to startoff would be Akerlof (1970). In markets with unobservable product quality (e.g., used cars), he finds asymmetric information between buyers and sellers causes adverse selection. Sellers of low-quality goods ("lemons") drive out high-quality goods because buyers cannot distinguish quality and only offer average prices.Hence, under these conditions, markets may unravel entirely or operate at suboptimal equilibria. Similarly, Stiglitz and Rothschild (1976) discuss formallized screening as a solution to asymmetric information. Hence, one thing is clear - information asymmetry can cause pareto inefficiency even in competitive markets. Hence, we move on to consider that not all agents possess the same information. There seem to be two distinct groups of traders: risk-averse agents and noise (liquidity) traders. Each agent's demand is $X_i \in \{X_I, X_U, X_N\}$, and the population sizes are $N_i \in \{N_I, N_U, N_N\}$. Noise traders submit $\hat{x} \sim \mathcal{N}(0, \sigma_x^2)$. There are no initial endowments and the risk-free rate is normalized to zero. The prior for the asset's fundamental value is $\hat{F} \sim \mathcal{N}(\bar{F}, \sigma_F^2)$. Priors represent beliefs before observing new information; posteriors incorporate private signals via Bayes' rule. At time $t = 0$, each agent receives a noisy signal $S \sim \mathcal{N}(F, \sigma_s^2)$.

Proposition 1.3 (Gaussian conjugate update for a scalar signal)**.** *Let the prior for a scalar fundamental F be $F \sim \mathcal{N}(\bar{F}, \sigma_F^2)$, and let a private signal satisfy $S \mid F \sim \mathcal{N}(F, \sigma_S^2)$, independent of other randomness. Then the posterior is Gaussian with*

$$\mathbb{E}[F \mid S] \;=\; \bar{F} \;+\; \frac{\sigma_F^2}{\sigma_F^2 + \sigma_S^2}\,(S - \bar{F}), \qquad \mathrm{Var}(F \mid S) \;=\; \frac{\sigma_F^2\,\sigma_S^2}{\sigma_F^2 + \sigma_S^2}.$$

Proof. Complete the square in the joint normal density of (F, S) or apply the linear regression formula $\mathbb{E}[F \mid S] = \bar{F} + \frac{\mathrm{Cov}(F,S)}{\mathrm{Var}(S)}(S - \mathbb{E}[S])$ with $\mathrm{Cov}(F, S) = \sigma_F^2$ and $\mathrm{Var}(S) = \sigma_F^2 + \sigma_S^2$; the conditional variance follows from the Schur complement. □

Chapter 2

Dynamic Information Revelation

Classical competitive equilibrium assumes that prices fully reflect all available information, ensuring informational efficiency. Yet, as Grossman and Stiglitz famously argued, this ideal cannot be sustained: if markets were perfectly revealing, no investor would have an incentive to incur the costs of acquiring information, and trade would vanish altogether (**?**). This paradox highlighted the inherent tension between incentives for information acquisition and the possibility of fully efficient markets.

2.1 Background and Motivation

In response, the market microstructure literature made the trading process itself explicit. Models such as Glosten and Milgrom demonstrated how order flow can act as a conduit for private information, and how adverse selection endogenously generates trading costs and bid–ask spreads even when dealers are risk-neutral and competitive (**?** **?**). These quote-driven frameworks explain spreads trade-by-trade, attributing them directly to the presence of better-informed traders. This then, should naturally raise the question: *why Kyle?* Kyle's (1985) auction-style formulation provides a complementary perspective. Instead of spreads, it emphasizes linear price impact and endogenous market depth as order-flow-based measures of illiquidity (**?** **?**). The framework became a workhorse for analyzing price discovery under asymmetric information, not only because of its tractability but also because later extensions preserved the linear structure while introducing stochastic noise-trading volatility, thereby capturing state-dependent liquidity and the empirically observed links between volume, volatility, and impact (**?**). At its core, Kyle's contribution was to embed a strategic, risk-neutral insider into a rational expectations setting with competitive market makers who observe only aggregate order flow, while noise trading sustains volume and camouflages informed trades. The resulting equilibrium is linear: prices equal the conditional

expectation of fundamentals given total flow, and the constant price impact parameter succinctly captures adverse selection.

2.2 Model

We consider a single risky asset whose terminal value v is uncertain. Before trading begins, this fundamental value v is drawn from a normal distribution with parameters $v \sim \mathcal{N}(\mu_0, \sigma_v^2)$, where $\mu_0 \in \mathbb{R}$ represents the common prior expectation about the asset's value and $\sigma_v^2 > 0$ captures the degree of fundamental uncertainty. These parameters are publicly known, i.e. they reflect the collective assessment of market participants about the asset before any private information acquisition takes place. The key innovation of Kyle's framework is the presence of an **informed trader** who, unlike other market participants, observes the true realization of v before trading. This trader essentially possesses perfect information about the fundamental value, creating a stark information asymmetry. However, this informational advantage comes with a strategic challenge: how to exploit private knowledge without fully revealing it through trading behavior. To make informed trading viable, Kyle introduces noise (liquidity) traders whose order u is distributed as $u \sim \mathcal{N}(0, \sigma_u^2)$, independent of the fundamental v. These traders represent participants who trade for reasons unrelated to the asset's fundamental value, they might be selling to meet liquidity needs, rebalancing portfolios, or responding to other non-informational motives. Crucially, the parameter $\sigma_u^2 > 0$ is known to all participants. Noise trading serves three essential functions in the model. First, it provides camouflage for informed orders: when market makers observe total order flow, they cannot perfectly distinguish between informed and uninformed components. Second, it ensures market viability: without noise, any order would immediately reveal the informed trader's signal, making information valueless. Third, it creates equilibrium depth: the presence of noise trading allows for a linear price impact that doesn't completely eliminate informed trading profits.

2.2.1 The Linear Equilibrium Ansatz

Competitive, risk-neutral market makers observe only the aggregate order flow $y = x + u$, where x is the informed trader's order. They cannot observe x and u separately—please note that this observational limitation is crucial for maintaining the information asymmetry that drives the model. Being competitive, market makers earn zero expected profits in equilibrium. Being risk-neutral, they set prices to equal their conditional expectation of the asset's value given the information available to them. This leads to the semi-strong efficient pricing condition

$$P(y) = \mathbb{E}[v \mid y]$$

This pricing rule reflects rational expectations: market makers use all available information (the order flow y) to form the best possible estimate of the fundamental value, and they set the price equal to this estimate. The model seeks a linear equilibrium where strategies take simple, tractable forms. The parameter β represents the trading intensity or how aggressively the informed trader responds to deviations of the fundamental from its prior mean. The parameter λ is the price impact coefficient, or how much prices move in response to each unit of order flow. The reciprocal $1/\lambda$ measures market depth i.e. the order size needed to move prices by one unit.

2.2.2 The Informed Trader's Problem

Now we turn to the informed trader's optimization problem, which embodies the central tension in the model: the desire to profit from private information versus the concern about moving prices adversely. The informed trader knows the true value v and conjectures that market makers will set prices according to $P(y) = \mu_0 + \lambda y$ with some positive λ. Given this pricing rule, the trader's profit from submitting order x is:

$$\begin{aligned}\pi &= x(v - P) \\ &= x\big(v - \mu_0 - \lambda(x + u)\big)\end{aligned}$$

The trader profits $x(v - \mu_0)$ from the difference between the true value and the prior expectation, but suffers a cost λx^2 from the price impact of their own trade, plus a random component $\lambda x u$ from the interaction with noise trading. Taking the conditional expectation given v (so that $\mathbb{E}[u \mid v] = 0$), we obtain

$$\mathbb{E}[\pi \mid v] = x(v - \mu_0) - \lambda x^2$$

This is a quadratic objective in x. The first term represents the expected gain from trading on the information advantage, while the second term represents the expected cost of price impact. The optimal trade balances these forces: trade more when the fundamental deviates further from the prior mean, but moderate the trade size to avoid excessive price impact.nThe first-order condition $\frac{\partial \mathbb{E}[\pi|v]}{\partial x} = 0$ yields:

$$v - \mu_0 - 2\lambda x = 0$$

Solving for x gives the insider's best response:

$$x(v) = \frac{v - \mu_0}{2\lambda} \tag{2.1}$$

Proposition 2.1 (Insider best response under linear pricing)**.** *Fix a conjectured linear pricing rule $P(y) = \mu_0 + \lambda y$ with $\lambda > 0$. Then the insider's optimal order given v is*

$$x(v) = \frac{v - \mu_0}{2\lambda},$$

so in any linear equilibrium one must have $\beta = \frac{1}{2\lambda}$.

This reveals the key insight: the informed trader's optimal strategy is indeed linear in the fundamental, with trading intensity $\beta = \frac{1}{2\lambda}$. The trader trades more aggressively (higher β) when price impact is low (low λ), and more conservatively when price impact is high. The factor of $\frac{1}{2}$ emerges from the quadratic nature of the price impact cost—this is the familiar result from monopolistic pricing where the markup is half the demand slope.

2.2.3 Market Makers and Linear Bayesian Updating

Having established the informed trader's optimal strategy, we now turn to the market makers' problem. Market makers must infer the fundamental value from the order flow they observe, knowing that this flow contains both informed and noise components.

When the informed trader uses the strategy $x(v) = \beta(v - \mu_0)$, the total order flow becomes $y = \beta(v - \mu_0) + u$. This creates a linear relationship between the unobservable fundamental v and the observable order flow y, contaminated by the noise term u.

Since both v and u are normally distributed and independent, the joint distribution of (v, y) is bivariate normal. This Gaussian structure allows us to apply the linear projection formula for conditional expectations. The market makers' optimal pricing rule is to set the price equal to the conditional expectation of the fundamental given the observed order flow:

$$P(y) = \mathbb{E}[v \mid y] = \mu_0 + \lambda y$$

To determine the slope coefficient λ, we use the fact that for jointly normal random variables, the conditional expectation is linear with slope equal to the ratio of covariance to variance:

$$\lambda = \frac{\mathrm{Cov}(v, y)}{\mathrm{Var}(y)}$$

Under the informed trading strategy $y = \beta(v - \mu_0) + u$, we can compute these moments. The covariance between v and y is $\mathrm{Cov}(v, y) = \beta\sigma_v^2$, since u is independent of v. The variance of the order flow is $\mathrm{Var}(y) = \beta^2\sigma_v^2 + \sigma_u^2$, reflecting both the variability induced by informed trading and the exogenous noise.

Therefore, the price impact coefficient is:

$$\lambda = \frac{\beta\sigma_v^2}{\beta^2\sigma_v^2 + \sigma_u^2}$$

The equilibrium requires that both the informed trader's best response and the market makers' pricing rule be mutually consistent. We have two equations: $\beta = \frac{1}{2\lambda}$ from the informed trader's optimization, and $\lambda = \frac{\beta\sigma_v^2}{\beta^2\sigma_v^2+\sigma_u^2}$ from the market makers' inference problem. These two conditions must be satisfied simultaneously.
Substituting the first into the second yields a quadratic equation that can be solved to obtain the unique positive solution:

$$\beta = \frac{\sigma_u}{\sigma_v}, \qquad \lambda = \frac{\sigma_v}{2\sigma_u}$$

These equilibrium values reveal important economic intuitions. The trading intensity β increases with noise variance σ_u^2 and decreases with fundamental variance σ_v^2. More noise provides better camouflage, encouraging more aggressive informed trading. Conversely, higher fundamental uncertainty makes each unit of information less precise, leading to more cautious trading.
The price impact λ decreases with noise variance and increases with fundamental variance. Market depth, measured by $1/\lambda = 2\sigma_u/\sigma_v$, is higher when there is more noise trading relative to fundamental uncertainty. This captures the intuitive idea that markets with more noise trading can absorb informed orders with less price movement.

2.2.4 Price Informativeness and Learning

A crucial question in any model of asymmetric information is how much private information gets revealed through the trading process. In Kyle's model, this can be measured by comparing the prior uncertainty about the fundamental with the posterior uncertainty after observing the order flow. The posterior variance of the fundamental given the order flow is calculated using the standard formula for conditional variance in the bivariate normal case

$$\mathrm{Var}(v \mid y) = \sigma_v^2 - \frac{\mathrm{Cov}(v, y)^2}{\mathrm{Var}(y)}$$

Substituting the equilibrium values, we find:

$$\mathrm{Cov}(v, y) = \beta\sigma_v^2 = \frac{\sigma_u}{\sigma_v} \cdot \sigma_v^2 = \sigma_u\sigma_v$$

$$\mathrm{Var}(y) = \beta^2\sigma_v^2 + \sigma_u^2 = \frac{\sigma_u^2}{\sigma_v^2} \cdot \sigma_v^2 + \sigma_u^2 = 2\sigma_u^2$$

Therefore:

$$\text{Var}(v \mid y) = \sigma_v^2 - \frac{(\sigma_u \sigma_v)^2}{2\sigma_u^2} = \sigma_v^2 - \frac{\sigma_v^2}{2} = \frac{\sigma_v^2}{2}$$

This remarkable result shows that a single trading round in the Kyle model reveals exactly half of the prior variance, regardless of the noise level σ_u^2. This invariance property is a distinctive feature of the Kyle equilibrium and reflects the endogenous adjustment of trading intensity to noise levels.

2.2.5 Profitability and Market Impact

The informed trader's expected profit provides another lens through which to understand the equilibrium. From the quadratic optimization problem, the conditional expected profit given the fundamental realization is:

$$\mathbb{E}[\pi \mid v] = \frac{(v - \mu_0)^2}{4\lambda}$$

Taking expectations over the fundamental gives the ex ante expected profit

$$\begin{aligned} \mathbb{E}[\pi] &= \frac{\mathbb{E}[(v - \mu_0)^2]}{4\lambda} \\ &= \frac{\sigma_v^2}{4\lambda} \end{aligned}$$

The expression reveals that expected profits increase with both fundamental uncertainty (more valuable information) and noise trading (better camouflage). The profit is proportional to the geometric mean of the two variance parameters, highlighting the complementary nature of information value and camouflage. The comparative statics of market depth deserve special attention. Since $\lambda = \sigma_v/(2\sigma_u)$, price impact decreases with noise variance, while market depth $1/\lambda = 2\sigma_u/\sigma_v$ increases linearly with noise variance. This endogenous relationship between noise trading and market liquidity is central to understanding how markets self-organize around information asymmetries.

Matcha with Ayyar

Over a warm cup of matcha, let's pause and think about something that often confuses students...

Question: "I keep hearing about different sources of trading costs in market microstructure models. In the Kyle model, where exactly do these costs come from? Are they the same as the inventory costs I read about in other papers?"

Great question! This is actually a subtle but important distinction that gets to the heart of what drives spreads and price impact in different market structures. In Kyle's model, the trading costs arise purely from **adverse selection**. Here's what's happening: the market makers know that some of the orders they see come from informed traders who know more about the asset's true value. This creates a classic "winner's curse" problem; when market makers get hit by a large order, it's more likely to be coming from someone who knows bad news (if it's a sell order) or good news (if it's a buy order). To protect themselves from this adverse selection, market makers build the expected cost into their pricing. This shows up as the price impact parameter $\lambda = \frac{\sigma_v}{2\sigma_u}$, which measures how much the price moves per unit of order flow. The key insight is that this impact exists even though market makers are risk-neutral and competitive, they're not worried about holding inventory per se, they're worried about being picked off by better-informed traders.

Now, **inventory costs** are a different animal entirely. They arise when market makers are risk-averse and worry about the risk of holding positions. Consider a dealer with CARA utility $U(w) = -\exp(-\gamma w)$ who holds inventory q. Their certainty equivalent from this position is

$$CE = q(\mu_0 - P) - \frac{\gamma}{2} q^2 \sigma_v^2$$

That second term $\frac{\gamma}{2} q^2 \sigma_v^2$ is pure inventory cost. It increases quadratically with position size and reflects the dealer's aversion to bearing risk. Crucially, this cost exists even if there's no asymmetric information at all! The beauty of Kyle's framework is that it isolates the adverse selection channel cleanly. The risk-neutral assumption strips away inventory concerns, leaving us with a pure laboratory to study how private information gets impounded into prices through strategic trading. In real markets, of course, both effects are likely present, but understanding them separately is crucial for empirical work that tries to decompose bid-ask spreads into their components.

The single-period Kyle model elegantly captures how informed trading, market making, and noise provision interact to incorporate private information into prices. Yet real markets are dynamic: information arrives over time, traders adapt their strategies, and strategic interactions evolve. Extending Kyle's framework to multiple periods transforms a static snapshot into a dynamic theory of information-based price discovery. Kyle himself introduced the multi-period extension in his 1985 Econometrica paper, showing that the insider's problem becomes one of dynamic programming: current profits must be balanced against the information revealed to market makers, which alters future opportunities. This temporal trade-off introduces genuine intertemporal strategy—far more than a repetition of the single-period game. Kyle's discrete-time formulation demonstrated linear equilibria through recursive difference equations.

2.3 Multiperiod Kyle

The breakthrough came with Back (1992), who proved that as trading intervals shrink, the discrete model converges to a tractable continuous-time limit. This insight provided the mathematical foundation for a generation of advances in dynamic microstructure theory. In continuous time, the Kyle framework has inspired extensive research: multiple insiders, dynamic information acquisition, stochastic noise volatility, funding constraints, disclosure requirements, and correlated signals across assets. The unifying theme is that informed traders manage information intertemporally—trading less aggressively early on to preserve private information, then accelerating as horizons shorten. The model predicts rich dynamics: market depth typically increases as the terminal date approaches; price informativeness rises as uncertainty resolves; and trading intensity follows time-varying patterns shaped by noise, horizon, and signal precision. Unlike the static case where price impact depends only on the signal-to-noise ratio, in multi-period settings impact itself becomes a forward-looking process, influenced by expectations of future order flow and information release. Crucially, the multi-period Kyle model retains linear equilibrium structure, allowing for closed-form characterizations despite the dynamic complexity. This combination of tractability and depth explains why it remains a cornerstone of market microstructure, with implications for optimal execution, high-frequency trading, and the design of modern electronic markets.

2.4 Model Setup

The multiperiod Kyle model preserves the three-player structure of the single-period version-informed trader, noise traders, and competitive market makers while introducing intertemporal dynamics that fundamentally alter strategy. Its power lies in combining simple

Gaussian-linear assumptions with dynamic optimization, allowing tractable analysis of how information gets revealed and prices adjust over time. A finite horizon is assumed, both for analytical convenience (backward induction via dynamic programming) and for economic realism: private information often expires (e.g., earnings announcements, merger outcomes, or patent approvals), creating urgency and shaping trading incentives. The model runs for T discrete periods, with a risky asset of terminal value $v \sim \mathcal{N}(\bar{v}, \sigma_v^2)$. At time zero, the informed trader learns the true v, while the market only knows the prior. Each period, the informed trader chooses an order x_t, noise traders submit independent demands $u_t \sim \mathcal{N}(0, \sigma_u^2)$, and market makers observe the total flow $y_t = x_t + u_t$. Prices update via conditional expectation, $p_t = \mathbb{E}[v \mid y_1, \ldots, y_t]$. The informed trader's challenge is dynamic: trading too aggressively reveals information and reduces future profits, while trading too cautiously underutilizes the informational advantage. The problem is thus an optimal control problem balancing immediate gains against preserving information rents across time.

2.4.1 Strategic Interaction

The informed trader solves a dynamic programming problem, choosing the sequence $(x_1, \ldots, x_T)$ to maximize expected cumulative profit. Current trades affect both immediate returns and the informativeness of future prices, creating intertemporal externalities. Optimal strategies typically imply declining trading intensity as the horizon shortens. Noise traders supply the camouflage that sustains informed trading. Their period-by-period independent orders represent liquidity needs unrelated to fundamentals, providing the randomness that prevents perfect inference by market makers. Market makers, observing only aggregate flows, update beliefs using Bayesian inference. Thanks to the Gaussian-linear structure, this process admits closed-form characterization through Kalman filter recursions. Prices form a martingale that gradually converges to the true value as information is revealed. The outcome is a dynamic process of price discovery that predicts how market depth, informativeness, and liquidity evolve over time.

Matcha with Ayyar

Let me pour some matcha and think about what changes when we go dynamic...

Question: "I understand the one-period Kyle model, but I'm confused about the multiperiod version. If the informed trader knows v from the beginning, why doesn't he just trade his entire position immediately in period 1 to maximize profits?"

Think of it this way: if the informed trader dumps his entire desired position in period 1, the massive order flow would cause a huge price movement. The market makers, seeing this

large order, would infer that someone has very strong information about the asset's value. This would cause prices to move most of the way to the fundamental value immediately! So while the trader gets high profits per unit traded in period 1 (since the price hasn't moved much yet), he's "killed the golden goose"—there's no information advantage left for periods 2 through T. The optimal strategy involves a delicate balance: trade enough today to capture some profits, but not so much that you give away all your informational advantage. It's like being a poker player who knows everyone's cards—you want to win money, but if you bet too aggressively on every hand, everyone will figure out that you're cheating! This creates a beautiful dynamic optimization problem where the informed trader is essentially deciding how fast to reveal his private information to the market.

2.4.2 Price Process and Information Revelation

A central insight of the multiperiod Kyle model is that prices are martingales under the public filtration generated by order flow. Let $\mathcal{F}_t = \sigma(y_1, \dots, y_t)$. With competitive, risk–neutral market makers,

$$p_t \;=\; \mathbb{E}x[v \mid \mathcal{F}_t] \quad \Rightarrow \quad \mathbb{E}[p_t \mid \mathcal{F}_{t-1}] = p_{t-1},$$

and under linear–Gaussian structure the pricing rule takes the form

$$p_t \;=\; p_{t-1} + \lambda_t\, y_t, \qquad y_t \;=\; x_t + u_t, \qquad u_t \sim \mathcal{N}(0, \sigma_u^2),$$

where $\lambda_t > 0$ is the period-t price–impact (inverse depth). The cumulative decomposition

$$v - \bar{v} \;=\; \sum_{t=1}^{T} \lambda_t\, y_t \;+\; \varepsilon_T$$

holds with a terminal residual ε_T that reflects remaining (posterior) uncertainty about v after T periods. The linear–Gaussian setting implies Kalman–filter updates for the posterior variance:

$$\sigma_t^2 \;\equiv\; \mathrm{Var}(v \mid \mathcal{F}_t) \;=\; \sigma_{t-1}^2 \;-\; \frac{\mathrm{Cov}(v, y_t \mid \mathcal{F}_{t-1})^2}{\mathrm{Var}(y_t \mid \mathcal{F}_{t-1})} \;=\; \sigma_{t-1}^2\, \frac{\sigma_u^2}{\beta_t^2 \sigma_{t-1}^2 + \sigma_u^2},$$

once we specify the insider's linear strategy $x_t = \beta_t\,(v - p_{t-1})$. Equivalently, the period-t *information revelation rate* is

$$\rho_t \;\equiv\; \frac{\sigma_{t-1}^2 - \sigma_t^2}{\sigma_{t-1}^2} \;=\; \frac{\beta_t^2 \sigma_{t-1}^2}{\beta_t^2 \sigma_{t-1}^2 + \sigma_u^2} \in (0, 1),$$

so $\sigma_t^2 = \sigma_{t-1}^2(1 - \rho_t)$ and hence $\sigma_T^2 = \sigma_v^2 \prod_{t=1}^{T}(1 - \rho_t)$.

2.4.3 The Informed Trader's Dynamic Problem

Let $\Delta_t \equiv v - p_t$ and $W_t(p_t, v)$ be the insider's continuation value from period t. A single trade in period t at order x_t yields *expected* one–period profit

$$\mathbb{E}_t\Big[(v - p_t)x_t - \lambda_t x_t^2 - \lambda_t x_t u_t\Big] \;=\; \Delta_t x_t \;-\; \lambda_t x_t^2,$$

and pushes the next price via $p_{t+1} = p_t + \lambda_t(x_t + u_t)$, so $\Delta_{t+1} = \Delta_t - \lambda_t(x_t + u_t)$. With discount factor $\delta \in (0, 1]$, the Bellman equation is

$$W_t(p_t, v) \;=\; \max_{x_t}\; \mathbb{E}_t\Big[\,\Delta_t x_t - \lambda_t x_t^2 + \delta\, W_{t+1}(p_{t+1}, v)\,\Big].$$

Proposition 2.2 (Quadratic value function and optimality condition)**.** *There exist coefficients $\{A_t, B_t\}_{t=1}^{T+1}$ with terminal condition $A_{T+1} = B_{T+1} = 0$ such that*

$$W_t(p_t, v) \;=\; A_t\,\Delta_t^2 + B_t.$$

Given A_{t+1} and λ_t, the insider's period-t optimal order is linear,

$$x_t^\star \;=\; \beta_t\,\Delta_t, \qquad \beta_t \;=\; \frac{2\delta A_{t+1}\lambda_t - 1}{2\lambda_t\,(1 - \delta A_{t+1}\lambda_t)},$$

and the value–function coefficient satisfies the backward recursion

$$A_t \;=\; \frac{1}{4\,\lambda_t\,(1 - \delta A_{t+1}\lambda_t)}.$$

Moreover, $B_t = \delta B_{t+1} + \frac{\delta A_{t+1}\lambda_t^2(\delta A_{t+1}\lambda_t - 1)}{1 - \delta A_{t+1}\lambda_t}\,\sigma_u^2$.

Proof. Plug the quadratic ansatz into the Bellman equation; take expectations using $\mathbb{E}[u_t] = 0$ and $\mathbb{E}[u_t^2] = \sigma_u^2$; maximize the resulting quadratic in x_t. The first–order condition yields the stated β_t. Substituting $x_t^\star$ back gives the recursions for A_t and B_t. □

2.4.4 Market Maker Pricing and Dynamic Consistency

With $x_t = \beta_t \Delta_{t-1}$, the covariance and variance terms are

$$\mathrm{Cov}(v, y_t \mid \mathcal{F}_{t-1}) = \beta_t \sigma_{t-1}^2, \qquad \mathrm{Var}(y_t \mid \mathcal{F}_{t-1}) = \beta_t^2 \sigma_{t-1}^2 + \sigma_u^2,$$

hence competitive pricing implies

$$\lambda_t \;=\; \frac{\beta_t \sigma_{t-1}^2}{\beta_t^2 \sigma_{t-1}^2 + \sigma_u^2} \quad \text{and} \quad p_t \;=\; p_{t-1} + \lambda_t y_t.$$

Equations in Proposition 2.2 together with the pricing and variance updates

$$\sigma_t^2 \;=\; \sigma_{t-1}^2 \frac{\sigma_u^2}{\beta_t^2 \sigma_{t-1}^2 + \sigma_u^2}$$

jointly characterize equilibrium via backward–forward recursion.

2.5 Equilibrium Characterization

Theorem 2.3 (Linear equilibrium: existence, uniqueness, and dynamics)**.** *Fix $\delta \in (0, 1]$, $\sigma_v^2 > 0$, and $\sigma_u^2 > 0$. There exists a unique linear equilibrium with strategies $x_t = \beta_t(v - p_{t-1})$ and prices $p_t = p_{t-1} + \lambda_t y_t$ such that for $t = 1, \ldots, T$:*

$$\beta_t \;=\; \frac{2\delta A_{t+1}\lambda_t - 1}{2\lambda_t\,(1 - \delta A_{t+1}\lambda_t)}, \quad A_t \;=\; \frac{1}{4\,\lambda_t\,(1 - \delta A_{t+1}\lambda_t)}, \quad \lambda_t \;=\; \frac{\beta_t \sigma_{t-1}^2}{\beta_t^2 \sigma_{t-1}^2 + \sigma_u^2}.$$

In the canonical case $\delta = 1$ with homoscedastic noise σ_u^2 and a single insider:

1. *Trading intensity is* increasing *over time: $\beta_1 < \beta_2 < \cdots < \beta_T$.*
2. *Price impact is* decreasing *over time: $\lambda_1 > \lambda_2 > \cdots > \lambda_T$.*
3. *Information revelation accelerates: $\rho_1 < \rho_2 < \cdots < \rho_T$ and $\rho_T = \frac{1}{2}$.*
4. *Period-t expected profit is*

$$\mathbb{E}[\pi_t] \;=\; \mathbb{E}[\Delta_{t-1} x_t - \lambda_t x_t^2] \;=\; \frac{\beta_t \sigma_{t-1}^2\, \sigma_u^2}{\beta_t^2 \sigma_{t-1}^2 + \sigma_u^2},$$

so total expected profit is $\sum_{t=1}^{T} \mathbb{E}[\pi_t]$.

Intuition. Early on, the insider protects future rents (trades cautiously), but the large residual uncertainty σ_{t-1}^2 makes each unit of order flow less informative, leading to higher depth (lower λ_t) later only after enough information is revealed. By the terminal period, the insider behaves as in a one–shot Kyle game with remaining variance σ_{T-1}^2, revealing exactly half of that variance.

Matcha with Ayyar

"Why does β_t rise over time while λ_t falls?"

Urgency grows as the horizon shrinks, pushing the insider to trade more aggressively (rising β_t). At the same time, previous trading has already reduced posterior variance, so each

unit of new order flow is *less* masked by noise relative to the shrinking uncertainty set. Market makers therefore need *less* slope to extract the same information (falling λ_t), and the terminal step always reveals half of what remains. The two forces—urgency vs. remaining uncertainty—jointly generate rising intensity but falling impact.

Two-Period Model ($T = 2$)

Backward induction yields:

$$\lambda_2 = \frac{\sigma_1}{2\sigma_u}, \qquad \beta_2 = \frac{\sigma_u}{\sigma_1}, \qquad \sigma_1^2 = \sigma_v^2 \, \frac{\sigma_u^2}{\beta_1^2 \sigma_v^2 + \sigma_u^2}, \qquad \lambda_1 = \frac{\beta_1 \sigma_v^2}{\beta_1^2 \sigma_v^2 + \sigma_u^2}.$$

The optimal β_1 maximizes $\mathbb{E}[\pi_1] + \mathbb{E}[\pi_2]$, delivering $\beta_2 > \beta_1$ and $\lambda_2 < \lambda_1$. Closed forms follow from the first–order condition but are omitted for brevity.

Continuous-Time Limit

Let $\Delta t = 1/T \to 0$. The discrete model converges to a continuous–time Kyle economy (Back, 1992) in which

$$dP_t = \Lambda(t)\, dY_t, \qquad dY_t = \beta(t)\,(v - P_t)\, dt + dU_t,$$

with $\{U_t\}$ a Brownian motion with variance rate σ_u^2. The residual variance $\sigma(t)^2 = \mathrm{Var}(v \mid \mathcal{F}_t)$ solves a Riccati–type ODE, and $\Lambda(t) = \frac{1}{2\sigma_u}\,\sigma(t)$ while $\beta(t)$ increases as time to maturity shrinks.

Chapter 3

One Ping Only

"Reverify our range to target. ONE PING only."

—Captain Marco Ramius

In the frigid depths of the North Atlantic, far from the familiar shores of Murmansk, Kapitan Marco Ramius faces an existential crisis of navigation and survival. His objective requires the determination of an exact distance to an unseen adversary, yet he is constrained by the most rudimentary of instruments: an acoustic wave projected into the abyss, followed by the harrowing wait for its return. The command for "one ping only" is not a gesture of aesthetic restraint or moral virtue; rather, it is a calculated response to the reality that every acoustic transmission is a hazardous declaration of presence. In the silence of sub-surface combat, to speak is to be found, yet to remain silent is to remain blind. This fundamental tension between the necessity of information and the cost of its acquisition provides a precise, if not immediately obvious, parallel to the modern electronic limit order book. Within the high-frequency environments that govern global equity trading, a comparable darkness persists. Institutional participants stand at a structural precipice, requiring a granular understanding of liquidity depth before committing to large-scale execution. However, the most vital information, the precise location and volume of hidden reserves, is frequently concealed behind the "acoustic silence" of iceberg orders. While the public data feed provides a superficial rendering of surface depth, beneath this visible layer lies an invisible architecture of dormant volume that radically alters execution dynamics in ways that traditional, passive models fail to interpret. This chapter seeks to establish this parallel through the lens of analogy, moving beyond mere mathematics to explore the philosophy of active measurement.

3.1 Information as Revelation

We begin by examining the mechanics of sonar, distinguishing between the fragmentary nature of passive observation and the definitive, albeit risky, certainty of active measurement. We then explore how tomographic principles allow for the reconstruction of complex internal structures from a series of simplified echo patterns. By tracing the evolution of the limit order book as a mechanism for price discovery, we demonstrate that these markets are best understood as queuing systems defined by a constraint: the majority of the queue remains hidden from view. Finally, we introduce the core intuition of our framework, proposing that by transmitting carefully calibrated "pings", in the form of discrete probe orders and analyzing the resulting "echoes" of execution timing, we can reconstruct the latent depth of the hidden queue with the same precision that active sonar reveals a submarine. Considering the epistemological challenge of extracting information from an opaque environment, one must first appreciate that in both the ocean and the exchange, measurement is not merely an observation; it is an act of revelation . Sonar, or "sound navigation and ranging," serves as the quintessential technology for navigating media where light is rendered useless by absorption and scattering . While acoustic waves possess the remarkable capacity to traverse thousands of kilometers with clarity, this physical advantage is governed by a strict constraint: silence yields no definitive data . To understand the mechanism of price discovery in an invisible queue, one must first distinguish between the two primary modes of acoustic interrogation.

3.1.1 Passive Measurement

The paradigm of passive measurement relies entirely upon the interpretation of ambient acoustic signatures . A sonar operator listens for the discrete artifacts of existence:rhythmic propeller signatures of distant vessels, the structural groans of hulls under pressure, or the industrial hum of the seabed. While this method is inherently silent and preserves the listener's stealth, the resulting information is fundamentally fragmentary . Signals are frequently refracted through varying thermal layers or degraded by distance, leaving the observer to rely on experience and pattern recognition rather than empirical certainty . In the context of the market, this is analogous to the trader who watches the public tape, attempting to infer hidden liquidity from the "noise" of visible trades without ever testing the depth themselves.

3.1.2 Active Alternative

In contrast, active measurement represents a decisive transition from inference to interrogation. Here, the observer actively projects a pulse of acoustic energy into the darkness

and awaits its reflection. Should an object reside at range r, the pulse will reflect off the surface and return to the transmitter. Utilizing the known speed of sound in water, $c \approx 1500$ meters per second, and the observed round-trip latency τ, the distance is derived through a mathematically unambiguous identity:$r = \frac{c\tau}{2}$. While this direct measurement eliminates the ambiguity of passive listening, it imposes a significant cost: the outgoing signal serves as a beacon, announcing the observer's presence and tactical intentions to the entire surrounding medium. In the naval theater, as in the limit order book, active measurement is an act of commitment. When the captain chooses to "ping," they are conceding the advantage of stealth in exchange for the absolute certainty of the target's position. This tactical trade-off, the exposure of one's own position for the sake of uncovering the adversary's, is the foundational logic of the probe-based execution strategies we develop hereafter. The utility of active measurement extends far beyond the determination of simple distance; it provides the raw data necessary to reconstruct hidden internal architectures . This principle is most clearly demonstrated in the field of medical imaging through computed tomography (CT). A CT scanner does not rely on a single perspective; instead, it projects a multitude of X-ray beams through a patient's body from a variety of angles . Each ray is absorbed at a different rate, contingent upon the integrated density of the tissue along its specific trajectory. While no single measurement reveals the entirety of the internal landscape, the collection of these one-dimensional observations allows for the mathematical inversion of the system, ultimately reconstructing a complete three-dimensional representation of structures that remain invisible to the naked eye . The fundamental catalyst for this reconstruction is the principle of conservation. When a ray of energy passes through a cross-section, its total absorption is the sum of the densities it encountered . By varying the angle of the "pings"—or in our case, the timing and placement of probe orders—the resulting pattern of reflections creates an overdetermined system . Just as modern active sonar can synthesize multiple echoes to generate a detailed acoustic image of canyons and ridges on the ocean floor, our tomographic framework utilizes the "echoes" of execution timing to map the dormant volume of the limit order book . The key to this precision is variation: by introducing structured differences in measurement, we can combine fragmentary reflections into a singular, high-fidelity model of hidden liquidity.

3.2 Limit Order Books

The application of sonar principles to the financial landscape requires a foundational understanding of the structural evolution of market mechanisms and the subsequent emergence of the queue-theoretic paradigm. For the vast majority of financial history, the intersection of buyer and seller was mediated through human agency— specifically, a specialist or mar-

ket maker standing upon a physical exchange floor who facilitated transactions through a combination of voice negotiation and hand signals. This manual era, exemplified by the late-twentieth-century New York Stock Exchange, relied on a specialist who maintained order books on physical media, such as paper ledgers or chalkboards, matching orders through informal convention and personal discretion. Such a process was constrained by the inherent biological limitations of human information processing and the deliberate asymmetric visibility of the specialist, who possessed the exclusive right to view the full depth of market interest without any obligation to disclose that structure to the public.

3.2.1 The Electronic Limit Order Book

The migration toward electronic markets fundamentally deconstructed this manual intermediary model, replacing human intuition with the rigid, sensor-driven logic of digital networks. Beginning with the introduction of electronic communication networks (ECNs) in the 1970s and accelerating with the expansion of the NASDAQ, the marketplace transformed into a limit order book: a sophisticated data structure that maintains an active queue of unfilled limit orders. In this environment, orders are not negotiated but sorted via the protocol of Price-Time Priority (PTP), which places the highest-priority orders at the vanguard of the queue. When a market order is deployed, it acts as a kinetic force that executes against these standing limit orders in strict sequence, depleting the queue as it progresses through the price levels. This transition from physical floor to digital ledger revealed a profound structural isomorphism: the problem of order execution in a high-frequency environment is mathematically identical to the classical queuing problems studied in operations research. For the institutional participant, the limit order book presents a binary tactical choice: the trader may either utilize a market order to achieve immediate execution at the cost of paying the spread, or submit a limit order to occupy a position within the queue. The latter strategy, while avoiding the immediate cost of the spread, introduces a stochastic wait time that is dictated by the arrival rate of market orders, the frequency of order cancellations, and the potential for new limit orders to "cut ahead" as price regimes shift . Consequently, the ability to execute successfully becomes a function of understanding the discipline of the queue—a discipline that remains robust in its First-In-First-Out (FIFO) nature, even as the volume within it becomes increasingly obscured

3.3 Stochastic Queue Theory

The formalization of order execution dynamics necessitates a transition from simple arithmetic to the rigorous framework of stochastic processes. Within this paradigm, the electronic limit order book is most elegantly modeled as an $M/M/1$ queuing system. In this environment, the

arrival of market orders, i.e. the "service" events is characterized as a Poisson process with an intensity rate of $\lambda_M 2$. Conversely, the departure of liquidity through order cancellations occurs at a rate defined by $\lambda_C 3$. The state of the system is defined by the queue depth Q, which serves as the primary determinant for the temporal dimension of execution4.

Proposition 3.1. *Under the assumption of locally stationary Poisson arrival and cancellation rates, the expected execution latency $\mathbb{E}[\tau]$ for an order positioned at depth Q is given by the identity*

$$\mathbb{E}[\tau] = \frac{Q}{\lambda_M + \lambda_C}$$

where the denominator represents the aggregate depletion intensity of the price level.

Traders have historically understood this relationship through intuition: a deeper queue necessitates a longer duration for service, whereas an acceleration in market activity facilitates a more rapid fill. However, this transparent model is predicated on the assumption that Q is fully observable, paradoxically an assumption that modern market microstructure systematically violates.

3.3.1 An Invisible Queue

The structural integrity of the simple $M/M/1$ model is compromised by the strategic use of iceberg orders. An iceberg order functions as a mechanism for concealment, where the vast majority of a participant's volume remains hidden from the public data feed, leaving only a nominal "visible portion" to be displayed. As the matching engine depletes this visible fraction, the exchange automatically replenishes it from the hidden reserve, creating the illusion of a shallow queue while maintaining a substantial, invisible barrier.This concealment is a rational response to the risk of information leakage. The public disclosure of a large institutional position invites adverse selection; savvy market participants can infer inventory pressure and adjust their pricing strategies accordingly, thereby widening spreads and inflating execution costs. While icebergs successfully mitigate this leakage, they introduce a fundamental opacity into the market's state variables. The true queue depth Q_{true} that dictates actual execution duration is no longer equivalent to the visible depth Q_{visible} reported by the exchange.

Remark 3.2. The discrepancy between visible and true depth can be formalized as:

$$Q_{\text{true}} = Q_{\text{visible}} + H$$

where $H \geq 0$ denotes the latent volume. Consequently, a naive estimator of execution time, $\widehat{\tau}_{\text{naive}}$, will systematically underestimate the true latency τ_{true} by a factor of $(1 + H/Q_{\text{visible}})$.

For execution algorithms such as TWAP or VWAP, this bias is catastrophic, leading to a profound underestimation of execution risk and a subsequent increase in implementation shortfall.

3.3.2 Order Book Tomography

To resolve this informational deficit, we return to the sonar analogy: just as a naval operator interrogates the darkness of the Atlantic to determine the range of a target, a trader must interrogate the limit order book to determine the true depth of the queue. This active measurement relies on the strict FIFO discipline that governs electronic exchanges. In this environment, a unit-sized limit order P_1 submitted at time t_1 serves as the initial "ping". As it joins the end of the visible queue, it initiates a measurement cycle that concludes at time T_1—the moment of the order's execution. However, a single pulse provides only a localized snapshot. To reconstruct the architecture of the queue, we deploy a second probe, P_2, at time $t_2 = t_1 + \delta$. During the controlled interval δ, the queue evolves: new visible orders arrive, existing orders are cancelled, and hidden icebergs may be injected. When P_2 executes at time T_2, the resulting execution interval $T_2 - T_1$ encodes the cumulative evolution of the queue. By invoking the Conservation Principle, we assert that the total volume removed from the queue between the two execution timestamps must equal the sum of the visible arrivals and the hidden components that stood between the probes25. Since the market tape provides an explicit record of every trade and cancellation between T_1 and T_2, the total depletion is an observable fact26. By subtracting the observable visible arrivals from this total depletion, we successfully isolate the hidden volume H, transforming market darkness into precise, actionable knowledge.

3.4 Theoretical Foundation

We formalize the limit order book (henceforth LOB) dynamics under the assumption of a Price-Time Priority matching engine, where orders at a given price level are executed according to a First-In-First-Out discipline.

3.4.1 Queue Geometry Mechanics

The state of the queue at price p and time t, denoted by $Q(t,p)$, evolves as a stochastic process driven by the interplay of liquidity provision as well as consumption. Specifically, the queue length is governed by the net aggregate of limit order arrivals, market order executions, and cancellations, given by

$$Q(t,p) = \sum_i v_i^L \mathbb{I}_{\{t_i^L \le t\}} - \sum_j v_j^M \mathbb{I}_{\{t_j^M \le t\}} - \sum_k v_k^C \mathbb{I}_{\{t_k^C \le t\}} \tag{3.1}$$

where v_i^L, v_j^M, and v_k^C represent the volumes of the i-th limit order, j-th market order, and k-th cancellation, respectively. Under PTP, the queue position is the primary determinant of execution quality. An order's position governs its execution probability, as front-of-queue orders are filled prior to those at the back; it dictates adverse selection risk, as orders deeper in the queue are more exposed to toxic flow and "picking-off" risks; and it defines the expected fill rate, directly impacting the opportunity cost of waiting. Crucially, the observable queue $Q_{visible}(t,p)$ reported by market data feeds is often a strict subset of the true liquidity available. The true queue depth $Q_{true}(t,p)$ accounts for hidden liquidity, commonly referred to as "iceberg" or reserve orders, such that:

$$Q_{true}(t,p) = Q_{visible}(t,p) + H(t,p) \tag{3.2}$$

where $H(t,p) \ge 0$ represents the latent volume concealed from the public tape. To estimate this latent component, we introduce a probe-based active inference mechanism. First, we submit two sequential limit orders, P_1 and P_2, both of unit quantity, at the best bid(resp. ask) price p. P_1 is submitted at $t = 0$, and P_2 is submitted after a deterministic latency δ, at $t = \delta$. Between the submission of P_1 and P_2, the visible queue expands due to new limit order arrivals. We define the *Gap Volume*, V_{gap}, as the cumulative visible volume added to the queue during the interval $[0, \delta]$. Consequently, the true distance between the queue positions of P_1 and P_2, denoted q_1 and q_2, is the sum of the visible gap volume and the unknown hidden volume accumulated in that interval:

$$q_2 - q_1 = V_{gap} + H_{gap} \tag{3.3}$$

where H_{gap} is the hidden volume added between the probes. The core objective of the Multidimensional Latency Tomography algorithm is to recover H_{gap} by analyzing the differential execution times of P_1 and P_2.

3.4.2 Multidimensional Intensity Framework

We model the order flow as a multivariate point process $N_t = (N_t^L, N_t^C, N_t^M)$, representing the counting processes for limit orders, cancellations, and market orders, respectively. The dynamics of these processes are characterized by their conditional execution intensities $\lambda(t)$, defined as the expected arrival rate conditioned on the filtration $\mathcal{F}_t$ of market history:

$$\lambda^X(t) = \lim_{\Delta t \to 0} \frac{\mathbb{E}[N_{t+\Delta t}^X - N_t^X \mid \mathcal{F}_t]}{\Delta t}, \quad X \in \{L, C, M\} \tag{3.4}$$

Following the microstructure models of Cont et al. (2010) and Bacry et al. (2015), we specify these intensities using a Hawkes process framework to capture the self-exciting and cross-exciting nature of order book events. The intensity for event type i is given by a baseline intensity μ_i augmented by a convolution of past events with an excitation kernel ϕ_{ij}:

$$\lambda_i(t) = \mu_i + \sum_j \int_0^t \phi_{ij}(t-s) dN_j(s) \tag{3.5}$$

For computational tractability, we employ an exponential decay kernel $\phi_{ij}(u) = \alpha_{ij} e^{-\beta_{ij} u}$, which allows for efficient recursive estimation of the intensities. The effective rate of queue depletion, which drives the execution of our probe orders, is the aggregate intensity of volume-removing events:

$$\lambda_{depletion}(t) = \lambda^M(t) + \lambda^C(t) \tag{3.6}$$

3.4.3 Conservation and Tomography

The theoretical anchor of the MDLT algorithm is a conservation law relating observable time intervals to latent volume. Let T_1 and T_2 denote the execution times of probes P_1 and P_2, respectively. Since P_2 cannot execute until all orders preceding it, both visible and hidden have been removed, the total volume depleted from the queue during the interval $[T_1, T_2]$ must exactly equal the volume standing between P_1 and P_2. We define the Observed Depletion, D_{obs}, as the cumulative volume of market orders and cancellations recorded on the public tape at price p between T_1 and T_2. This yields the fundamental conservation equation:

$$D_{obs} = \int_{T_1}^{T_2} (dM_t + dC_t) = V_{gap} + H_{gap} \tag{3.7}$$

rearranging this identity allows us to solve for the unobservable hidden volume H_{gap} in closed form:

$$H_{gap} = D_{obs} - V_{gap} \tag{3.8}$$

From this, we derive the *Iceberg Density Coefficient*, ρ, which quantifies the ratio of hidden to visible liquidity in the local order book:

$$\rho = \frac{H_{gap}}{V_{gap}} = \frac{D_{obs}}{V_{gap}} - 1 \tag{3.9}$$

A value of $\rho \approx 0$ indicates a transparent order book, while $\rho > 0$ signals the presence of iceberg orders. This coefficient is then used to construct the MDLT Priority Metric, Q_{MDLT}, a rigorous estimate of the true effective queue position facing a new limit order:

$$Q_{MDLT}(t) = Q_{visible}(t) \cdot (1 + \bar{\rho}_t) \tag{3.10}$$

where $\bar{\rho}_t$ is an exponentially weighted moving average of the iceberg density, smoothing out microstructure noise. This metric provides a corrected input for optimal execution algorithms, replacing the naive $Q_{visible}$ with a latency-adjusted measure of queue priority.

3.4.4 Probe Order Placement

To actively interrogate the queue structure, we employ a differential latency measurement technique using paired probe orders. Let the current time be t_0. We define a probe pair as a sequence of two limit orders, denoted P_1 and P_2, submitted to the same side of the book (e.g., best bid) with identical unit quantity size $s_p = 1$. The submission mechanism follows a strict temporal discipline:

1. **Probe P_1:** Submitted at time $t_1 = t_0$. Upon acceptance by the matching engine, it is assigned a queue position $q_1 = Q_{true}(t_1, p) + 1$.
2. **Probe P_2:** Submitted at time $t_2 = t_0 + \delta$, where $\delta > 0$ is a deterministic inter-arrival gap. Upon acceptance, it is assigned a queue position $q_2 = Q_{true}(t_2, p) + 1$.

During the interval $(t_1, t_2]$, the queue dynamics continue to evolve. New limit orders may arrive, adding to the visible depth, while hidden orders (icebergs) may also be injected into the queue. We define the *Visible Gap Volume*, V_{gap}, as the cumulative size of all visible limit orders arriving at price p between the two probe submissions:

$$V_{gap} = \sum_k v_k^L \cdot \mathbb{I}_{\{t_1 < t_k^L \leq t_2\}} \tag{3.11}$$

Similarly, let H_{gap} denote the unobservable hidden volume arriving during this same interval. The fundamental geometric relationship between the queue positions of the two probes is thus:

$$q_2 - q_1 = V_{gap} + H_{gap} \tag{3.12}$$

This equation establishes that the "distance" between our probes in the execution queue is strictly equal to the sum of visible and hidden liquidity added during the inter-arrival latency δ.

3.4.5 Iceberg Density Ratio

We invoke the principle of volume conservation to derive the hidden liquidity parameters. Let T_1 and T_2 denote the stochastic execution timestamps of probes P_1 and P_2, respectively. Under the assumption of a FIFO matching algorithm, P_2 executes only after all orders preceding it in the queue have been depleted. Therefore, the total volume removed from the book between T_1 and T_2 must exactly match the queue volume standing between the two probes. We define the Observed Depletion, D_{obs}, as the integral of the order flow depletion rate (market orders and cancellations) over the execution interval $[T_1, T_2]$. Since market data feeds report these trades and cancellations explicitly, D_{obs} is a fully observable quantity:

$$D_{obs} = \int_{T_1}^{T_2} (\lambda^M(t) + \lambda^C(t))\, dt = \sum_j v_j^M \mathbb{I}_{\{T_1 \leq t_j^M \leq T_2\}} + \sum_k v_k^C \mathbb{I}_{\{T_1 \leq t_k^C \leq T_2\}} \tag{3.13}$$

By equating the volume depleted to the volume separating the probes, we obtain the Conservation Law of Queue Tomography:

$$D_{obs} = q_2 - q_1 = V_{gap} + H_{gap} \tag{3.14}$$

This identity allows us to isolate the unknown latent variable H_{gap}. When we send the two probe orders, P_1 and P_2, a certain amount of volume sits between these two probe orders. This particular "gap" is made us of visible orders or V_{gap} as well as hidden orders H_{gap}. Rearranging Equation (3.14), we solve for the hidden volume:

$$H_{gap} = D_{obs} - V_{gap} \tag{3.15}$$

To generalize this finding across different market regimes and asset classes, we define the *Iceberg Density Coefficient*, ρ, as the ratio of hidden volume to visible volume added. This

dimensionless metric normalizes the hidden liquidity relative to the observable order flow:

$$\rho = \frac{H_{gap}}{V_{gap}} \tag{3.16}$$

Substituting the expression for H_{gap}, we arrive at the operational formula for the MDLT estimator:

$$\rho = \frac{D_{obs} - V_{gap}}{V_{gap}} = \frac{D_{obs}}{V_{gap}} - 1 \tag{3.17}$$

Hence, by observing only public data (D_{obs} and V_{gap}), we can recover the scalar parameter ρ that characterizes the hidden depth of the limit order book. A value of $\rho \approx 0$ implies $D_{obs} \approx V_{gap}$, consistent with a fully lit market. Conversely, $\rho > 0$ provides a direct measure of dark liquidity intensity.

The Priority Metric

While the iceberg density coefficient ρ provides an instantaneous snapshot of hidden liquidity, raw measurements derived from individual probe pairs are subject to stochastic microstructure noise, arising from latency jitter and transient liquidity fluctuations. To construct a robust estimator suitable for algorithmic execution, we employ an Exponentially Weighted Moving Averag to smooth the density sequence. Let ρ_k denote the raw density estimate derived from the k-th probe pair. The smoothed density state variable, $\bar{\rho}_k$, evolves according to the recursive filter:

$$\bar{\rho}_k = \alpha\rho_k + (1 - \alpha)\bar{\rho}_{k-1} \tag{3.18}$$

where $\alpha \in (0, 1)$ is the decay factor controlling the memory of the estimator. A higher α increases responsiveness to regime shifts in hidden liquidity usage, while a lower α enhances stability against measurement noise.

We define the *MDLT Priority Metric*, denoted as $Q_{MDLT}(t, p)$, as the effective queue position adjusted for this latent volume. This metric transforms the observable queue depth reported by the exchange into a "virtual" queue depth that reflects the true liquidity barrier facing a new limit order. For a visible queue size $Q_{visible}(t, p)$, the effective position is given by:

$$Q_{MDLT}(t, p) = Q_{visible}(t, p) \cdot (1 + \bar{\rho}_t) \tag{3.19}$$

This formulation implies that for every unit of visible volume, the market participant must anticipate competing against an additional $\bar{\rho}_t$ units of hidden volume. Under the assumption that order arrivals follow a locally stationary Poisson process with depletion

intensity $\lambda_{depletion} = \lambda^M + \lambda^C$, we can derive the expected time-to-fill, $\mathbb{E}[\tau]$, for a newly submitted limit order. Standard queueing theory dictates that the wait time is the ratio of the queue length to the service rate. Substituting our adjusted metric yields:

$$\mathbb{E}[\tau_{fill}] = \frac{Q_{MDLT}(t,p)}{\lambda^M(t) + \lambda^C(t)} = \frac{Q_{visible}(t,p)(1+\bar{\rho}_t)}{\lambda_{depletion}(t)} \tag{3.20}$$

This equation highlights the critical deficiency of naive models: strategies relying solely on $Q_{visible}$ systematically underestimate execution latency by a factor of $(1+\bar{\rho}_t)$, leading to optimal execution schedules that are overly passive and prone to adverse selection. Q_{MDLT} corrects this bias, providing a mathematically consistent basis for execution logic.

Expected Wait Time

We formalize the execution latency, τ, as the first passage time of the cumulative depletion process reaching the order's effective queue position. Let $D(t)$ represent the cumulative volume removed from the queue via market orders and cancellations over the interval $[0,t]$. For a limit order positioned at queue depth Q, the execution time is the stochastic stopping time defined by:

$$\tau(Q) = \inf\{t > 0 : D(t) \geq Q\} \tag{3.21}$$

Under the assumption that the depletion process $D(t)$ follows a compound Poisson process with a constant aggregate intensity $\lambda_{depletion} = \lambda^M + \lambda^C$ and unit volume increments, the expectation of the stopping time is linear with respect to the queue depth. Standard queueing theory yields the first moment:

$$\mathbb{E}[\tau(Q)] = \frac{Q}{\lambda_{depletion}} \tag{3.22}$$

In a market regime characterized by hidden liquidity, utilizing the observable queue depth $Q_{visible}$ yields a *naive* wait time estimator, $\hat{\tau}_{naive}$. However, as derived in the previous section, the true barrier to execution is Q_{MDLT}. Consequently, the corrected MDLT wait time estimator, $\hat{\tau}_{MDLT}$, is given by:

$$\hat{\tau}_{MDLT} = \frac{Q_{MDLT}}{\lambda^M + \lambda^C} = \frac{Q_{visible}(1+\bar{\rho})}{\lambda^M + \lambda^C} \tag{3.23}$$

The discrepancy between these two estimators represents the *Hidden Latency Bias*. We can express the relationship between the true and naive expectations as:

$$\hat{\tau}_{MDLT} = \hat{\tau}_{naive} \cdot (1+\bar{\rho}) \tag{3.24}$$

This multiplicative relationship highlights the non-linear risk of ignoring iceberg orders. In regimes where $\bar{\rho} \approx 1$ (hidden volume equals visible), the naive model underestimates the time-to-fill by 50%. Such underestimation directly impacts optimal execution logic, particularly for Almgren-Chriss style trajectories, where the estimated variance of execution cost is a function of time. By substituting $\hat{\tau}_{MDLT}$ into the cost function, traders can accurately price the risk of "resting" in the queue versus paying the spread, thereby minimizing the implementation shortfall caused by unexpected delays.

3.5 Regime Normalization

A critical challenge in latency tomography is decoupling the structural properties of the queue (depth) from the stochastic intensity of the arrival process (speed).

3.5.1 Corrected Latency Normalization

The raw execution latency $T = T_2 - T_1$ is inversely proportional to the queue depletion rate. Consequently, a decrease in T could ambiguously signal either a shallower queue or a surge in market aggressiveness. To resolve this ambiguity, we control for the Order Flow Imbalance, which acts as a good measure for short-term buying or selling pressure. We define the Order Flow Imbalance over an interval Δt as the net flow of liquidity demanding events:

$$OFI_t = \sum_{t-\Delta t<s\leq t} v_s^M \cdot \mathbb{I}_{\{dir_s=buy\}} - \sum_{t-\Delta t<s\leq t} v_s^M \cdot \mathbb{I}_{\{dir_s=sell\}} \tag{3.25}$$

High-magnitude OFI regimes are characterized by elevated arrival intensities $\lambda^M(t)$, which systematically bias raw latency measurements downward. To isolate the queue depth contribution, we introduce the Normalized Latency, τ_{norm}. This metric rescales the raw time-domain measurement into "volume-time" units, effectively normalizing for the varying speed of market depletion:

$$\tau_{norm} = (T_2 - T_1) \cdot \left(\hat{\lambda}^M + \hat{\lambda}^C\right) \tag{3.26}$$

By multiplying the time duration by the estimated depletion intensity, τ_{norm} approximates the total volume processed by the market during the probe interval. Unlike raw latency, this quantity is invariant to changes in trading tempo and provides a more stable basis for estimating the effective queue size Q_{MDLT} across different volatility regimes.

3.5.2 Regime-Dependent Density

Empirical evidence suggests that the presence of iceberg orders is not uniform but highly state-dependent. Institutional algorithms tend to vary their concealment logic based on market urgency and volatility. Therefore, a global average $\bar{\rho}$ may lack the specificity required for precision execution. To address this, we adopt a regime-switching framework conditioned on the OFI distribution. We partition the trading day into K distinct regimes based on the quintiles of the OFI distribution, denoted as $\mathcal{R}_k$ for $k \in \{1, \dots, 5\}$. We maintain separate exponentially weighted moving averages for the iceberg density coefficient within each regime. Let $\bar{\rho}^{(k)}$ represent the density estimator specific to the k-th OFI quintile. The update rule is applied conditionally:

$$\bar{\rho}_t^{(k)} = \begin{cases} \alpha \rho_t + (1-\alpha)\bar{\rho}_{t-1}^{(k)} & \text{if } OFI_t \in \mathcal{R}_k \\ \bar{\rho}_{t-1}^{(k)} & \text{otherwise} \end{cases} \tag{3.27}$$

The final Priority Metric is then constructed dynamically by selecting the density coefficient corresponding to the current market regime:

$$Q_{MDLT}(t) = Q_{visible}(t) \cdot \left(1 + \sum_{k=1}^{5} \mathbb{I}_{\{OFI_t \in \mathcal{R}_k\}} \bar{\rho}_{t-1}^{(k)}\right) \tag{3.28}$$

This stratified approach allows the MDLT algorithm to adapt to changing market microstructures, applying a higher "hidden liquidity penalty" in regimes known to feature heavy iceberg usage (e.g., low-volatility accumulation periods) while relaxing the penalty in high-velocity trends where liquidity is predominantly visible.

3.5.3 Real-Time Intensity

Algorithm 1 Estimate Order Flow Intensities

Input: Live market feed $\{(t_i, \text{type}_i, v_i, p_i)\}_{i=1}^{n}$, lookback window T_{win}
Output: Intensity vector $[\lambda_L, \lambda_C, \lambda_M]$
Initialize counters: $N_L \leftarrow 0, N_C \leftarrow 0, N_M \leftarrow 0$
for each event i in feed **do**
    if $t_{\text{now}} - t_i < T_{\text{win}}$ **then**
        if $\text{type}_i =$ "Limit" **then**
            $N_L \leftarrow N_L + 1$
        else if $\text{type}_i =$ "Cancel" **then**
            $N_C \leftarrow N_C + 1$
        else if $\text{type}_i =$ "Trade" **then**
            $N_M \leftarrow N_M + 1$
        end if
    end if
end for
$\lambda_L \leftarrow N_L / T_{\text{win}}$
$\lambda_C \leftarrow N_C / T_{\text{win}}$
$\lambda_M \leftarrow N_M / T_{\text{win}}$
Return $[\lambda_L, \lambda_C, \lambda_M]$

3.5.4 Probe Pair Execution with Tomographic Scan

Algorithm 2 MDLT Probe Pair Execution

Input: Best bid price p^*, gap δ (ms), quantity $q = 1$
Output: $(T_1, T_2, D_{\text{obs}}, V_{\text{gap}})$
Observe $Q_{\text{visible}} \leftarrow$ current LOB depth at p^*
Submit P_1: Limit Buy, Qty=1, Price=p^*
Wait for fill, record $T_1 \leftarrow$ execution timestamp
δ milliseconds
P_2 Limit Buy, Qty=1, Price=p^*
Wait for fill, record $T_2 \leftarrow$ execution timestamp
Scan market tape during $[T_1, T_2]$:
    $D_{\text{obs}} \leftarrow 0$
for each event e in $[T_1, T_2]$ **do**
    if type(e) = "Trade" and price(e) = p^* **then**
        $D_{\text{obs}} \leftarrow D_{\text{obs}} + \text{volume}(e)$
    else if type(e) = "Cancel" and price(e) = p^* **then**
        $D_{\text{obs}} \leftarrow D_{\text{obs}} + \text{volume}(e)$
    end if
end for
Calculate visible adds:
    $V_{\text{gap}} \leftarrow$ sum of Limit orders at p^* during $(0, \delta)$
Compute $\rho \leftarrow (D_{\text{obs}}/V_{\text{gap}}) - 1$
Update rolling average:
    $\rho_{\text{smooth}} \leftarrow 0.9 \times \rho_{\text{smooth}} + 0.1 \times \rho$
Return $(T_1, T_2, D_{\text{obs}}, V_{\text{gap}})$

3.5.5 Priority Metric Calculation

Algorithm 3 Compute Q_MDLT

Input: Q_{visible}, ρ_{smooth}, $[\lambda_L, \lambda_C, \lambda_M]$
$Q_{\text{MDLT}} \leftarrow Q_{\text{visible}} \times (1 + \rho_{\text{smooth}})$
$\mu_{\text{depletion}} \leftarrow \lambda_M + \lambda_C$
$\mathbb{E}[\tau] \leftarrow Q_{\text{MDLT}}/\mu_{\text{depletion}}$
Return $(Q_{\text{MDLT}}, \mathbb{E}[\tau])$

3.6 Numerical Example

In this section, we ground the abstract principles developed thus far in a concrete market scenario. We present a detailed worked example showing how the tomographic measurement principle operates in practice, from the submission of probe orders through the calculation of hidden liquidity and the implications for execution strategy. This example is not merely illustrative; it demonstrates the mechanical operation of the MDLT framework and validates the claim that passive observation of the order book leaves critical information hidden.

3.6.1 Scenario

We consider a liquid equity market at mid-morning trading hours, when volatility is moderate and order flow is predictable. The conditions are as follows:

Table 3.1: Market Conditions at Probe Submission Time

Parameter	**Value**
Security	Apple Inc. (AAPL)
Best Bid Price	\$100.00
Best Ask Price	\$100.01
Bid-Ask Spread	\$0.01 (1 cent)
Visible Queue Depth at Bid	500 shares
Market Time	10:30:00.000 (mid-morning)
Market Regime	Moderate volatility, normal activity

The visible queue of 500 shares represents limit buy orders placed at the best bid price of \$100.00. These are the orders that any market participant can observe through the public order book feed. However, as discussed in Section **??**, this visible depth likely understates the true queue depth because of iceberg orders. Our goal is to measure this hidden component through active probing.

3.6.2 Probe Sequence and Execution Timeline

We now trace the sequence of events as our two probe orders proceed through the matching engine. Each probe is a limit buy order of unit size (one share) submitted to the best bid price. The temporal spacing between submissions is critical: it defines the window over which we will observe queue dynamics.

Table 3.2: Probe Order Timeline: Submission and Execution

Time (HH:MM:SS.mmm)	**Event**	**Details**
10:30:00.000	Submit P_1	Limit Buy 1 share @ \$100.00
10:30:00.025	Submit P_2	Limit Buy 1 share @ \$100.00 ($\delta = 25$ ms)
10:30:00.058	P_1 executes	Execution time $T_1 = 58$ ms after submission
10:30:00.087	P_2 executes	Execution time $T_2 = 87$ ms after P_1 submission

The inter-probe gap is $\delta = 25$ milliseconds. This gap is chosen to be long enough to allow meaningful market activity (new limit orders, cancellations, market orders) to occur between submissions, but short enough that market regime (volatility, order flow intensity) remains approximately stationary. The execution times $T_1 = 58$ ms and $T_2 = 87$ ms reflect the time elapsed from the initial submission of P_1 until each probe fills.

The key observation is that P_1 and P_2 do not execute instantaneously. Each must wait for all orders ahead of it in the FIFO queue to be removed through either market order execution or cancellation. The wait time for P_1 is 58 milliseconds. By the time P_2 executes, an additional 29 milliseconds have passed. This additional waiting time encodes information about the queue state at the moment P_2 was submitted.

3.6.3 Inter-Execution Market Activity

Between the execution of P_1 (at 58 ms) and the execution of P_2 (at 87 ms), the order book is not quiescent. Market orders arrive and execute against standing limit orders. Some traders cancel their orders. The public market tape records all of these events. We now enumerate what occurred during this 29-millisecond interval.

Table 3.3: Market Tape Events in the Interval $[T_1, T_2]$ (Execution Interval)

Event Type at \$100.00 Bid	**Volume (shares)**	**Cumulative Volume**
Market Sell @ \$100.00	80	80
Market Sell @ \$100.00	120	200
Cancel (Limit Order) @ \$100.00	50	250
Market Sell @ \$100.00	90	340
Market Sell @ \$100.00	60	400
Total Volume Removed		**400**

The table above represents the complete market activity at the best bid price during the execution interval. A market sell is an aggressive order that executes immediately against

the best standing bid, removing shares from the queue. A cancellation is a limit order withdrawal, also removing shares from the queue but not resulting in a transaction.
We aggregate across event types to obtain the total observed depletion:

$$D_{\text{obs}} = (\text{market orders executed}) + (\text{limit orders cancelled}) = 350 + 50 = 400 \text{ shares} \quad (3.29)$$

This quantity D_{obs} is fully observable from the market data feed. Every trade is time-stamped and reported. Every cancellation is announced to the market. Therefore, $D_{\text{obs}} = 400$ shares is a fact, not an estimate or inference.

3.6.4 Observable Queue Additions

While market activity removes volume from the queue during the interval $[T_1, T_2]$, other market participants are adding volume to the queue. Specifically, new limit orders arrive at the best bid price after P_1 is submitted but before P_2 executes. These arrivals are equally observable from the market data feed. We define the gap volume as the cumulative size of all limit orders that arrive at the best bid price during the inter-probe interval $[0, \delta]$, where time zero is the submission of P_1 and time $\delta = 25$ ms is the submission of P_2:

Table 3.4: Limit Order Arrivals During the Probe Gap $[0, \delta]$

Time (HH:MM:SS.mmm)	**Event: Limit Buy Arrivals at $100.00**
10:30:00.005	Arrival of 20 shares
10:30:00.018	Arrival of 30 shares
Total Gap Volume	**50 shares**

These arrivals represent new buy-side limit orders placed at the best bid price. They become part of the queue at the bid price, appearing in the public order book for all market participants to see. Thus, the gap volume $V_{\text{gap}} = 50$ shares is also fully observable.

3.6.5 Conservation Principle

We now invoke the conservation principle introduced in Section **??**. This principle states that the volume removed from the queue between the execution times of the two probes must equal the distance separating those probes in the queue. Formally, the distance between P_1 and P_2 in the execution queue is the sum of two components: the visible volume that arrived between their submission times, plus any hidden volume from iceberg orders:

$$\text{Distance between } P_1 \text{ and } P_2 = V_{\text{gap}} + H_{\text{gap}} \tag{3.30}$$

Here, V_{gap} is the observable gap volume (which we computed as 50 shares), and H_{gap} is the unobservable hidden volume from iceberg orders in the same interval.

Now, a fundamental fact about FIFO queue discipline: an order cannot execute until all orders ahead of it have been removed. When P_2 executes at time T_2, this means all volume separating P_1 from P_2 must have been depleted between the execution times T_1 and T_2. The volume depleted is precisely what we observe from the market tape: $D_{\text{obs}} = 400$ shares.

By conservation:

$$D_{\text{obs}} = V_{\text{gap}} + H_{\text{gap}} \tag{3.31}$$

Rearranging to solve for the hidden component:

$$H_{\text{gap}} = D_{\text{obs}} - V_{\text{gap}} = 400 - 50 = 350 \text{ shares} \tag{3.32}$$

This is the key result. Between the submission of our two probes, hidden iceberg orders concealed 350 shares of volume. This volume was never visible in the public order book, yet it constrained execution, added to the effective queue depth, and affected the execution dynamics of any trader trying to execute at the best bid.

3.6.6 Iceberg Density Estimation

We now normalize the hidden volume relative to the visible volume to create a dimensionless measure of iceberg intensity. The iceberg density coefficient ρ is defined as the ratio of hidden to visible volume:

$$\rho = \frac{H_{\text{gap}}}{V_{\text{gap}}} \tag{3.33}$$

Equivalently, substituting our expression for H_{gap}:

$$\rho = \frac{D_{\text{obs}}}{V_{\text{gap}}} - 1 \tag{3.34}$$

In our numerical example:

$$\rho = \frac{400}{50} - 1 \tag{3.35}$$

$$= 8 - 1 \tag{3.36}$$

$$= 7.0 \tag{3.37}$$

This result indicates that for every one share of visible liquidity in this interval, seven shares of hidden liquidity existed. Stated differently, the hidden volume is 700% of the visible volume, or equivalently, the true queue is eight times deeper than the visible queue suggests.

On Iceberg Density

An iceberg density of 7.0 is high, indicating unusually heavy use of hidden orders during this interval. In normal market conditions, typical values of ρ range from 0.2 to 0.6, indicating that hidden volume is 20% to 60% of visible volume. The elevated value in our scenario suggests one of several possibilities: (a) a large institutional investor is executing a significant block trade and has hidden most of their order; (b) market makers are using iceberg orders to manage inventory risks during a volatile period; or (c) the visible queue is unusually shallow due to earlier trading activity, making hidden orders appear more prominent.

The interpretation is straightforward: $\rho = 0$ would mean the order book is fully transparent, with no hidden liquidity. $\rho > 0$ indicates the presence of iceberg orders. Higher values of ρ indicate heavier reliance on concealment strategies.

3.6.7 Adjusted Queue Position

Having measured the iceberg density from our probe pair, we can now apply this information to refine our understanding of the queue depth at subsequent times. Suppose that at time $t = 87$ ms (the moment when P_2 executes), a trader wishes to submit a new large limit order at the same price level. The trader observes from the public order book that the visible queue depth is $Q_{\text{visible}} = 500$ shares.

If the trader naively assumes that this visible depth is the true queue depth, they will make execution decisions under the assumption that the queue is shallow. However, our probe measurement has just revealed that during the recent interval, the iceberg density was $\rho = 7.0$. Assuming this density persists (an assumption we will refine in Chapter Two through smoothing), the true effective queue depth is:

$$Q_{\text{MDLT}} = Q_{\text{visible}} \times (1 + \rho) \tag{3.38}$$
$$= 500 \times (1 + 7.0) \tag{3.39}$$
$$= 500 \times 8 \tag{3.40}$$
$$= 4000 \text{ shares} \tag{3.41}$$

The MDLT metric adjusts the visible queue by the factor $(1 + \rho)$ to account for hidden liquidity. In this case, the adjustment is substantial: a visible queue of 500 shares becomes an effective queue of 4000 shares. This adjustment captures the intuition that hidden icebergs act as additional layers of queueing depth, even though they are not visible.

3.6.8 Wait Time Estimation

With an adjusted queue position in hand, we can now estimate expected execution times using queueing theory. Recall that under the M/M/1 queue model, the expected wait time for an order at queue position Q is

$$\mathbb{E}[\tau] = \frac{Q}{\lambda_M + \lambda_C} \tag{3.42}$$

where λ_M is the rate of market order arrivals and λ_C is the rate of cancellations (both measured in shares per second). For our scenario, we estimate from recent market data that the combined depletion rate is $\lambda_M + \lambda_C = 50$ shares per second.

Naive Estimate

A trader who observes only the visible queue would estimate:

$$\mathbb{E}[\tau_{\text{naive}}] = \frac{Q_{\text{visible}}}{\lambda_M + \lambda_C} \tag{3.43}$$
$$= \frac{500}{50} \tag{3.44}$$
$$= 10 \text{ seconds} \tag{3.45}$$

This estimate suggests that the queue will clear in 10 seconds, a reasonable wait time. Based on this estimate, the trader might decide that joining the queue at the best bid is preferable to paying the spread through a market order.

Multidimensional Latency Estimates

Our measurement, however, reveals a different picture:

$$\mathbb{E}[\tau_{\text{MDLT}}] = \frac{Q_{\text{MDLT}}}{\lambda_M + \lambda_C} \tag{3.46}$$

$$= \frac{4000}{50} \tag{3.47}$$

$$= 80 \text{ seconds} \tag{3.48}$$

$$\approx 1.3 \text{ minutes} \tag{3.49}$$

The MDLT estimate suggests that the order will wait approximately 80 seconds—a much longer duration. This dramatic difference arises entirely from the hidden liquidity revealed by our probes.

Decision Rule

The trader now faces a different calculus. An 80-second wait exposes the position to significant price risk. If the market price moves by even a few cents against the position during that wait, the cost of the move will exceed the spread savings from joining the limit order queue. The decision rule might be structured as follows:

First, we classify order sizes into categories based on the expected wait time and associated risks:

(a) **Small Orders** ($N < 100$ shares): Even with an 80-second wait, the order is small enough that it likely clears quickly from the queue. The decision is to join the queue at the best bid. Expected wait time is less than 1-2 seconds even after MDLT adjustment.

(b) **Medium Orders** ($100 \leq N \leq 1000$ shares): The wait time becomes material. The trader should consider a time-weighted average price (TWAP) algorithm that spreads the execution across a longer time horizon (e.g., 10-15 minutes), reducing the impact of any single segment of the order joining the queue at a given moment.

(c) **Large Orders** ($N > 1000$ shares): The wait time in a queue with depth equivalent to 4000 shares is prohibitive. The trader is better served by using market orders (paying the spread immediately) or seeking out hidden liquidity pools and alternative trading venues where the queue structure may be different.

You Can't Ignore Me

To quantify the cost of ignoring the hidden liquidity, consider a specific scenario. Suppose the trader places a 1000-share order at the best bid, intending to wait for execution. Under the naive model, the trader expects execution in 20 seconds (1000/50). However, the MDLT model reveals the true wait time is 160 seconds. During that additional 140-second wait, the market price might move. If the midpoint price rises by just 0.05 (five cents), the trader loses $1000 \times 0.05 = \$50$ due to the price move, an amount that vastly exceeds the \$0.01 spread savings from using a limit order. Conversely, if the trader had used the MDLT measurement to inform the execution strategy, they might have chosen to (a) submit smaller segments of the order across multiple price levels, (b) access hidden liquidity through alternative venues, or (c) use market orders to ensure immediate execution at a known price. Each of these alternatives protects against the risk of unexpected price movement during the wait.

3.7 Risk Analysis

Every active measurement carries an economic cost. Unlike passive inference, which requires only data observation, active probing requires sending orders into the market. These orders must execute to generate the signal we need, and execution incurs trading costs. Understanding and managing these costs is critical to ensuring that the value of measurement exceeds its price.The MDLT framework provides a principled approach to measuring hidden queue depth, yet like all measurement systems operating in complex environments, it is subject to costs, model assumptions, and failure modes. This section systematically examines these constraints and proposes mitigation strategies. Understanding these limitations is essential: a robust measurement system is one that explicitly acknowledges where it may fail and implements safeguards accordingly.

Spread Cost Per Probe Pair

The direct cost of submitting a probe pair arises from the bid-ask spread. When we submit a limit buy order at the best bid price, it executes at that bid price. We thus "pay" the full bid-ask spread in the sense that we sell to the market at the bid price, which is lower than the contemporaneous ask price. For a probe order of unit size (one share), the cost is the spread itself:

$$C_{\text{probe pair}} = 2 \times \frac{s}{2} = s \tag{3.50}$$

where s denotes the bid-ask spread. Each of our two probes costs half the spread (since we execute at the bid and the ask midpoint is halfway between bid and ask). Summing both

probes yields a total cost equal to the full spread. For highly liquid securities such as AAPL, which typically trade with spreads of one penny, the cost per probe pair is:

$$C_{\text{probe pair}} = \$0.01 \tag{3.51}$$

This cost is minimal in absolute terms. However, for the measurement to create positive economic value, the information gain must justify this cost. We therefore define a break-even condition.

Break-Even Analysis

The information extracted from a probe pair is valuable only if it prevents greater losses in the subsequent main order execution. Let Δ_{slippage} denote the per-share reduction in slippage (measured in dollars per share) that results from using MDLT-informed execution versus naive execution. For a main order of size N shares, the total benefit from improved execution is:

$$\text{Benefit} = N \times \Delta_{\text{slippage}} \tag{3.52}$$

For the measurement to be economical, the benefit must exceed the cost:

$$N \times \Delta_{\text{slippage}} > C_{\text{probe pair}} \tag{3.53}$$

Rearranging to solve for the break-even order size:

$$N_{\text{break-even}} = \frac{C_{\text{probe pair}}}{\Delta_{\text{slippage}}} = \frac{s}{\Delta_{\text{slippage}}} \tag{3.54}$$

To make this concrete, consider a realistic scenario. Suppose accurate queue depth measurement prevents one basis point (0.01%) of slippage per share. For AAPL trading at approximately \$150 per share, one basis point is $150 \times 0.0001 = \$0.015$ per share. With a probe cost of \$0.01 per pair:

$$N_{\text{break-even}} = \frac{0.01}{0.015} \approx 667 \text{ shares} \tag{3.55}$$

Alternatively, if we estimate more conservatively that MDLT prevents 1 basis point of slippage in dollar terms (not percentage terms), then:

$$N_{\text{break-even}} = \frac{0.01}{0.0001} = 100 \text{ shares} \tag{3.56}$$

In practice, institutional investors executing orders of 100 to 10,000 shares are common in equity markets. The break-even threshold of 100–700 shares is well within the range of

institutional order sizes. For smaller retail orders (fewer than 100 shares), the measurement cost exceeds the likely benefit. For institutional orders, the measurement is economical.

Probe Non-Execution

A subtle but important risk arises if probe orders fail to execute promptly. Our methodology assumes that both P_1 and P_2 execute within a short time window (typically tens to hundreds of milliseconds). If the market price moves away from the best bid during the measurement interval, our limit buy orders will sit unfilled in the queue without contributing to the measurement signal. Specifically, if the security's price rises above our limit bid price (e.g., if the best bid moves from \$100.00 to \$100.01), our limit orders become "out of the money" and will not execute until the price falls back. This creates two problems. First, we have submitted orders but received no signal; the measurement is incomplete. Second, if the price does later drop back, our old orders may execute far later than intended, at a time when market conditions have changed and the measurement signal has become stale.nTo mitigate this risk, we recommend using immediate-or-cancel (IOC) orders for probes rather than persistent limit orders. An IOC probe is a limit order that executes any portion that matches immediately, and any remainder is automatically cancelled. We would typically set a timeout window (e.g., 200 milliseconds) within which the probe must execute. If it does not execute within that window, it is cancelled, and we attempt a fresh probe in the next measurement cycle. The tradeoff is that IOC probes may not execute at all if market conditions are adverse (e.g., large spreads, shallow depth). In that case, we obtain no measurement signal. However, a non-signal in an adverse market regime is arguably more informative than a delayed signal that reflects stale conditions. We recommend monitoring the probe execution rate: if the fraction of probe pairs that execute drops below 80%, this indicates either a regime change (wider spreads, lower liquidity) or technical issues with order submission, both of which warrant immediate recalibration.

Chapter 4

Effective Liquidity Imbalance

> "Depth is the amount of order flow required to change prices by a given amount...
> A market is liquid if it is deep." — Albert S. Kyle,1985

In the previous chapter, we established that a single probe pair ($N = 1$) acts as a scalar measurement of the instantaneous effective queue depth, Q_{MDLT}. However, we posit that liquidity is not a static variable; it is a dynamic process characterized by flow. A single measurement tells us where the wall is, but not if it is moving, crumbling, or reinforcing itself. To capture the dynamics of hidden liquidity, specifically its decay and replenishment rate; we must extend the tomographic framework from a single snapshot to a temporal sequence. By utilizing a Triple Probe Pair ($N = 3$) structure, we can measure not just the state of the order book, but its first and second derivatives with respect to time. This allows us to calculate a metric we define as the Effective Liquidity Imbalance (ELI), that predicts near-term order book stability.

4.1 Liquidity Kinematics

We define a "Probe Pair" $\mathcal{P}_k$ as a set of two atomic orders sent to the same price level p with a micro-separation ϵ. The execution latency difference ΔT_k reveals the instantaneous intensity λ_k and effective depth Q_k. To reconstruct such a dynamic state, we deploy a sequence of three probe pairs spaced by a sampling interval δ:

1. $\mathcal{P}_1$ at t_0: Returns effective depth $Q(t_0)$.
2. $\mathcal{P}_2$ at $t_0 + \delta$: Returns effective depth $Q(t_0 + \delta)$.
3. $\mathcal{P}_3$ at $t_0 + 2\delta$: Returns effective depth $Q(t_0 + 2\delta)$.

Using the MDLT estimator derived in Equation (2.19), we obtain a discrete sequence of observations:

$$\mathbf{Q}_{obs} = [Q_{MDLT}(t_0), Q_{MDLT}(t_0 + \delta), Q_{MDLT}(t_0 + 2\delta)]$$

In the preceding chapter, we established that a single probe pair provides a scalar measurement of the instantaneous effective queue depth, Q_{MDLT}. However, we posit that liquidity is not a static variable; rather, it is a stochastic process characterized by continuous flow. A single measurement reveals the location of the resistance wall, but it remains silent on whether that wall is reinforcing itself through hidden replenishment or crumbling under the weight of toxic flow. To capture the dynamics of hidden liquidity, specifically its rates of decay and regeneration, we must extend the tomographic framework from a static snapshot to a temporal sequence.

4.1.1 The Latent Liquidity State Function

We model the true resistance of the limit order book not as a discrete set of quantities, but as a continuous, twice-differentiable function of time,

$$\mathcal{L} : \mathbb{R}^+ \to \mathbb{R}^+$$

. The value $\mathcal{L}(t)$ represents the effective depth (visible as well as hidden) available to absorb market orders at time t. The fundamental epistemological challenge inherent in market microstructure is that $\mathcal{L}(t)$ is unobservable. The market data feed provides only a discrete, noisy sampling of the visible component, often masking the underlying trends of institutional inventory management. To reconstruct the trajectory of the total liquidity function, we invoke the principle of local polynomial approximation. For a sufficiently small time interval δ, the evolution of any smooth physical system can be approximated by its Taylor series expansion. By expanding $\mathcal{L}(t)$ around a reference time t_0, we obtain a local approximation of the liquidity surface

$$\mathcal{L}(t) \approx \mathcal{L}(t_0) + \mathcal{L}'(t_0)(t - t_0) + \frac{1}{2}\mathcal{L}''(t_0)(t - t_0)^2 + \mathcal{O}((t - t_0)^3) \tag{4.1}$$

This extensive expansion reveals that the state of liquidity is governed by three primary components, which we define as the Liquidity Moments. $\mathcal{L}(t_0)$ represents the instantaneous position or depth; $\mathcal{L}'(t_0)$ represents the velocity, or the rate of change in depth; and $\mathcal{L}''(t_0)$ represents the acceleration, or the curvature of the liquidity function. To solve for these three unknowns, a single measurement point is mathematically insufficient. We require a system of at least three data points to fit the parabolic curve defined by the second-order expansion, necessitating the development of the Triple Probe Pair system.

4.1.2 Triple Probe Pair

We define a "Triple Probe Pair" sequence as a set of three distinct tomographic measurements, $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3$, submitted at times t_0, $t_0 + \delta$, and $t_0 + 2\delta$ respectively. Using the MDLT estimator derived in Chapter 3, these probes yield a vector of observed effective depths

$$\mathbf{Q}_{obs} = [Q_{MDLT}(t_0),\ Q_{MDLT}(t_0 + \delta),\ Q_{MDLT}(t_0 + 2\delta)]^T$$

. By aligning these discrete observations with the continuous Taylor expansion, we can isolate the kinematic properties of the queue through finite difference methods.

Liquidity Velocity

The first derivative of the liquidity state function, $\mathcal{L}'(t)$, represents the net flow of volume into or out of the price level. It captures the struggle between liquidity consumption (driven by market orders and cancellations) and liquidity provision (driven by new limit orders and iceberg refills).

Proposition 4.1 (Liquidity Velocity)**.** *Let Q_1 and Q_2 be the effective depths measured at t_0 and $t_0 + \delta$. The Liquidity Velocity, denoted $\hat{v}_{liq}$, is the discrete estimator of the first derivative $\mathcal{L}'(t_0)$, given by the forward difference:*

$$\hat{v}_{liq}(t_0) \equiv \frac{Q_2 - Q_1}{\delta} \tag{4.2}$$

The interpretation of $\hat{v}_{liq}$ provides the first layer of directional signal . A negative velocity indicates a state of Net Depletion, where the rate of consumption exceeds the rate of replenishment, signaling a potentially collapsing queue. Conversely, a positive velocity indicates Net Replenishment, suggesting that hidden liquidity is refilling the level faster than aggressive orders can deplete it. This metric allows the algorithm to distinguish between a static queue and one that is actively being reinforced or dismantled.

Liquidity Acceleration

While velocity indicates the direction of liquidity flow, it does not distinguish between a sustainable trend and a transient shock. To separate panic-induced withdrawals from resilient market making, we must measure the curvature of the liquidity function, formalized as the second derivative.

Proposition 4.2 (Liquidity Acceleration)**.** *Let Q_1, Q_2, and Q_3 be the effective depths measured at intervals of δ. The Liquidity Acceleration, denoted $\hat{a}_{liq}$, is the discrete estimator*

of the second derivative $\mathcal{L}''(t_0)$, given by the second-order finite difference:

$$\hat{a}_{liq}(t_0) \equiv \frac{Q_3 - 2Q_2 + Q_1}{\delta^2} \tag{4.3}$$

Acceleration here serves as a crucial convexity metric for the order book, distinguishing between two distinct market regimes. The first regime is that of an Accelerating Collapse, characterized by negative velocity and negative acceleration ($\hat{v}_{liq} < 0, \hat{a}_{liq} < 0$). In this scenario, the queue is decaying at an increasing rate, often a signature of a "liquidity vacuum" where market makers pull quotes in anticipation of an adverse price shift. The second regime is that of an Elastic Defense, where velocity is negative but acceleration is positive ($\hat{v}_{liq} < 0, \hat{a}_{liq} > 0$). Here, although the queue is shrinking, the rate of decay is slowing. This convexity implies that as the visible depth collapses, new volume is stepping in to stabilize the level, signaling strong hidden support and "impact resilience." By projecting this quadratic model forward, we can calculate the Time to Liquidity Failure (TTLF), providing a superior metric for execution timing than static queue size alone.

4.1.3 An Intuitive Example

To operationalize the kinematic principles derived above, we consider a representative market microstructure scenario involving a Market Maker (henceforth, MM) providing liquidity at the Best Bid. Suppose the public order book displays a static visible queue of 500 shares. To the passive observer, this liquidity appears constant and stable. However, an execution algorithm must determine whether this depth is real, i.e. it is supported by substantial iceberg reserves or alternately fragile, representing a facade that will crumble under execution pressure. To resolve this ambiguity, we deploy the Triple Probe Pair sequence with an inter-arrival latency of $\delta = 50ms$. The resulting measurements of effective depth allow us to distinguish between two radically different liquidity regimes.

4.1.4 The Accelerating Collapse

In the first scenario, our probe sequence reveals a rapid deterioration of the liquidity surface. At time $t = 0ms$, the initial probe detects a robust effective depth of $Q_1 = 2500$ shares, indicating strong hidden backing. However, at $t = 50ms$, the second measurement yields $Q_2 = 2000$ shares, representing a depletion of 500 shares and a negative velocity of -10 shares per millisecond. Crucially, at $t = 100ms$, the third measurement returns $Q_3 = 1000$ shares. While the visible quote remains effectively unchanged to the public, the hidden liquidity has plummeted by an additional 1000 shares. Calculating the liquidity acceleration

reveals the structural weakness of this queue:

$$\hat{a}_{liq} = \frac{1000 - 2(2000) + 2500}{50^2} = \frac{-500}{2500} = -0.2 \text{ shares}/ms^2 \tag{4.4}$$

The negative acceleration coefficient ($\hat{a}_{liq} < 0$) serves as a quantitative signature of a "panic withdrawal." It indicates that the rate of liquidity consumption is not constant but increasing; the market maker is actively pulling hidden backing in anticipation of an adverse price move. The static visible quote of 500 shares is effectively a mirage, masking a collapsing interior. Consequently, this kinematic profile generates an **Aggressive Sell** signal, advising the algorithm to execute immediately at market rather than resting in a deteriorating queue.

4.1.5 An Elastic Defense

In the second scenario, the initial conditions appear identical: $Q_1 = 2500$ at $t = 0ms$, followed by a drop to $Q_2 = 2000$ at $t = 50ms$. The divergence occurs in the third interval. At $t = 100ms$, the effective depth is measured at $Q_3 = 1900$ shares. While the queue is still shrinking, the magnitude of the loss has slowed dramatically, dropping only 100 shares in the second interval compared to 500 in the first.
The acceleration metric captures this stabilization:

$$\hat{a}_{liq} = \frac{1900 - 2(2000) + 2500}{2500} = \frac{400}{2500} = +0.16 \text{ shares}/ms^2 \tag{4.5}$$

Here, the positive acceleration ($\hat{a}_{liq} > 0$) indicates convexity in the liquidity function. Although the net flow is still negative, the deceleration of decay suggests "impact resilience." This dynamic implies that as the queue is hit, new iceberg orders are being reloaded to absorb the selling pressure, effectively hardening the support level. Unlike the first scenario, this profile suggests a sustainable floor. The resulting strategic signal is a **Passive Buy**, validating the decision to join the bid and capture the spread, protected by the verified hidden depth behind the order.

4.2 Generalized Extensions

In our humble opinion, the kinematic framework established in the previous section need not be limited to quadratic approximations derived from a Triple Probe Pair. The logic of measuring liquidity derivatives can be generalized to an arbitrary set of N probe pairs (comprising $2N$ atomic orders) to reconstruct the liquidity surface with higher-order fidelity. By increasing the sampling resolution, we aim to move beyond simple velocity and acceleration to capture higher-moment dynamics such as "jerk", which often signals abrupt regime shifts in

high-frequency market making algorithms. Specifically, let $\mathbf{T} = \{t_0, t_0 + \delta, \ldots, t_0 + (N-1)\delta\}$ denote the discrete transmission times of N sequential probe pairs. Correspondingly, let $\mathbf{Q} = [Q_1, Q_2, \ldots, Q_N]^T$ be the vector of observed effective depths returned by the MDLT estimator at each timestamp. We seek to approximate the true Liquidity State Function $\mathcal{L}(t)$ via a polynomial of degree $K = N - 1$. This approximation takes the form:

$$\mathcal{L}(t) = \sum_{k=0}^{K} c_k (t - t_0)^k \tag{4.6}$$

To solve for the unknown coefficients c_k, which encode the structural properties of the queue, we formulate the observations as a linear system. Since the sampling intervals are deterministic multiples of δ, the system maps to a classic Vandermonde matrix structure, $\mathbf{Vc} = \mathbf{Q}$:

$$\begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & \delta & \delta^2 & \ldots & \delta^{N-1} \\ 1 & 2\delta & (2\delta)^2 & \ldots & (2\delta)^{N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & (N-1)\delta & ((N-1)\delta)^2 & \ldots & ((N-1)\delta)^{N-1} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{N-1} \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \\ \vdots \\ Q_N \end{bmatrix} \tag{4.7}$$

Solving for the coefficient vector $\mathbf{c} = \mathbf{V}^{-1}\mathbf{Q}$ yields the coefficients of the Taylor expansion of the liquidity function. From the definition of the Taylor series, these coefficients are directly linked to the derivatives of the liquidity state at time t_0. Specifically, the k-th derivative is given by $k!c_k$. This inversion allows us to extract the complete kinematic profile of the order book, namely (a) c_0, the Instantaneous Depth (State), equivalent to the static MDLT metric. (b) c_1 the Liquidity Velocity (Trend), representing the net flow rate, (c) $2c_2$ the Liquidity Acceleration (Curvature), representing the stability of the flow and finally (d) $k!c_k$: The higher-order derivatives (Jerk, Snap), representing complex algorithmic reactions.

4.2.1 Effective Liquidity Imbalance

To synthesize these kinematic measurements into a single actionable signal, we introduce the **Effective Liquidity Imbalance**. Traditional microstructure metrics, such as the Order Book Imbalance, rely on static snapshots of visible depth. These metrics fail because they treat a collapsing queue (negative velocity) as identical to a refilling queue (positive velocity) if their instantaneous depths are equal. We resolve this by defining ELI not as a function of current depth, but as a function of the projected depth over a forward-looking horizon τ. We define the projected depth, $\tilde{Q}(t + \tau)$, as the expected effective liquidity available at time $t + \tau$, extrapolated using the measured velocity and acceleration. To ensure physical realism

(as liquidity cannot be negative), we apply a non-negative constraint:

$$\tilde{Q}(t+\tau) = \max\left(0, \hat{Q}_t + \hat{v}_{liq}\tau + \frac{1}{2}\hat{a}_{liq}\tau^2\right) \tag{4.8}$$

The Effective Liquidity Imbalance is then calculated as the normalized difference between the projected depths on the bid and ask sides:

$$ELI(\tau) = \frac{\tilde{Q}_{bid}(t+\tau) - \tilde{Q}_{ask}(t+\tau)}{\tilde{Q}_{bid}(t+\tau) + \tilde{Q}_{ask}(t+\tau)} \tag{4.9}$$

The metric fundamentally captures the trajectory of the order book rather than its state. By incorporating the derivatives of liquidity, ELI correctly identifies the strength of a side based on its momentum. A queue that is ostensibly deep but rapidly collapsing ($\hat{v} \ll 0$) will yield a low projected depth $\tilde{Q}$, correctly identifying it as a weak support level. Conversely, a shallow queue that is being aggressively replenished ($\hat{v} \gg 0$) will yield a high projected depth, identifying it as a hidden wall of liquidity. The above kinematic adjustment solves the fundamental limitation of static OBI/OFI metrics, allowing the execution algorithm to align with the true flow of institutional inventory.

Bibliography

[1] Arrow, K. J., & Debreu, G. (1954). Existence of an equilibrium for a competitive economy. *Econometrica*, 22(3), 265–290.

[2] Debreu, G. (1959). *Theory of value: An axiomatic analysis of economic equilibrium.* Yale University Press.

[3] Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3), 393–408.

[4] Walras, L. (1954). *Elements of pure economics* (W. Jaffé, Trans.). George Allen and Unwin. (Original work published 1874).

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

[6] Kyle, A. S. (1989). Informed speculation with imperfect competition. *The Review of Economic Studies*, 56(3), 317–355.

[7] Glosten, L. R. (1994). Is the electronic open limit order book inevitable? *Journal of Finance*, 49(4), 1127–1161.

[8] O'Hara, M. (1995). *Market microstructure theory.* Blackwell.

[9] Almgren, R., & Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of Risk*, 3, 5–40.

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

[11] Cartea, Á., Jaimungal, S., & Penalva, J. (2015). *Algorithmic and high-frequency trading.* Cambridge University Press.

[12] Bouchaud, J.-P., & Potters, M. (2002). *Theory of financial risk and derivative pricing: From statistical physics to risk management.* Cambridge University Press.

[13] Cont, R., Stoikov, S., & Talreja, R. (2010). A stochastic model for order book dynamics. *Operations Research*, 58(3), 549–563.

[14] Cont, R., Kukanov, A., & Stoikov, S. (2014). The price impact of order book events. *Journal of Financial Econometrics*, 12(1), 47–88.

[15] Bacry, E., Mastromatteo, I., & Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01), 1550005.

[16] Rosu, I. (2009). A dynamic model of the limit order book. *Review of Financial Studies*, 22(11), 4601–4641.

[17] Parlour, C. A., & Seppi, D. J. (2008). Limit order markets: A survey. In A. V. Thakor & A. W. A. Boot (Eds.), *Handbook of Financial Intermediation and Banking* (pp. 63–96). North-Holland.

[18] Christensen, C., & Woodmansey, P. (2013). Detecting iceberg orders. *Market Microstructure Knowledge Base*.

[19] Frey, S., & Sandås, P. (2017). The impact of iceberg orders in limit order books. *Review of Finance*, 21(2), 773–800.

[20] Moro, E., Vicente, J., Moyano, L. G., Gerig, A., Farmer, J. D., Vaglica, G., Lillo, F., & Mantegna, R. N. (2009). Market impact and trading profile of hidden orders in stock markets. *Physical Review E*, 80(6), 066102.

[21] Zotikov, D., & Antonov, A. (2019). CME iceberg order detection and prediction. *arXiv preprint arXiv:1909.09495*.

[22] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, 22(6), 2201–2238.

[23] Hasbrouck, J. (2007). *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*. Oxford University Press.

[24] Moallemi, C. C., & Yuan, K. (2016). A model for queue position valuation in a limit order book. *Working Paper*, Columbia University.

.1 Proofs

This section presents the various proofs of the different concepts mentioned in this book.

.1.1 Conservation Law

Queue at price p^* evolves as:

$$Q(t) = Q(0) + \int_0^t dN_L(s) - \int_0^t dN_M(s) - \int_0^t dN_C(s)$$

For probe P_1 at position q_1, execution at T_1 implies:

$$\int_0^{T_1} (dN_M(s) + dN_C(s)) = q_1$$

For P_2 at position $q_2 = q_1 + V_{\text{gap}} + H_{\text{gap}}$:

$$\int_0^{T_2} (dN_M(s) + dN_C(s)) = q_2 = q_1 + V_{\text{gap}} + H_{\text{gap}}$$

Subtracting:

$$\int_{T_1}^{T_2} (dN_M(s) + dN_C(s)) = V_{\text{gap}} + H_{\text{gap}}$$

But LHS is observable:

$$D_{\text{obs}} = \int_{T_1}^{T_2} dN_M(s) + \int_{T_1}^{T_2} dN_C(s)$$

Hence:

$$H_{\text{gap}} = D_{\text{obs}} - V_{\text{gap}}$$

.1.2 Unbiasedness Under Poisson Assumptions

Theorem: If $N_M(t), N_C(t)$ are Poisson with constant rates λ_M, λ_C, and icebergs refill uniformly in time, then $\mathbb{E}[\hat{H}_{\text{gap}}] = H_{\text{gap}}$.

Proof:

$$\mathbb{E}[D_{\text{obs}}] = \mathbb{E}\left[\int_{T_1}^{T_2} dN_M(s) + dN_C(s)\right] = (\lambda_M + \lambda_C)\mathbb{E}[T_2 - T_1]$$

By definition, $T_2 - T_1$ is the time to deplete q_2:

$$\mathbb{E}[T_2 - T_1] = \frac{q_2}{\lambda_M + \lambda_C} = \frac{V_{\text{gap}} + H_{\text{gap}}}{\lambda_M + \lambda_C}$$

Substituting:

$$\mathbb{E}[D_{\text{obs}}] = (\lambda_M + \lambda_C) \cdot \frac{V_{\text{gap}} + H_{\text{gap}}}{\lambda_M + \lambda_C} = V_{\text{gap}} + H_{\text{gap}}$$

Hence:

$$\mathbb{E}[\hat{H}_{\text{gap}}] = \mathbb{E}[D_{\text{obs}} - V_{\text{gap}}] = (V_{\text{gap}} + H_{\text{gap}}) - V_{\text{gap}} = H_{\text{gap}}$$