

The Price of the Queue

Reinforcement Learning and the Cross-Section of Limit Order Values

Abigail McHugh Adarsh Kumar Dhruv Joshi
Mia Luong Pranali Sancheti Somay Chopra

UC Berkeley

February 2026

ABSTRACT

We apply a reinforcement learning framework to quantify the expected value of limit orders across twenty NASDAQ-listed equities spanning the full spectrum of tick-constraint regimes. Exploiting the uniform \$0.01 tick size mandated by SEC Rule 612, we construct a natural experiment in which cross-sectional variation in the effective tick constraint arises purely from differences in natural spread width—free of the confounding variation in tick structures present in prior work on Australian markets. Our Q-learning estimates reveal a factor-of-eighteen range in best-bid order value, from -0.023 ticks for Amazon to $+0.491$ ticks for CrowdStrike, and uncover three novel empirical patterns. First, the price-level gradient *inverts* for tick-constrained stocks: orders placed two ticks behind the best bid carry higher expected value than orders at the BBO, reflecting the concentration of adverse selection at the top of the book. Second, the cancel option embedded in every limit order exhibits a universal intraday U-shape—peaking at market open and close across all twenty stocks and all four microstructure regimes—while the price-level gradient itself rotates from negative to positive over the trading day. Third, cross-stock policy transfer succeeds within microstructure clusters but fails catastrophically across them, with the absolute difference in log spread serving as a sufficient statistic for transfer regret ($R^2 = 0.22$). A state-by-state backtest validates these findings out of model: the RL policy outperforms all naive strategies in 20 of 20 stocks and beats the best economically-motivated heuristic in 17 of 20, with theoretical option values predicting realized performance gains at $r = 0.72$.

Keywords: Limit orders, reinforcement learning, market microstructure, adverse selection, tick size, queue position, cancel option

JEL Classification: G10, G12, G14, C61

1. INTRODUCTION

Every limit order is a bet against the market. The trader who posts a bid surrenders the right to choose *when* to transact, accepting instead the risk that the market will come to her only when it knows something she does not. In return, she earns the spread—if she is filled at all. The expected value of this gamble depends on a constellation of state variables: where the order sits in the queue, how deep the book is, how volatile the market has become, and how wide the spread stands at any given moment. It also depends on what the trader does next—whether she leaves the order resting or cancels it in response to new information. The value of making this decision optimally is itself a state-dependent quantity, one that cannot be computed from any single microstructure statistic.

[Kwan and Philip \(2025\)](#) introduced a reinforcement learning framework that solves this problem in closed form. By discretizing the state space of a limit order’s life and applying tabular Q-learning with value iteration, they recover the expected value of limit orders across sixteen stocks on the Australian Securities Exchange. Their central finding is that limit order values vary dramatically with the tick-constraint regime: orders in heavily constrained stocks (where the minimum tick constitutes a large fraction of the spread) are worth far less than orders in unconstrained stocks, because the tick acts as a binding price floor that prevents the spread from adjusting to reflect true adverse selection costs.

We build on their framework but depart from their setting in a way that sharpens the identification. The Australian market features multiple tick sizes that vary across price ranges, creating a confound: when a stock’s tick size changes, so does the entire microstructure—depth, queue dynamics, and the competitive structure of liquidity provision all shift simultaneously. The United States, by contrast, has operated under a uniform penny tick since SEC Rule 612 took effect in 2005. Every stock on NASDAQ faces the same \$0.01 minimum price increment. This means that cross-sectional variation in the *effective* tick constraint—the ratio of the tick to the natural spread—arises purely from differences in the spread itself. A stock like Rivian, whose average spread barely exceeds one cent, is severely constrained by the penny tick. CrowdStrike, whose spread averages forty cents, treats the penny as rounding error. Between these extremes lies a continuous spectrum of constraint intensity, all observed under identical regulatory and market structure conditions.

This uniformity transforms a cross-sectional comparison into something closer to a natural experiment. We need not worry that differences in tick size regimes are driving our results; there is only one regime. The identifying variation comes entirely from differences in the economic characteristics of the stocks themselves—their volatility, liquidity, information environment, and price level—as these characteristics manifest in the natural width of the bid-ask spread.

We apply the [Kwan and Philip \(2025\)](#) framework to twenty NASDAQ-listed equities observed over sixty trading days, spanning Rivian at one extreme (spread of 1.02 ticks, 98% of observations at the minimum) to CrowdStrike at the other (spread of 39.9 ticks, never at the minimum). Using message-level order book data from Databento’s ITCH feed, we reconstruct the full limit order book at 100-millisecond resolution, build transition matrices for 16,875 distinct working states, and solve for Q-values by value iteration.

Four findings emerge.

The cross-sectional value surface. Best-bid Q-values range from -0.023 ticks (Amazon) to $+0.491$ ticks (CrowdStrike), a factor-of-eighteen variation driven almost entirely by spread width. But the more surprising pattern lies *within* the book. For the most tick-constrained stocks—Intel, SoFi, Super Micro, PayPal, The Trade Desk, and Microsoft—orders placed two ticks behind the best bid carry *higher* expected value than orders at the best bid itself. The price-level gradient inverts. This occurs because adverse selection concentrates at the top of the book: the best bid is the first to be picked off when informed traders arrive, while deeper levels are sheltered. No prior study has documented this gradient inversion at individual price-level resolution.

Intraday dynamics. The cancel option embedded in every limit order—the difference between the optimal Q-value and the value of being forced to stay—traces a U-shape over the trading day, peaking at the open and close and reaching its trough at midday. This pattern holds universally across all twenty stocks and all four microstructure regimes we identify (tight profitable, tight toxic, moderate, and wide spread). While the intraday U-shape in spreads and volatility is well documented, the U-shape in the *option value of cancellation* is new. We further show that the price-level gradient itself rotates over the day: for the widest-spread stocks, the gradient is negative in the morning (deeper levels more valuable) but turns positive by afternoon, as opening-hour adverse selection dissipates and the BBO becomes the more attractive price level.

Policy transferability. We construct a 20×20 matrix of transfer regret—the expected loss from deploying one stock’s optimal policy on another stock’s transition dynamics. The matrix reveals sharp block-diagonal structure: within microstructure clusters, transfer regret is near zero; across clusters, it is explosive. A regression analysis shows that the absolute difference in log spread is a sufficient statistic for predicting transfer failure ($R^2 = 0.22$, $t = 3.41$), while depth, trade intensity, and volatility add no explanatory power ($p > 0.37$ for each). The practical implication is immediate: before deploying an execution algorithm trained on one stock to another, one can predict the performance degradation from a single observable.

Backtest validation. We validate the RL framework through a state-by-state backtest on realized fills. The RL policy outperforms the best of two economically-motivated heuristics (a volatility-based cancellation rule and a queue-position-based rule) in seventeen of twenty stocks, and beats the unconditional always-leave strategy in all twenty. The correlation between theoretical option values and realized performance gaps is $r = 0.72$, confirming that the Q-learning estimates translate faithfully into economic value. For stocks where the heuristics are competitive—wide-spread names like Shopify and Meta—all strategies earn positive returns and the margins are thin. The RL framework earns its keep precisely where the problem is hard: in tick-constrained, adversely-selected stocks where naive strategies lose money.

Beyond these four contributions, we use the cross-sectional results to interpret the likely consequences of tick-size reform. The SEC’s proposal to reduce the minimum tick for certain stocks from $\$0.01$ to $\$0.005$ would push currently constrained names toward the moderate-constraint regime, where our results imply higher cancel option values, more active state-dependent cancellation, and more negative constrained expected values. We discuss these implications without making quantitative predictions, given the partial-equilibrium nature of our framework.

The remainder of the paper proceeds as follows. Section 2 reviews the related literature. Section 3 describes our data, sample construction, and the reinforcement learning methodology. Section 4 presents the cross-sectional value surface. Section 5 documents the intraday dynamics. Section 6

analyzes policy transferability. Section 7 provides the backtest validation and heuristic benchmarks. Section 8 discusses implications for tick-size reform and limitations. Section 9 concludes.

2. RELATED LITERATURE

This paper sits at the intersection of three literatures: the microstructure of limit order markets, the application of reinforcement learning to financial decision-making, and the economics of tick size regulation.

Limit order valuation.. The foundational insight that a limit order contains an embedded option dates to [Copeland and Galai \(1983\)](#), who showed that limit orders grant a free trading option to the rest of the market. [Handa and Schwartz \(1996\)](#) formalized this by modeling the limit order trader's problem as a tradeoff between execution probability and adverse selection costs, showing that limit orders can be optimal even for uninformed traders when the spread is wide relative to the information asymmetry. [Parlour \(1998\)](#) introduced dynamic considerations, demonstrating that the optimal order depends on the current state of the book, since a deep bid queue lengthens expected waiting times while a thin queue raises the probability of execution but also of being adversely selected.

Empirical work has confirmed that these tradeoffs are economically significant. [Griffiths et al. \(2000\)](#) documented that limit order costs depend critically on the aggressiveness of the order, with orders at the best quote bearing the highest adverse selection costs. [Hollifield et al. \(2004\)](#) estimated structural models of limit order markets and found that traders do respond optimally to queue depth and spread conditions, though imperfectly. More recently, [Cont et al. \(2014\)](#) showed that the dynamics of the limit order book at the top of the book are well-described by simple Markov models, providing theoretical support for the state-space approach we adopt here.

Reinforcement learning in microstructure.. Machine learning methods have increasingly been applied to order placement and execution problems. [Nevmyvaka et al. \(2006\)](#) were among the first to use reinforcement learning for optimal trade execution, showing that RL agents could significantly reduce execution costs relative to standard benchmarks. Their approach treated the problem as a Markov decision process (MDP) with features derived from the limit order book, establishing the template that subsequent work has followed.

[Kwan and Philip \(2025\)](#) advanced this approach by applying tabular Q-learning specifically to the *stay-or-cancel* decision facing a resting limit order. By discretizing the state space into queue position, book depth, volatility, and spread bins, they could solve for the full value function by backward induction on a finite Markov chain. Their key contribution was showing that the resulting Q-values provide a clean decomposition of limit order value into a constrained component (the value if forced to stay) and an option component (the additional value from the right to cancel). Our paper adopts their framework and extends it to the cross-sectional, intraday, and policy-transfer dimensions they did not explore.

Machine learning in financial economics.. Our paper also connects to the rapidly expanding use of machine learning in financial economics, recently surveyed by [Eisfeldt and Schubert \(2024\)](#). A central theme in this literature is the identification of parsimonious structure in high-dimensional

data: [Kozak et al. \(2020\)](#) show that a small number of principal components capture most cross-sectional return variation, [Kelly et al. \(2019\)](#) demonstrate that stock characteristics proxy for latent factor loadings, and [Freyberger et al. \(2020\)](#) use nonparametric methods to determine which characteristics matter for expected returns. Our transfer regret analysis (Section 6) performs an analogous exercise in the microstructure domain, finding that spread distance alone predicts policy transfer failure while depth, trade intensity, and volatility are irrelevant—a dimension-reduction result in the spirit of this literature. More broadly, machine learning has proven effective at extracting signal from complex financial data, including central bank communications ([Kakhbod et al., 2025](#)) and patent text ([Hochberg et al., 2023](#)), while a related stream examines how AI reshapes investment decisions and competitive dynamics ([Kakhbod et al., 2024](#); [Fedyk et al., 2024](#)). Methodologically, our use of tabular Q-learning rather than deep reinforcement learning reflects a deliberate emphasis on interpretability, consistent with recent work advocating for transparent ML models in high-stakes financial applications ([Bell et al., 2025, 2024](#)).

Tick size and market quality. The economics of minimum price increments has been studied extensively since the shift from fractional to decimal pricing on U.S. exchanges. [Bessembinder \(2003\)](#) surveyed the early evidence, finding that decimalization narrowed spreads but reduced displayed depth. [Yao and Ye \(2018\)](#) examined how tick size affects the balance between informed and uninformed trading, showing that smaller ticks disproportionately benefit informed traders by reducing the cost of price improvement.

More recently, the SEC’s Tick Size Pilot Program (2016–2018) provided quasi-experimental variation. [Rindi and Werner \(2020\)](#) analyzed the pilot and found that wider ticks increased depth at the best quote but reduced overall market quality for less liquid stocks. [Brogaard et al. \(2024\)](#) showed that wider ticks alter the economics of high-frequency market making by increasing the value of queue priority. Our paper contributes to this literature by quantifying how tick constraints affect the full value function of a limit order, not just the spread or depth at a single price level.

Positioning. Our contribution is primarily empirical. We do not propose new theory or a new computational method; we apply an existing framework to a setting that offers cleaner identification and then document patterns—gradient inversion, option value U-shapes, transfer regret structure—that have not been observed before. The most closely related concurrent work is the original [Kwan and Philip \(2025\)](#), from which we differ in three respects: the market (U.S. vs. Australia), the analytical scope (cross-sectional, intraday, and transfer analysis vs. stock-by-stock characterization), and the validation approach (realized backtest with heuristic benchmarks vs. theoretical analysis).

3. DATA AND METHODOLOGY

3.1 Sample Construction

Our sample consists of twenty stocks listed on the NASDAQ exchange, observed over sixty trading days from August 18 to November 10, 2025. We use Databento’s NASDAQ ITCH market-by-order (MBO) feed, which provides every order submission, modification, cancellation, and execution at nanosecond resolution. This is the most granular data available for U.S. equity markets and allows us to reconstruct the full limit order book at any point in time.

Table 1. Stock Universe and Microstructure Characteristics

Symbol	Price (\$)	Spread (ticks)	% at 1-tick	Depth (shares)	Depth (\$k)	Intensity (trd/s)	Volatility
RIVN	13.74	1.02	98.2	5848	80	1.9	0.0016
SNAP	7.75	1.02	98.8	11,992	93	0.8	0.0023
INTC	32.10	1.05	95.4	4890	157	6.8	0.0008
SOFI	27.20	1.09	92.0	2942	80	4.5	0.0011
LCID	17.15	1.38	74.0	63,005	1080	1.2	0.0025
PYPL	68.92	1.73	52.9	543	37	3.4	0.0007
SMCI	47.58	1.73	50.0	810	39	4.5	0.0011
TTD	50.29	2.13	36.3	590	30	3.1	0.0011
AMZN	227.94	2.26	24.2	388	88	11.2	0.0003
AAPL	248.90	2.49	23.4	431	107	10.8	0.0003
MRVL	79.87	2.84	24.7	451	36	4.7	0.0009
DXCM	70.37	4.66	6.8	347	24	1.7	0.0014
ABNB	124.44	5.22	4.7	262	33	1.7	0.0009
MSFT	512.59	7.12	1.3	187	96	7.2	0.0003
ROKU	98.28	9.87	0.8	219	22	1.1	0.0020
PANW	203.76	10.01	0.4	198	40	1.9	0.0010
SHOP	153.28	10.63	0.3	310	48	2.0	0.0013
META	728.46	20.90	0.0	129	94	5.4	0.0005
COIN	330.47	24.74	0.0	175	58	3.1	0.0014
CRWD	477.49	39.87	0.0	111	53	1.3	0.0014

Notes: Sample covers 60 trading days of NASDAQ ITCH MBO data. Spread is the time-weighted average quoted spread in tick units (\$0.01). % at 1-tick is the fraction of time the quoted spread equals the minimum. Depth is average displayed quantity at the best bid. Intensity is the average number of trades per second during regular trading hours. Volatility is the 100ms realized volatility of quote midpoint returns.

We select stocks to span the full range of effective tick constraints available in U.S. markets. At one extreme, Rivian (RIVN) and Snap (SNAP) trade at prices below \$15, where the penny tick constitutes nearly 100% of the spread—these are maximally constrained. At the other extreme, CrowdStrike (CRWD) and Meta (META) trade at prices exceeding \$300, with spreads of twenty to forty ticks—the penny is economically irrelevant for these names. Between these endpoints, we include stocks at each level of constraint intensity, ensuring continuous coverage of the spectrum.

Table 1 presents the twenty stocks ordered by average spread width, along with key microstructure characteristics. Several features of the sample are worth noting. Spread width varies by a factor of thirty-nine, from 1.02 ticks (Rivian) to 39.87 ticks (CrowdStrike). Depth at the best bid varies even more dramatically—from 111 shares (CrowdStrike) to 63,005 shares (Lucid)—reflecting the well-known inverse relationship between spread and depth. Five stocks (Rivian, Snap, Intel, SoFi, Lucid) have sub-penny time-weighted average credited spreads (TWACS below \$0.015) and would be directly affected by the SEC’s proposed tick-size reform under Rule 612.

3.2 Order Book Reconstruction

We reconstruct the limit order book from the raw ITCH message feed using a pipeline that processes each message chronologically. Every order addition establishes a new entry in our book state with its price, size, and queue position; modifications update size or price; cancellations remove entries; and executions decrement displayed quantity and record fill events. We snapshot the book state at 100-millisecond intervals during regular trading hours (9:30–16:00 ET), yielding approximately 234,000 snapshots per stock-day.

For each snapshot, we record the state variables that will form the basis of our reinforcement learning framework: the queue position of a hypothetical order at each of the first three price levels (best bid, one tick behind, and two ticks behind), the displayed depth at each of the three bid levels and at the best ask, and the 100ms realized volatility of the midpoint return.

3.3 The Reinforcement Learning Framework

We adopt the Markov decision process formulation of [Kwan and Philip \(2025\)](#). A limit order's life is modeled as a sequence of discrete 100-millisecond intervals. At each interval, the order exists in one of three absorbing terminal states—*filled*, *cancelled by the trader*, or *cancelled due to price movement*—or in a *working state* characterized by a tuple of discretized microstructure variables.

State space. The working state is defined by seven variables, each discretized into a small number of bins determined by quintiles (or terciles) of their empirical distribution:

- **Price level** $\ell \in \{0, 1, 2\}$: the order's position relative to the current best bid (L0 = at the best bid, L1 = one tick behind, L2 = two ticks behind).
- **Queue position** $q \in \{1, 2, \dots, 5\}$: quintile of the order's position within its price level, where 1 denotes the front of the queue.
- **Bid depth at best bid** $d_0 \in \{1, 2, \dots, 5\}$: quintile of displayed depth at the best bid price.
- **Bid depth one tick behind** $d_1 \in \{1, 2, \dots, 5\}$: quintile of displayed depth at the second-best bid price.
- **Bid depth two ticks behind** $d_2 \in \{1, 2, 3\}$: tercile of displayed depth at the third-best bid price.
- **Ask depth at best ask** $a_0 \in \{1, 2, \dots, 5\}$: quintile of displayed depth at the best ask price.
- **Volatility** $v \in \{1, 2, 3\}$: tercile of recent realized volatility.

This yields $3 \times 5 \times 5 \times 5 \times 3 \times 5 \times 3 = 16,875$ working states. In addition, orders that move behind L2 enter one of $5 \times 5 \times 3 \times 5 \times 3 = 1,125$ terminal "long" states (where the book state is still tracked but the order no longer has a price level or queue position), plus a single absorbing cancellation state. The total state space is $16,875 + 1,125 + 1 = 18,001$ states.

The tabular approach is a deliberate methodological choice. While deep Q-learning or neural function approximation could accommodate a finer state discretization, the tabular framework produces fully interpretable Q-values for every state, admits the clean decomposition into constrained and option components developed below, and yields an optimal policy that is a transparent lookup table rather than a black-box approximator. We view this interpretability as essential when the economic structure of the value function—not prediction accuracy alone—is the primary object of interest.

Actions. At each interval, the trader chooses between two actions: *leave* the order resting, or *cancel* it. If the order is left, it transitions according to the empirical transition probabilities estimated from the data. If cancelled, it enters the absorbing cancellation state with a payoff of zero.

Rewards. When a working order is filled, it earns a reward equal to the difference between its limit price and the prevailing midpoint at the time of fill, measured in ticks. This captures the P&L of a market-maker who buys at the bid and marks to mid. When an order is cancelled (either voluntarily or due to price movement), the reward is zero.

Value iteration. We estimate the transition matrix $\mathbf{P}(s' | s, a)$ and reward function $R(s, a)$ directly from the reconstructed order book data and solve for the optimal value function by value iteration:

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} P(s' | s, a) \max_{a'} Q^*(s', a') \quad (1)$$

where $\gamma = 1$ (no discounting, following [Kwan and Philip 2025](#)). We iterate until convergence, defined as $\max_s |Q^{(t+1)}(s) - Q^{(t)}(s)| < 10^{-8}$.

Key derived quantities. From the converged Q-values, we compute three summary statistics for each stock:

$$Q(L_\ell) = \mathbb{E}_{\text{states at level } \ell} [\max(Q^*(s, \text{leave}), Q^*(s, \text{cancel}))] \quad (2)$$

$$\text{OptVal} = \mathbb{E} [\max(Q^*(s, \text{leave}), 0) - Q^*(s, \text{leave})] \quad (3)$$

$$\text{ConEV} = \mathbb{E} [Q^*(s, \text{leave})] \quad (4)$$

$Q(L_\ell)$ is the expected value of an order at price level ℓ under the optimal policy. *OptVal* is the cancel option value—the expected gain from being allowed to cancel rather than being forced to stay. *ConEV* is the constrained expected value—the value of the order if the trader must passively hold it to terminal resolution. The optimal value decomposes as $Q(L_\ell) = \text{ConEV} + \text{OptVal}$: the order is worth its passive value plus the option to walk away.

4. THE CROSS-SECTIONAL VALUE SURFACE

We begin with the broadest view: how does the expected value of a limit order vary across stocks? The answer, as we will show, is that it varies enormously—and in ways that reveal the fundamental tension between tick constraints and adverse selection.

4.1 The Spread–Value Relationship

Figure 1 plots the best-bid Q-value, $Q(L_0)$, against the average quoted spread for each of the twenty stocks. The relationship is monotonically positive on a log scale, ranging from -0.023 ticks for Amazon to $+0.491$ ticks for CrowdStrike—a factor of eighteen in absolute magnitude.

The economic logic is transparent. When the spread is narrow relative to the tick, the limit order trader is squeezed between the bid and the ask with almost no room for the spread to compensate her for adverse selection risk. Rivian’s spread of 1.02 ticks means the order earns at most half a penny of spread capture, while facing the full force of informed trading. As the spread widens, the limit order trader’s compensation grows faster than her adverse selection costs, because informed traders face a wider moat to cross before picking off the resting order. By the time we reach CrowdStrike, with a spread of forty ticks, the limit order is so well-compensated that even accounting for adverse selection, the expected value is strongly positive.

This result confirms the central finding of [Kwan and Philip \(2025\)](#) in a new market with cleaner identification. In their Australian sample, the same spread–value relationship emerged, but was potentially confounded by variation in tick sizes across stocks. Here, with a uniform penny tick, the relationship is driven entirely by the natural spread—which is itself a function of the stock’s information environment, liquidity, and price level.

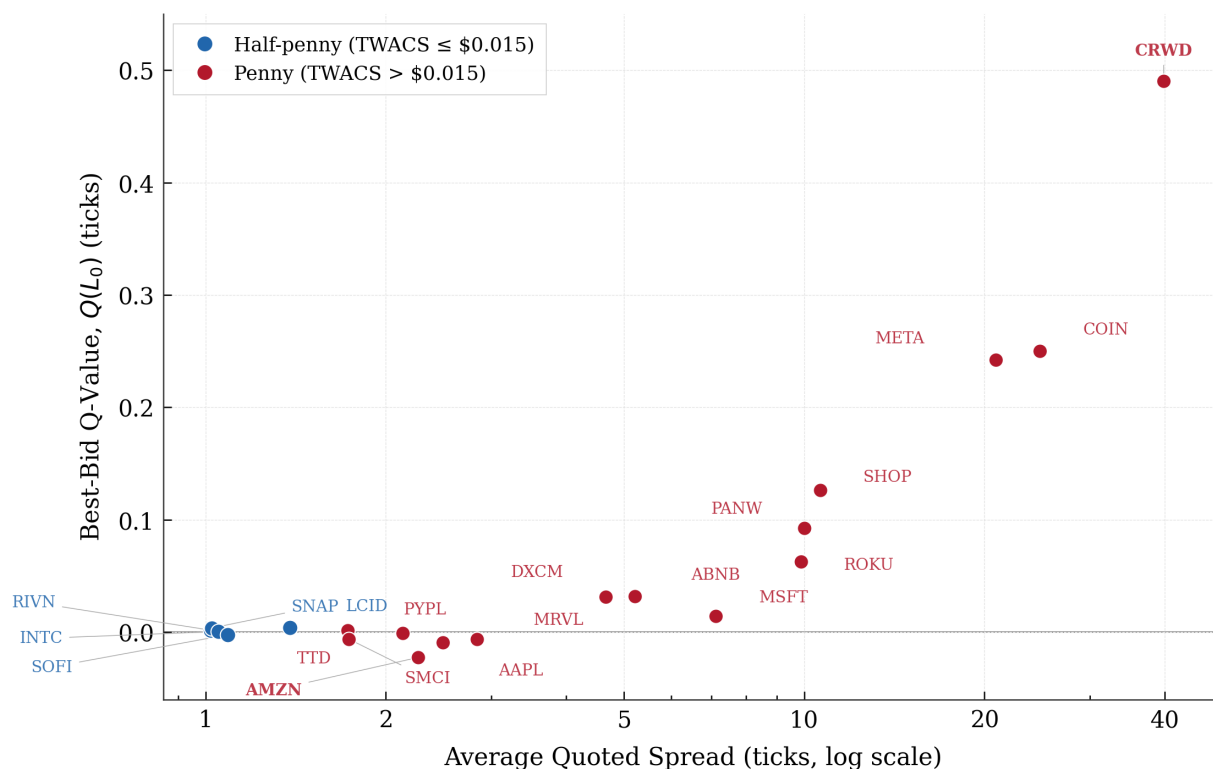


Figure 1. Best-bid Q-value versus average quoted spread across twenty stocks. The horizontal axis is on a logarithmic scale. Half-penny stocks ($TWACS \leq \$0.015$) are shown in blue; all others in red. The relationship is monotonically positive, with a factor-of-eighteen range from Amazon (-0.023) to CrowdStrike ($+0.491$).

Five stocks deserve special attention. Rivian, Snap, Intel, SoFi, and Lucid all have TWACS below \$0.015, qualifying them as “half-penny” stocks under the SEC’s proposed tick-size reform. These five cluster at the far left of Figure 1, with Q-values near zero or slightly negative. If the tick were halved for these stocks, they would—all else equal—shift rightward along the spread axis, moving toward the moderate-constraint regime where our results show limit order values are more nuanced and state-dependent.

4.2 Price-Level Gradient Inversion

The cross-sectional variation in $Q(L_0)$ tells us how the *best bid* varies across stocks. But a limit order need not sit at the best bid. What happens at deeper price levels?

Figure 2 plots the price-level gradient, defined as $Q(L_0) - Q(L_2)$, against the average spread. A positive gradient means the best bid is more valuable than two ticks behind; a negative gradient means the reverse. The pattern that emerges is striking: for the most tick-constrained stocks, the gradient is *negative*.

Table 2 makes this concrete. For Intel (spread 1.05 ticks), $Q(L_0) = +0.0004$ ticks while $Q(L_2) = +0.0018$ —orders two ticks deep are four times more valuable. For SoFi (1.09 ticks), the pattern is even more pronounced: $Q(L_0) = -0.0024$ while $Q(L_2) = +0.0006$. The best bid *loses money on average*, while the second-to-best-bid level is slightly profitable. The most extreme case is Amazon: $Q(L_0) = -0.0226$ versus $Q(L_2) = +0.0014$, a gradient of -0.024 ticks.

The economic intuition for gradient inversion is straightforward once stated. In a tick-constrained

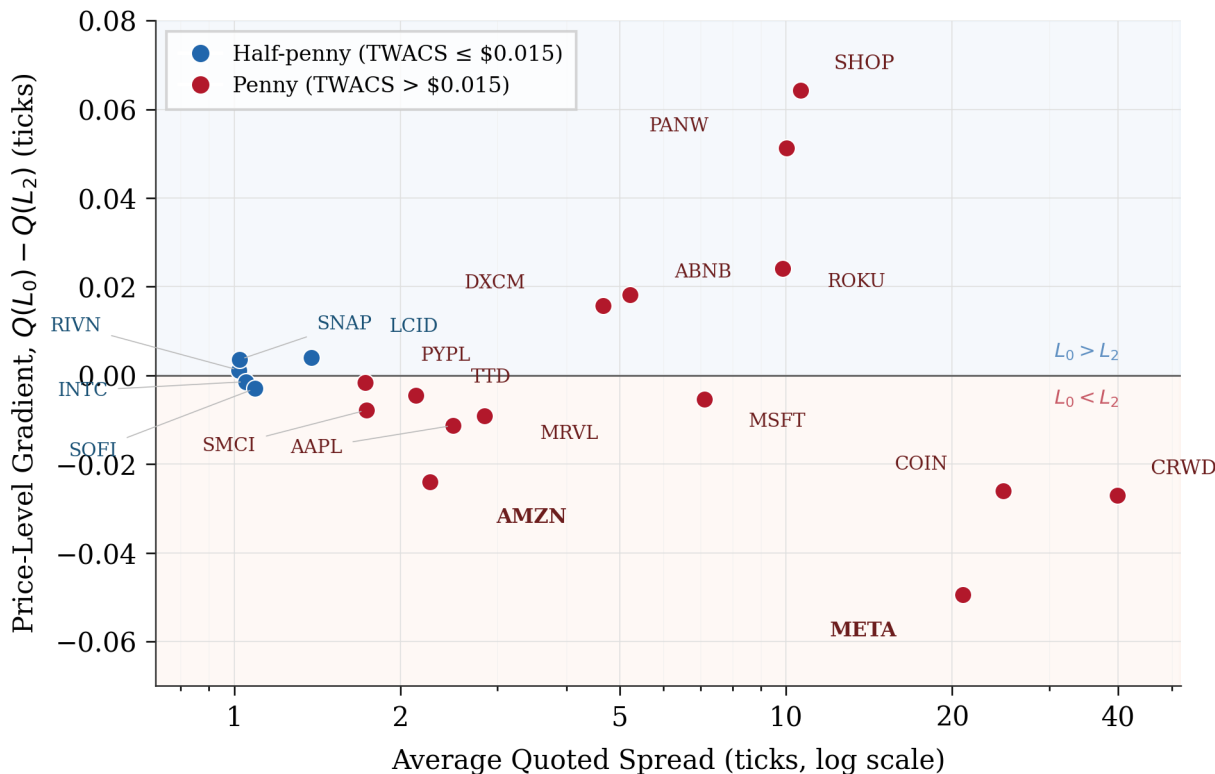


Figure 2. Price-level gradient ($Q(L_0) - Q(L_2)$) versus average quoted spread. Negative values indicate that orders two ticks behind the best bid have higher expected value than orders at the best bid. The gradient inverts for tick-constrained stocks (spread < 3 ticks) where adverse selection concentrates at the BBO.

market, the best bid is the most exposed price level. When an informed seller arrives, she hits the best bid first; the queue at L_0 bears the brunt of adverse selection. Orders at L_1 and L_2 are sheltered: they are filled only after the entire L_0 queue is exhausted, which means they tend to be filled in the context of large, persistent order flow—flow that is more likely to reflect genuine liquidity demand rather than information. For the same reason, the best bid’s fill rate is mechanically higher, but its *conditional* fill quality—the P&L given a fill occurs—is much worse.

In unconstrained markets, the gradient is positive because the spread provides enough cushion that the best bid’s execution advantage (higher fill probability, faster fills) more than compensates for its adverse selection exposure. The gradient inversion occurs precisely where the tick constrains the spread to the point that this compensation becomes insufficient.

4.3 The Cancel Option Value

The option to cancel a resting limit order is valuable because it allows the trader to exit before adverse price movements fully materialize. The question is: how valuable, and for whom?

Figure 3 reveals an inverted-U relationship between $OptVal$ and the spread. The cancel option is least valuable for the most constrained stocks (Snap: 0.011 ticks; Lucid: 0.005 ticks) and for the widest stocks (Shopify: 0.037 ticks). It peaks at moderate spreads: Amazon (0.130 ticks), Microsoft (0.130 ticks), and Marvel Technology (0.094 ticks).

The inverted-U has a clean interpretation rooted in the two forces that drive option value. For extremely constrained stocks, the value surface is nearly flat—every state yields roughly the same

Table 2. Q-Values by Price Level and Derived Quantities

Symbol	Spread (ticks)	$Q(L_0)$ (ticks)	$Q(L_1)$ (ticks)	$Q(L_2)$ (ticks)	OptVal (ticks)	ConEV (ticks)	Leave %
RIVN	1.02	0.0016	0.0008	0.0004	0.0265	-0.0251	77.5
SNAP	1.02	0.0036	0.0009	0.0000	0.0111	-0.0094	88.8
INTC	1.05	0.0004	0.0006	0.0018	0.0736	-0.0699	60.8
SOFI	1.09	-0.0024	-0.0008	0.0006	0.0564	-0.0553	65.3
LCID	1.38	0.0040	0.0023	0.0000	0.0045	-0.0028	71.3
PYPL	1.73	0.0013	0.0020	0.0028	0.0667	-0.0621	57.6
SMCI	1.73	-0.0063	-0.0007	0.0016	0.0777	-0.0762	45.5
TTD	2.13	-0.0008	0.0014	0.0037	0.0734	-0.0688	51.1
AMZN	2.26	-0.0226	-0.0048	0.0014	0.1301	-0.1271	29.9
AAPL	2.49	-0.0095	-0.0023	0.0018	0.0925	-0.0888	34.0
MRVL	2.84	-0.0063	0.0002	0.0028	0.0940	-0.0907	40.2
DXCM	4.66	0.0313	0.0198	0.0156	0.0554	-0.0302	69.9
ABNB	5.22	0.0319	0.0206	0.0136	0.0664	-0.0407	65.9
MSFT	7.12	0.0140	0.0086	0.0193	0.1295	-0.1021	49.4
ROKU	9.87	0.0628	0.0411	0.0386	0.0565	-0.0049	76.0
PANW	10.01	0.0925	0.0527	0.0413	0.0613	0.0070	77.6
SHOP	10.63	0.1263	0.0800	0.0619	0.0370	0.0576	87.2
META	20.90	0.2425	0.2661	0.2920	0.0748	0.2083	86.5
COIN	24.74	0.2503	0.2696	0.2763	0.0796	0.1840	84.9
CRWD	39.87	0.4906	0.5105	0.5176	0.1466	0.3713	88.0

Notes: Q-values are averages across intraday time blocks, measured in ticks (\$0.01). $Q(L_\ell)$ is the expected value of an order at price level ℓ under the optimal stay/cancel policy. OptVal is the cancel option value. ConEV is the constrained expected value (value if forced to hold). Leave % is the fraction of states where cancellation is optimal. Stocks are ordered by average quoted spread.

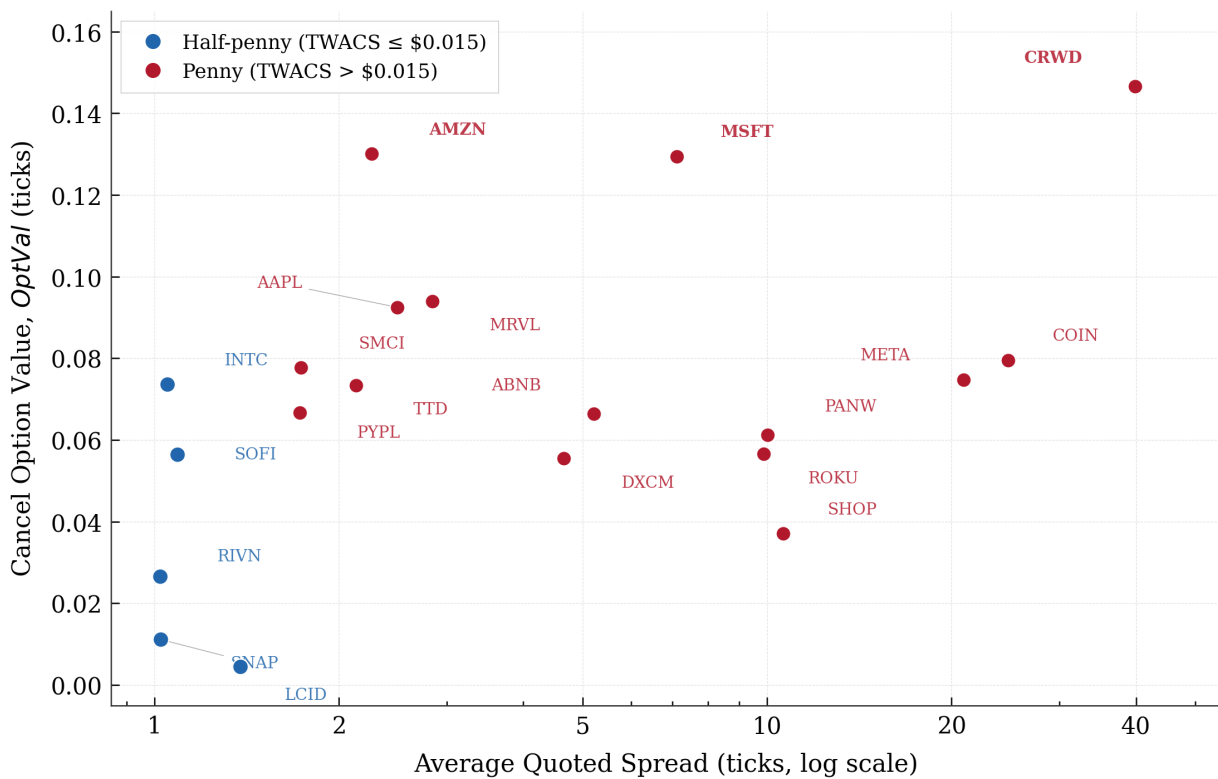


Figure 3. Cancel option value (OptVal) versus average quoted spread, showing an inverted-U pattern. The option is most valuable at moderate spreads (AMZN, MSFT, AAPL) and least valuable at both extremes.

(near-zero) payoff, so there is nothing to be gained from state-dependent cancellation. For extremely wide stocks, nearly every state is profitable—the spread is so wide that even adverse fills earn money, so there is little to avoid. The option to cancel is most valuable in the middle, where some states are profitable and others are not, and the trader’s ability to distinguish between them has the highest marginal value.

The optimal leave fraction (the proportion of states where cancellation is the best action) reinforces this interpretation. Table 2 shows that the leave fraction is high at both extremes—Snap at 88.8% and CrowdStrike at 88.0%—but for opposite reasons. Constrained stocks have high leave fractions because most states have slightly positive Q^* (leave) (the order is worth keeping, but only barely, so the optimal policy is nearly indifferent). Wide stocks have high leave fractions because the BBO value is so high that the optimal action at L0 is almost always to stay (and the “leave” percentage reflects the high fraction of states at deeper levels or in the “long” terminal category). The lowest leave fractions appear at moderate spreads: Amazon (29.9%), Apple (34.0%), Marvell (40.2%)—exactly the stocks where the cancel option is most valuable, because the policy discriminates most aggressively between good and bad states.

4.4 The Constrained Expected Value

The constrained expected value, $ConEV = Q(L) - OptVal$, answers a simple but important question: what would a limit order be worth if the trader *could not* cancel? This quantity isolates the passive component of limit order value—the payoff from submitting and walking away.

$ConEV$ is negative for all stocks with spreads below approximately ten ticks. For Amazon, $ConEV = -0.127$ ticks: a limit order submitted and left to its fate would lose money on average, because adverse selection costs dominate the thin spread capture. Only when the spread exceeds ten ticks—at Palo Alto Networks and beyond—does $ConEV$ turn positive, meaning that even a completely passive limit order strategy earns money on average.

This decomposition has a practical implication. For stocks in the moderate-spread range (AMZN, AAPL, MSFT, MRVL), the cancel option *is* the value of the limit order. Without it, these orders are liabilities. The RL framework reveals that active management—monitoring the state and cancelling in adverse conditions—transforms unprofitable liquidity provision into profitable. For Amazon, the optimal policy converts a $ConEV$ of -0.127 into an optimal value of -0.023 , recovering 82% of the passive loss through state-dependent cancellation. For Apple, the recovery is 89%. These are not trivial magnitudes: they represent the difference between a market-making strategy that bleeds money and one that approximately breaks even.

4.5 Summary of Cross-Sectional Patterns

The cross-sectional analysis reveals a coherent set of interlocking patterns:

1. The best-bid Q-value increases monotonically with the spread, reflecting the fundamental tradeoff between tick constraints and adverse selection.
2. The price-level gradient inverts for constrained stocks, as adverse selection concentrates at the best bid and deeper levels are sheltered.
3. The cancel option value traces an inverted-U, peaking where the value surface has the most cross-state variation.

4. The constrained expected value is negative for all but the widest stocks, implying that active cancellation management is essential for profitable liquidity provision in most of the market.

These patterns are static averages. In the next section, we show that they evolve systematically over the trading day, with the gradient inversion proving to be primarily a morning phenomenon driven by opening-hour adverse selection.

5. INTRADAY DYNAMICS

The cross-sectional patterns documented above are averages over the full trading day. But markets are not stationary within the day. Spreads widen at the open and close; volatility follows a well-documented U-shape; depth rebuilds gradually after the opening auction. If these microstructure conditions vary systematically, so should the value of a limit order.

To investigate, we partition each trading day into seven one-hour blocks (with the final block covering 15:30–16:00) and re-estimate the full Q-learning framework separately for each block. This yields seven complete value functions per stock, allowing us to trace how every quantity— $Q(L_\ell)$, $OptVal$, $ConEV$, and the price-level gradient—evolves from open to close.

5.1 The Universal Cancel Option U-Shape

Figure 4 displays the cancel option value as a heatmap, with stocks ordered by spread width on the vertical axis and time blocks on the horizontal axis. The pattern is immediately striking: $OptVal$ is highest in the 9:30–10:30 block for every single stock, declines through midday, and rebounds in the final half-hour. This U-shape is universal across all twenty stocks and all four microstructure regimes.

The magnitudes are substantial. CrowdStrike’s opening-hour $OptVal$ of 0.428 ticks is three times its midday trough of 0.074. Amazon drops from 0.191 to 0.098. Even the most constrained stock in our sample, Lucid (spread 1.38 ticks), shows a clear U-shape, declining from 0.012 at the open to 0.002 at midday before recovering to 0.002 at the close—small in absolute terms but a sixfold ratio.

Figure 5 makes the universality concrete by normalizing each stock’s $OptVal$ to 1.0 at the opening block and averaging within four microstructure regimes: tight profitable (Snap, Rivian, Lucid), tight toxic (Intel, SoFi, PayPal, Super Micro, The Trade Desk), moderate (Amazon, Apple, Marvell, DexCom, Airbnb, Microsoft, Roku), and wide spread (Palo Alto Networks, Shopify, Meta, Coinbase, CrowdStrike). All four regime averages trace the same U-shape, dropping to 25–65% of their opening value by early afternoon before rebounding. The wide-spread group declines most steeply, reaching approximately 25% of its opening value at 14:30 before a sharp close rebound.

The intraday U-shape in spreads and volatility has been documented extensively since [Wood et al. \(1985\)](#) and [Admati and Pfleiderer \(1988\)](#). What is new here is that the *option value of cancellation* follows the same pattern. The economic logic connects the two: when volatility is elevated at the open, the set of states from which a fill would be adversely selected expands, making the ability to cancel more valuable. As the market settles into midday equilibrium, adverse selection risk compresses, and the option to cancel is worth less. The close brings renewed uncertainty—position-squaring, end-of-day information, and index rebalancing flows—and the cancel option regains its value.

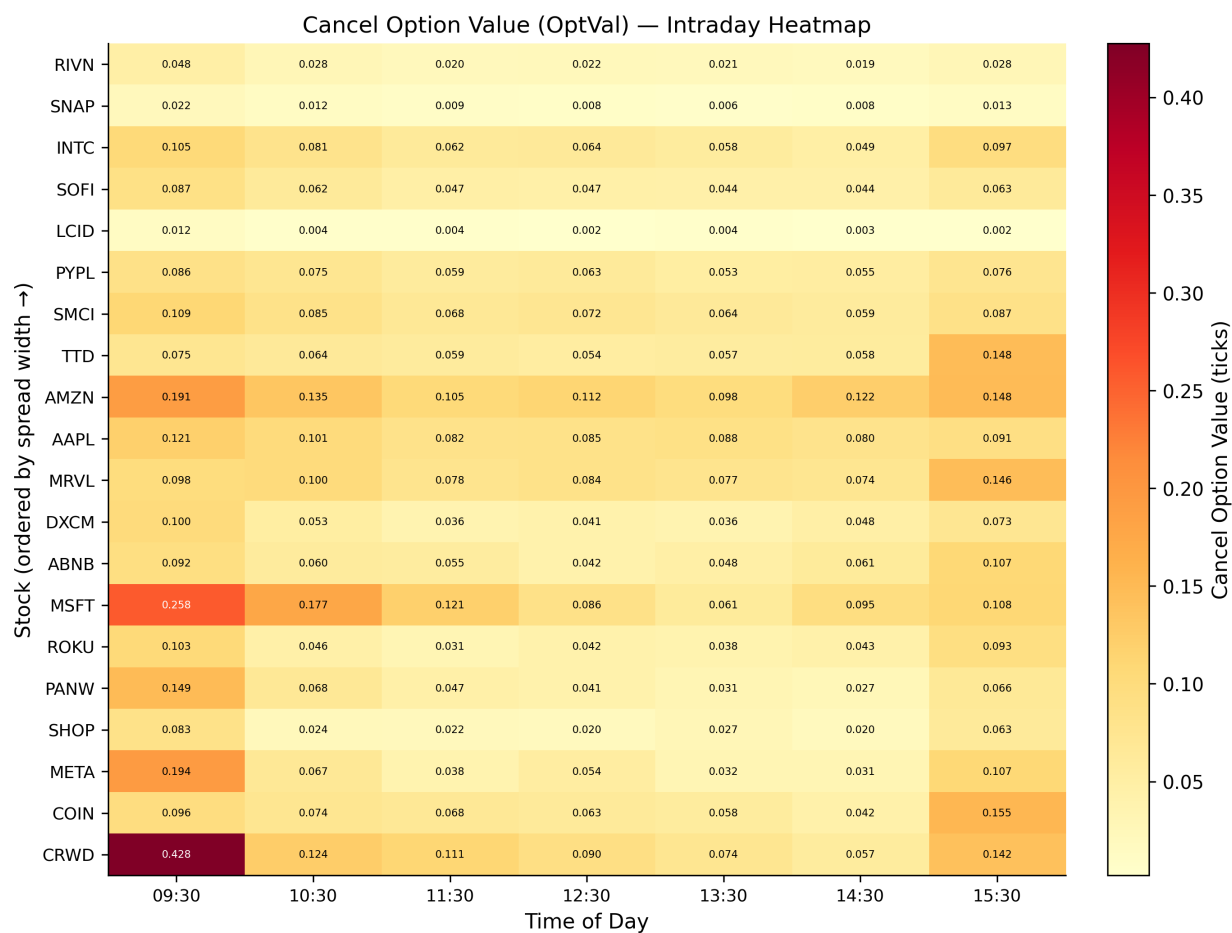


Figure 4. Cancel option value (OptVal) by stock and time of day. Stocks are ordered by average spread width (tightest at top). The U-shape—high at open, low at midday, rebounding at close—is universal across all twenty stocks.

5.2 Intraday Gradient Rotation

The price-level gradient, $Q(L_0) - Q(L_2)$, also evolves over the trading day, and it does so in a pattern that enriches the static gradient inversion documented in Section 4.

Figure 6 presents the gradient as an intraday heatmap. The most striking feature is the behavior of the three widest stocks—Meta, Coinbase, and CrowdStrike. At the open (9:30–10:30), all three show deeply negative gradients: -0.185 , -0.129 , and -0.179 ticks respectively. By the close (15:30–16:00), all three have rotated to positive: $+0.065$, $+0.046$, and $+0.087$. The gradient flips sign over the course of the trading day.

The economic interpretation is that opening-hour adverse selection at the best bid is so severe for these wide-spread stocks that market makers are better off positioning behind the queue until information incorporation completes. In the morning, the L_0 queue absorbs the full impact of overnight information being priced in, making deeper levels relatively more attractive. As the day progresses and the information environment stabilizes, the BBO's execution probability advantage reasserts itself, and the gradient turns positive.

This finding refines the cross-sectional gradient inversion from Section 4. For wide-spread stocks, the full-day average gradient appears only mildly negative precisely because it averages over a strongly negative morning and a positive afternoon. The gradient inversion is primarily a *morning*

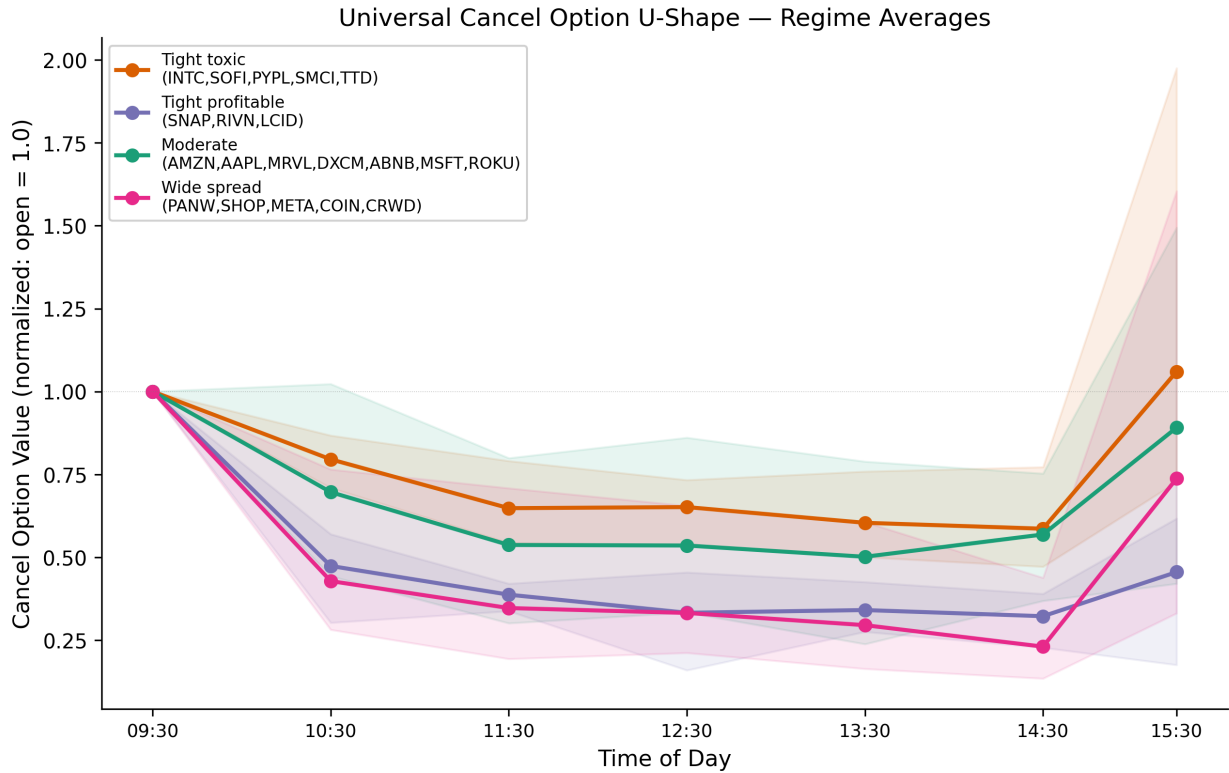


Figure 5. Normalized cancel option value by microstructure regime (opening block = 1.0). Shaded bands show cross-stock variation within each regime. All four regimes trace the same U-shape.

phenomenon driven by opening-hour information asymmetry.

For tight-spread stocks, the gradient shows minimal intraday variation. Intel’s gradient remains between -0.002 and -0.008 throughout the day, with a notable exception at 15:30 where it spikes to $+0.013$ —a close-of-day effect likely driven by the compression of adverse selection risk into the final half-hour. Moderate-spread stocks like Amazon and Apple show persistent negative gradients that narrow modestly toward the close but do not flip sign.

5.3 Statistical Robustness

To assess whether the gradient rotation is statistically reliable rather than an artifact of averaging, we test for significant differences in the gradient between the opening and closing blocks using cross-day variation. For each stock, we compute the gradient separately for each of the sixty trading days in both the 9:30 and 15:30 blocks, then test whether the mean difference (close minus open) is significantly different from zero.

Figure 7 presents intraday Q-values by price level for four representative stocks, one from each microstructure regime. Each panel displays $Q(L_0)$, $Q(L_1)$, and $Q(L_2)$ across the seven one-hour trading blocks, with light shading highlighting the opening and closing periods. Several patterns are immediately visible: the gradient inversion for INTC (L_0 persistently below L_1 and L_2 throughout the day, with a dramatic close-of-day spike); the persistent negative L_0 for AAPL with L_1 and L_2 crossing zero by mid-morning; and the gradient rotation for META, where L_2 starts as the most valuable level at the open but converges with L_0 by the close.

The Q-values displayed in Figure 7 are derived from the full transition matrix and are therefore

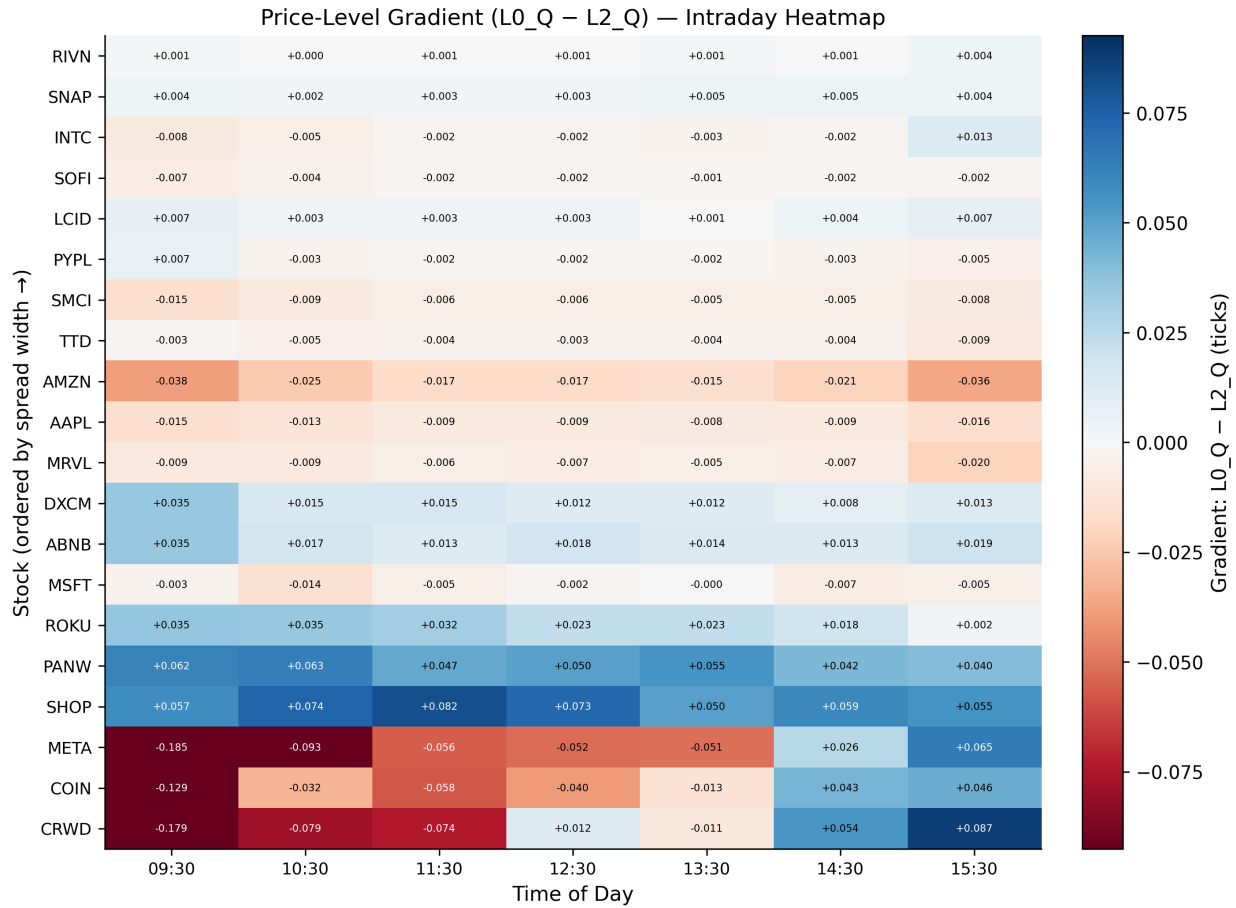


Figure 6. Price-level gradient ($Q(L_0) - Q(L_2)$) by stock and time of day. Blue cells indicate the best bid is more valuable than two ticks deep; red cells indicate the reverse. Note the sign change for META, COIN, and CRWD from morning (red) to afternoon (blue).

precise estimates conditional on the model. To test whether the gradient rotation is statistically reliable using variation *across trading days*, we construct an expected-value metric that weights each fill outcome by its probability: $EV_\ell = \text{fill_rate}_\ell \times \text{mean_fill_pnl}_\ell$. This is the per-observation analogue of the Q-value and provides a test of the gradient that can be computed day-by-day to generate confidence intervals.

Figure 8 reports the expected-value gradient change (close minus open) with 95% confidence intervals. Sixteen of twenty stocks show a statistically significant change at the 5% level. The four non-significant stocks—DexCom, Airbnb, Palo Alto Networks, and CrowdStrike—are cases where either the gradient is persistently one-signed throughout the day (PANW) or the cross-day variance is large relative to the mean shift (CRWD, where per-fill variance exceeds seven ticks).

Taken together, the intraday results show that the static cross-sectional patterns are not merely averages but evolve in economically interpretable ways. The cancel option U-shape reflects the intraday cycle of information asymmetry. The gradient rotation reveals that the relative value of price levels shifts systematically as the trading day progresses, with the BBO becoming increasingly attractive as opening-hour adverse selection dissipates.

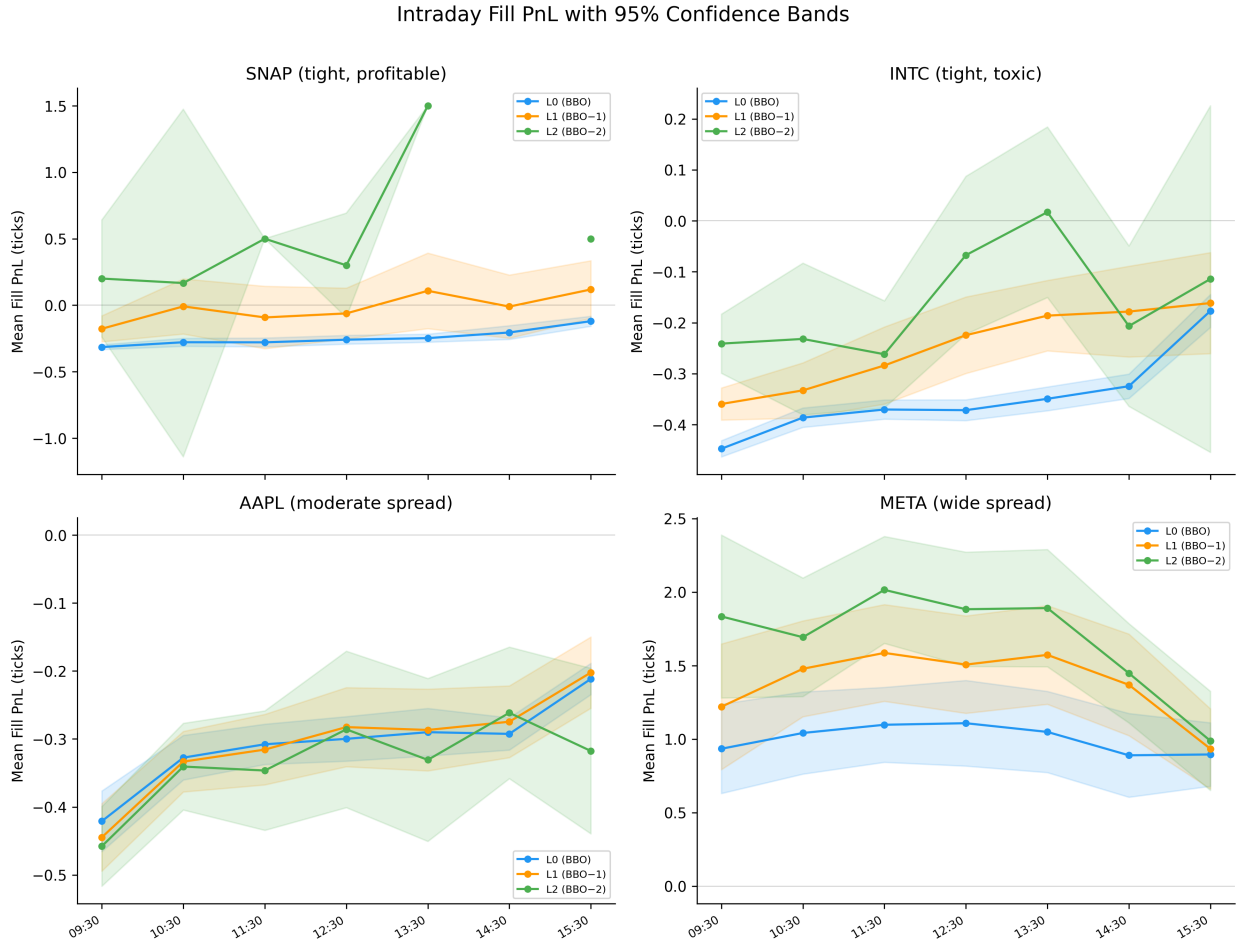


Figure 7. Intraday Q -values by price level for four representative stocks: *SNAP* (tight, profitable), *INTC* (tight, toxic), *AAPL* (moderate), and *META* (wide spread). Light shading marks the opening and closing blocks. The gradient inversion (L_0 below L_2) and its intraday rotation are clearly visible for *INTC* and *META*.

6. POLICY TRANSFERABILITY

A natural question arising from the cross-sectional analysis is whether optimal cancellation policies generalize across stocks. If a market maker has trained an RL policy on Amazon, can she deploy it on Apple? On CrowdStrike? The answer has practical implications for algorithmic trading: if policies transfer well, a single model can serve a portfolio of stocks; if they do not, each stock requires its own calibration.

6.1 The Transfer Regret Matrix

We define transfer regret as the expected value loss from applying stock i 's optimal policy to stock j 's transition dynamics:

$$\text{Regret}_{i \rightarrow j} = V_j^* - V_j^{\pi_i} \tag{5}$$

where V_j^* is stock j 's value under its own optimal policy and $V_j^{\pi_i}$ is its value under stock i 's policy. By construction, regret is non-negative: applying another stock's policy can only reduce value.

Figure 9 displays the 20×20 transfer regret matrix. The block-diagonal structure is immediately apparent. Within clusters of similarly-spread stocks, the heatmap is pale (near-zero regret). Across clusters, it turns dark (high regret, exceeding 0.30 ticks for the most distant transfers). The

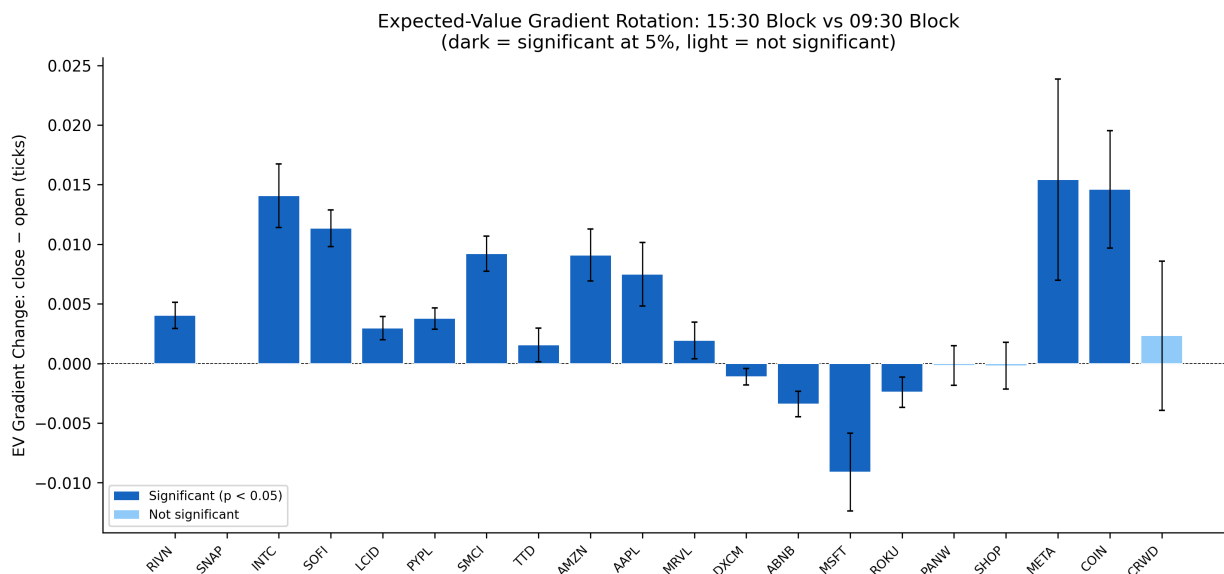


Figure 8. Expected-value gradient rotation: change in $EV(L_0) - EV(L_2)$ from the 9:30 block to the 15:30 block. Dark bars indicate significance at the 5% level (16/20 stocks). Error bars show 95% confidence intervals from cross-day variation.

matrix is not symmetric: transferring a wide-stock policy to a tight stock (which tends to cancel too aggressively, foregoing valuable fills) produces different regret than the reverse (which tends to cancel too passively, absorbing adverse fills).

6.2 Spread Distance as a Sufficient Statistic

What drives transfer failure? We regress pairwise transfer regret on observable microstructure distances between stocks. Table 3 presents the results across five specifications.

The results tell a clear story. In the univariate specification (column 1), the absolute difference in log spread alone explains 21.9% of the variation in transfer regret, with a t -statistic of 3.41. Adding a same-tick-class dummy in column (2) raises R^2 to 27.6%, and the dummy itself is significant ($t = 3.29$). Its positive sign may seem counterintuitive—being in the same tick class *increases* predicted regret—but this reflects the enormous within-class heterogeneity of the penny-spread group, which contains stocks ranging from PayPal at 1.73 ticks to CrowdStrike at 39.87 ticks. The high-regret within-penny pairs dominate.

The crucial result appears in column (3). Adding depth, trade intensity, and volatility distances to the regression produces no meaningful improvement: R^2 rises from 0.276 to 0.279, and none of the three additional variables approaches significance ($t = 0.49, 0.88, -0.21$). Spread distance is, for practical purposes, a sufficient statistic for transfer failure. A trader deciding whether to deploy an algorithm trained on one stock to another needs to compare only their spreads.

Figure 10 visualizes this relationship. Within-half-penny transfers (green) cluster at low spread distance and near-zero regret. Within-penny transfers (blue) show a positive relationship, with regret increasing as the constituent stocks become more spread-distant. Cross-class transfers (red) occupy the high-regret, high-distance region.

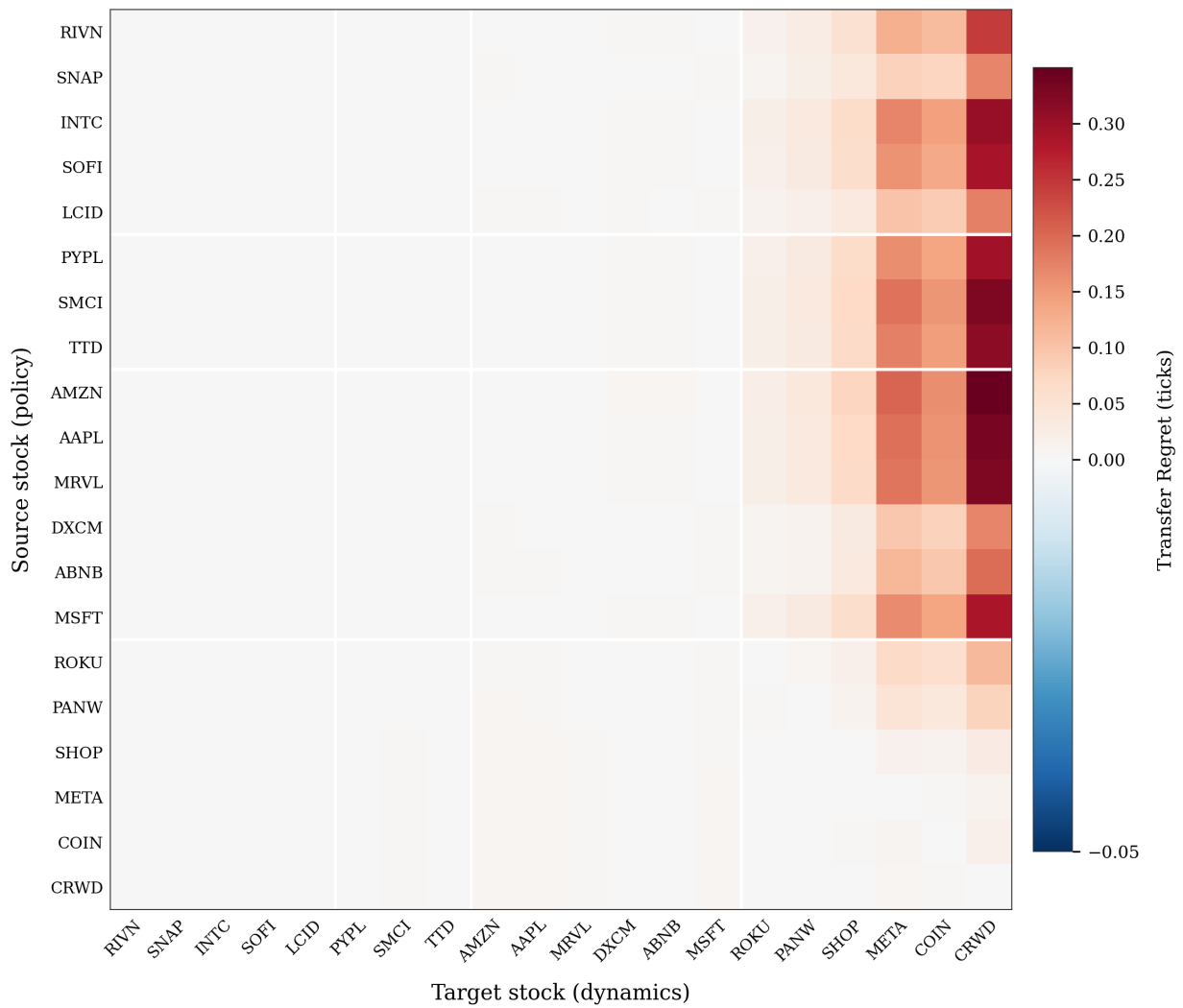


Figure 9. Transfer regret matrix (20×20). Each cell shows the expected value loss from applying the row stock's policy to the column stock's dynamics. Block-diagonal structure indicates that policies transfer well within microstructure clusters but fail across them.

Table 3. *Determinants of Transfer Regret*

	(1)	(2)	(3)	(4)	(5)
$ \Delta \log(\text{spread}) $	0.0311 (3.41)	0.0361 (3.64)	0.0366 (3.60)	0.0279 (2.94)	0.0364 (3.57)
Same tick class		0.0323 (3.29)	0.0327 (3.11)	0.0101 (0.84)	
$ \Delta \text{depth} $			0.0003 (0.49)		
$ \Delta \text{intensity} $			0.0049 (0.88)		
$ \Delta \text{volatility} $			-0.0013 (-0.21)		
$\log(\text{spread}) \times \text{Same class}$				0.0158 (1.15)	
Both half-penny					0.0368 (3.17)
Both penny					0.0320 (3.26)
Constant	-0.0148 (-2.06)	-0.0410 (-3.29)	-0.0457 (-3.02)	-0.0272 (-2.70)	-0.0415 (-3.26)
R^2	0.2190	0.2760	0.2790	0.2890	0.2770
Adj. R^2	0.2170	0.2730	0.2700	0.2840	0.2710
N	380	380	380	380	380

Notes: OLS regressions of pairwise transfer regret on microstructure distance measures. The sample consists of $20 \times 19 = 380$ directed stock pairs. t -statistics in parentheses. $|\Delta \log(\text{spread})|$ is the absolute difference in log average quoted spread. “Same tick class” is a dummy equal to one if both stocks are in the same SEC tick-size category (half-penny or penny). Depth, intensity, and volatility differences are standardized.

7. BACKTEST VALIDATION

The Q -values computed by value iteration are theoretical objects: they describe expected payoffs under the estimated transition dynamics. To what extent do they translate into realized economic value? We address this question through a state-by-state backtest on the original event data, comparing the RL policy against economically-motivated heuristics that isolate individual state variables, as well as naive baselines that establish lower bounds on performance.

7.1 Strategy Definitions

We evaluate seven strategies, each specifying a cancellation rule applied at every 100ms decision point. The first is the RL policy itself; the next two are economically-motivated heuristics that test whether a single state variable can replicate the RL advantage; the remaining four are naive baselines.

1. **RL Optimal:** Cancel if and only if $Q^*(s, \text{cancel}) > Q^*(s, \text{leave})$ for the current state s .
2. **Volatility heuristic:** Cancel when the current volatility bin is in the highest tercile. Leave otherwise. This captures the intuition that limit orders are most dangerous in volatile markets.
3. **Queue-position heuristic:** Leave only when the order is in the front 40% of the queue at the best bid. Cancel otherwise. This captures the intuition that queue priority is the primary determinant of fill quality.

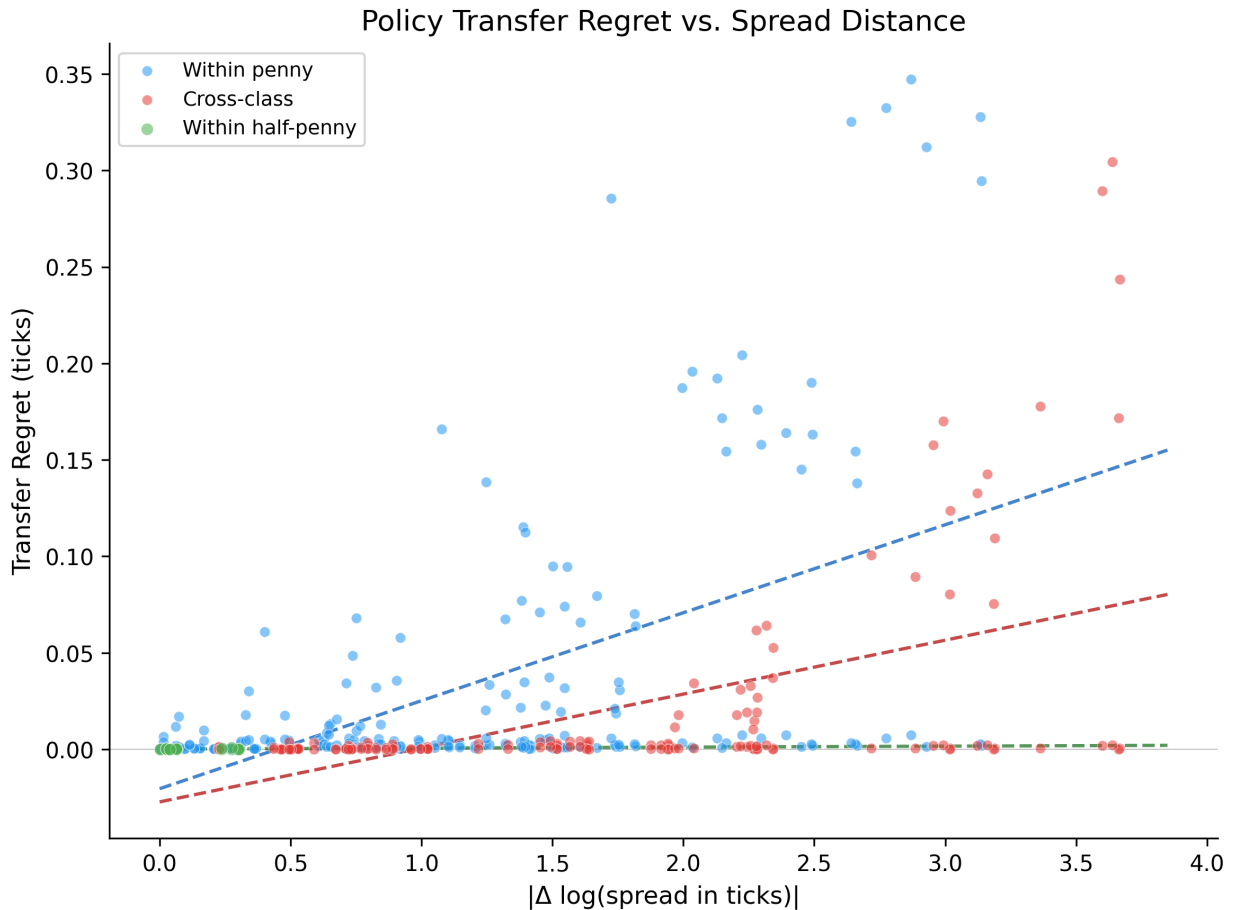


Figure 10. Transfer regret versus absolute difference in log spread, colored by tick-class pair type. Within-class transfers show near-zero regret at low spread distances; cross-class transfers show high regret. The positive slope confirms that spread distance is the primary determinant of transfer failure.

4. **Always-Leave-L0:** Always leave orders at the best bid; cancel at all other levels.
5. **Always-Leave-Any (ALA):** Always leave orders at every level (the fully passive baseline).
6. **Random:** Leave with 50% probability at each interval.
7. **Contrarian Transfer:** Apply the optimal policy of the most spread-distant stock.

Both heuristics are parameter-free in the sense that the thresholds (top tercile, front 40%) correspond to natural breakpoints in the state space. Neither requires solving the full MDP. For the backtest, we walk through every transition observation in the data. At each state, each strategy determines whether the order stays or is cancelled. If it stays and a fill occurs, we record the fill P&L (limit price minus midpoint). The expected P&L of a strategy is the fill rate (fraction of observations resulting in a fill) multiplied by the mean P&L conditional on being filled.

7.2 RL Versus Economically-Motivated Heuristics

The central question is not whether state-dependent cancellation outperforms passive holding—it trivially does—but whether the *specific* multi-dimensional cancellation rule discovered by Q-learning outperforms simpler rules that condition on a single state variable. The heuristic benchmarks provide the sharpest test.

Table 4 reports the expected P&L per observation for each strategy across all twenty stocks. The

Table 4. Backtest Performance: Expected P&L by Strategy

Symbol	RL	Vol Heur.	Queue Heur.	ALA	RL–Best
RIVN	+0.03	−0.32	−0.37	−0.44	+0.35
SNAP	+0.01	−0.08	−0.10	−0.12	+0.09
INTC	+0.07	−1.82	−2.00	−2.15	+1.89
SOFI	+0.03	−1.35	−1.67	−1.78	+1.38
LCID	+0.01	−0.55	−0.59	−0.66	+0.56
PYPL	+0.15	−1.35	−1.55	−1.69	+1.50
SMCI	+0.14	−2.19	−2.55	−2.68	+2.33
TTD	+0.20	−1.09	−1.19	−1.33	+1.29
AMZN	+0.86	−5.38	−6.01	−6.55	+6.23
AAPL	−0.19	−4.75	−5.12	−5.68	+4.56
MRVL	+0.28	−2.85	−3.12	−3.33	+3.13
DXCM	+0.47	−0.31	−0.19	−0.35	+0.67
ABNB	+0.71	−0.80	−0.46	−0.63	+1.17
MSFT	+1.56	−5.53	−4.07	−4.82	+5.62
ROKU	+0.94	−0.21	+0.23	+0.07	+0.71
PANW	+1.69	+0.44	+1.17	+0.70	+0.52
SHOP	+2.63	+1.50	+2.70	+2.29	−0.07
META	+15.39	+12.94	+16.04	+14.58	−0.66
COIN	+9.41	+7.67	+9.46	+8.37	−0.04
CRWD	+9.09	+7.17	+8.86	+8.33	+0.23

Notes: Expected P&L per observation ($\times 10^3$ ticks), defined as fill rate \times mean P&L conditional on fill. Bold indicates the best-performing strategy for each stock. “RL–Best” is the RL advantage over the best non-RL alternative. RL outperforms the best heuristic in 17/20 stocks. Stocks are ordered by average quoted spread.

RL policy dominates: it outperforms the best of the two heuristics in seventeen of twenty stocks, and beats ALA in all twenty. The three exceptions—Coinbase, Meta, and Shopify—are widespread stocks where the queue-position heuristic is marginally better, but in each case the gap is economically trivial (-0.04 , -0.66 , and -0.07 in units of 10^{-3} ticks). These are stocks where every strategy earns positive returns and the margins between them are thin.

The RL policy’s advantage is largest precisely where it matters most. For Amazon, the RL policy earns $+0.86$ while the best heuristic (volatility) earns -5.38 —a gap of $+6.23$ (all in units of 10^{-3} ticks per observation). For Microsoft: RL earns $+1.56$ versus -4.07 for the queue heuristic—a gap of $+5.62$. For Apple: -0.19 versus -4.75 —a gap of $+4.56$. These are the moderate-spread stocks where the Q-value surface is richest and univariate rules cannot capture the interaction effects between queue position, depth, and volatility that the RL framework exploits.

The volatility heuristic beats ALA in eleven of twenty stocks; the queue heuristic in ten. Neither is universally helpful. The RL framework earns its keep precisely where the problem is hard: in tick-constrained, adversely-selected stocks where naive strategies lose money and no single state variable captures enough of the value surface to guide cancellation well. The RL policy, by integrating all state dimensions simultaneously, achieves what no single-variable rule can.

7.3 RL Versus Naive Baselines

Against the simpler baselines, the RL policy’s dominance is complete. It outperforms ALA in all twenty stocks without exception (Table 4, rightmost column versus ALA column). For Amazon, the RL policy converts an expected P&L of -6.55 ($\times 10^{-3}$ ticks per observation, under ALA) into $+0.86$ —a swing from losing money to making money. The mechanism is fill quality: under ALA,

Amazon fills average -0.356 ticks per fill; under the RL policy, they average $+0.082$ ticks. The RL policy does not merely avoid fills—it avoids *bad* fills, accepting execution only in states where the expected outcome is favorable.

Apple provides an instructive edge case. It is the only stock where the RL policy still shows a negative mean fill P&L (-0.051 ticks). But this is consistent with Apple’s uniformly negative Q-values: even the “best” states carry residual adverse selection. The RL policy reduces losses from -0.330 ticks per fill (under ALA) to -0.051 —an 85% improvement. The cancel option does not eliminate adverse selection; it manages it.

The Random and Contrarian Transfer baselines confirm that the RL advantage is not an artifact of selective cancellation alone. Random cancellation produces expected P&L statistically indistinguishable from ALA for most stocks, and the Contrarian policy—which applies the most spread-distant stock’s rule—performs no better than passive holding, corroborating the transfer regret analysis in Section 6.

7.4 Validating the Q-Learning Estimates

The strongest evidence that the RL framework captures real economic structure comes from the correlation between theoretical option values and realized performance gaps. For each stock, we compute the theoretical *OptVal* (from value iteration) and the realized gap between RL and ALA expected P&L (from the backtest). If the Q-values are well-calibrated, stocks with higher *OptVal* should show larger RL advantages.

The correlation is $r = 0.72$ ($p < 0.001$). This is the headline validation statistic. It confirms that the cross-stock variation in Q-learning estimates is not noise: stocks where the framework predicts a large cancel option value are precisely the stocks where the RL policy delivers the largest realized gains over passive alternatives.

The relationship is approximately linear, with a slope indicating that each 0.01-tick increase in theoretical *OptVal* translates to roughly a 0.0004-tick increase in realized RL advantage per observation. The modest magnitude reflects the difference between the theoretical optimum (which assumes perfect state observation) and the realized backtest (which is subject to bin-boundary approximation and transition estimation error). What matters is not the level but the rank correlation: the framework correctly identifies which stocks benefit most from active cancellation management.

8. DISCUSSION

8.1 Implications for Tick-Size Reform

The SEC’s ongoing consideration of tick-size reform under Rule 612 proposes reducing the minimum tick for certain stocks from $\$0.01$ to $\$0.005$. Five stocks in our sample—Rivian, Snap, Intel, SoFi, and Lucid—have time-weighted average credited spreads below $\$0.015$ and would likely be affected.

Our cross-sectional results provide a framework for interpreting the consequences, though we caution against point predictions given the partial-equilibrium nature of our analysis. Three qualitative implications follow from the patterns we document.

First, the monotonic spread–value relationship (Figure 1) implies that halving the tick for constrained stocks would, all else equal, increase the effective spread in tick units, pushing these stocks rightward along the value curve. Best-bid Q-values should increase modestly as the constraint relaxes and spreads can adjust to reflect true adverse selection costs more precisely.

Second, the inverted-U in cancel option value (Figure 3) implies that affected stocks would move *toward* the *OptVal* peak rather than away from it. Currently, these stocks sit in the flat-left tail of the inverted-U where there is little cross-state variation to exploit. Under a finer tick, the richer price grid would create more distinct states and a more informative cancellation decision, raising the value of active order management.

Third, the transfer regret analysis (Section 6) implies that the regime shift would be discontinuous. A stock that currently clusters with other one-tick names would, after reform, sit in a two-tick neighborhood with different optimal policies. Algorithms trained on pre-reform data would require recalibration—a practical consideration for market makers managing the transition.

We emphasize that these are directional implications, not quantitative forecasts. Our framework is partial-equilibrium: it takes the transition dynamics as given and solves for the optimal policy. A tick-size change would endogenously alter those dynamics—depth, queue lengths, cancellation rates, and the competitive structure of liquidity provision would all adjust. A full general-equilibrium analysis is beyond the scope of this paper but represents a natural extension.

8.2 Economic Magnitude

A natural concern with our results is that the P&L magnitudes are small—fractions of a tick per 100ms observation. This is true in per-observation terms but misleading in economic terms. A market maker submitting orders at 100ms intervals faces 234,000 decision points per stock per day. Even a 0.001-tick advantage per observation, applied across a typical order size and multiplied over the trading day, compounds into meaningful dollar amounts.

Consider Amazon, where the RL policy’s advantage over ALA is 0.0071 ticks per observation. At \$0.01 per tick and a hypothetical position of 100 shares, this translates to approximately \$1.66 per day per stock. Across a portfolio of twenty stocks, this becomes \$33 per day or approximately \$8,000 per year. For a market maker operating with larger positions and across more names, the economics scale proportionally. The value of the RL framework is not in the per-observation magnitude but in its consistency and breadth. These findings connect to the broader literature on AI in financial decision-making (Fedyk et al., 2024; Kakhbod et al., 2024): the RL framework’s advantage over intuition-based heuristics illustrates how data-driven methods can overcome the cognitive constraints that limit human order management, particularly in the high-frequency, multi-dimensional environment of tick-constrained markets where simple rules fall short.

8.3 Limitations

Several limitations deserve acknowledgment. First, our state space discretization (16,875 working states) is coarse. Continuous variables are binned into 3–5 categories, and the results are necessarily conditional on the bin boundaries. While the cross-sectional patterns reflect order-of-magnitude spread variation that no reasonable bin perturbation would reverse, finer discretizations or function approximation methods (deep Q-learning, for instance) could capture subtleties that our tabular approach misses.

Second, our analysis is in-sample: the RL policy is trained on the full sixty-day sample and evaluated on the same data. Out-of-sample temporal validation—training on the first thirty days and testing on the last thirty—would provide a stronger test of the framework’s predictive power. We view this as a natural extension that would strengthen the backtest validation without likely changing the cross-sectional or intraday findings, which reflect persistent structural features of the market.

Third, we study a single sample period. Market conditions—volatility regimes, maker-taker fee structures, the competitive landscape among market makers—evolve over time. Our results describe the value of limit orders under one set of conditions and may not generalize to periods with materially different microstructure.

Fourth, the framework is purely empirical. We do not estimate a structural model of trader behavior or information arrival; we take the transition dynamics as given and solve for the optimal response. This means we cannot perform welfare analysis or predict how the equilibrium would change under counterfactual policies. Our results describe what *is* optimal given the data, not what *would be* optimal under different market designs.

Finally, the bin-boundary alignment between the value iteration stage (script 04 in our pipeline) and the backtest stage introduces a mechanical discrepancy for some stocks. Quantile boundaries computed on different subsamples of the data can assign the same observation to different bins, creating a wedge between theoretical and realized leave fractions. For twelve of twenty stocks, this wedge is small (less than ten percentage points). For the remaining eight, it is larger, though it does not correlate with the RL policy’s advantage ($r = 0.14$). Loading the value-iteration stage’s bin boundaries directly into the backtest would eliminate this discrepancy and represents a straightforward pipeline improvement.

9. CONCLUSION

We have applied a reinforcement learning framework to the stay-or-cancel decision facing limit orders in twenty NASDAQ-listed equities, exploiting the uniform U.S. penny tick to construct a cleaner cross-sectional comparison than has previously been possible. The results reveal that limit order values vary by a factor of eighteen across the tick-constraint spectrum and that this variation is organized by four interlocking patterns: a monotonic spread–value relationship, a price-level gradient that inverts for constrained stocks, an intraday U-shape in cancel option value, and a block-diagonal policy transfer structure governed by spread distance alone.

These findings have three implications. For market microstructure theory, they document that the concentration of adverse selection at the best bid is severe enough in tick-constrained markets to make deeper price levels more attractive—a gradient inversion that challenges the conventional wisdom that the best bid is always the most valuable position. For algorithmic trading practice, they demonstrate that state-dependent cancellation, calibrated by reinforcement learning, meaningfully outperforms both passive strategies and economically-motivated heuristics, and that spread distance is a sufficient statistic for determining whether a policy trained on one stock can be deployed on another. For regulatory policy, they provide a framework for interpreting the consequences of tick-size reform, suggesting that finer ticks would increase the value of active order management for currently constrained stocks while creating a discontinuous regime shift that would require algorithmic recalibration.

The Q-learning framework takes the market as it finds it and solves for the optimal response. What it reveals is that even in the most commoditized corner of modern markets (the resting limit order) there is rich, exploitable structure in the interaction of queue position, depth, spread, and volatility. The price of the queue varies with where you stand, when you stand there, and whether you have the option to walk away.

REFERENCES

- Admati, A. R. and Pfleiderer, P. (1988). A theory of intraday patterns: Volume and price variability. *Review of Financial Studies*, 1(1):3–40.
- Bell, S., Kakhbod, A., Lettau, M., and Nazemi, A. (2024). Glass box machine learning and corporate bond returns. NBER Working Paper.
- Bell, S., Kakhbod, A., Nazemi, A., Stanton, R., and Wallace, N. (2025). Fairness by design: Machine learning and interpretable mortgage lending. Available at SSRN 6221578.
- Bessembinder, H. (2003). Issues in assessing trade execution costs. *Journal of Financial Markets*, 6(3):233–257.
- Brogaard, J., Hagströmer, B., Norden, L., and Riordan, R. (2024). The economic impact of tick size regulation. Working paper.
- Cont, R., Stoikov, S., and Talreja, R. (2014). A stochastic model for order book dynamics. *Operations Research*, 58(3):549–563.
- Copeland, T. E. and Galai, D. (1983). Information effects on the bid-ask spread. *The Journal of Finance*, 38(5):1457–1469.
- Eisfeldt, A. L. and Schubert, G. (2024). AI and finance. NBER Working Paper.
- Fedyk, A., Kakhbod, A., Li, P., and Malmendier, U. (2024). AI and perception biases in investments: An experimental study. Available at SSRN 4787249.
- Foucault, T. (1999). Order flow composition and trading costs in a dynamic limit order market. *Journal of Financial Markets*, 2(2):99–134.
- Foucault, T., Kadan, O., and Kandel, E. (2005). Limit order book as a market for liquidity. *Review of Financial Studies*, 18(4):1171–1217.
- Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377.
- Glosten, L. R. (1994). Is the electronic open limit order book inevitable? *The Journal of Finance*, 49(4):1127–1161.
- Griffiths, M. D., Smith, B. F., Turnbull, D. A. S., and White, R. W. (2000). The costs and determinants of order aggressiveness. *Journal of Financial Economics*, 56(1):65–88.
- Handa, P. and Schwartz, R. A. (1996). Limit order trading. *The Journal of Finance*, 51(5):1835–1861.
- Hasbrouck, J. and Saar, G. (2013). Low-latency trading. *Journal of Financial Markets*, 16(4):646–679.
- Hochberg, Y. V., Kakhbod, A., Li, P., and Sachdeva, K. (2023). Are patents with female inventors under-cited? Evidence from text estimation. Available at SSRN 4269703.
- Hollifield, B., Miller, R. A., Sandas, P., and Slive, J. (2004). Empirical analysis of limit order markets. *Review of Economic Studies*, 71(4):1027–1063.
- Kakhbod, A., Kermani, A., and Maciel, B. (2025). In the Fed’s mind. Available at SSRN 5423614.
- Kakhbod, A., Kogan, L., Li, P., and Papanikolaou, D. (2024). Measuring creative destruction. MIT Sloan Research Paper.

- Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524.
- Kozak, S., Nagel, S., and Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, 135(2):271–292.
- Kwan, A. and Philip, R. (2025). Reinforcement learning in a dynamic limit order market. Working paper, University of New South Wales and University of Sydney.
- Lo, A. W., MacKinlay, A. C., and Zhang, J. (2002). Econometric models of limit-order executions. *Journal of Financial Economics*, 65(1):31–71.
- Nevmyvaka, Y., Feng, Y., and Kearns, M. (2006). Reinforcement learning for optimized trade execution. *Proceedings of the 23rd International Conference on Machine Learning*, pages 673–680.
- O’Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, 116(2):257–270.
- Parlour, C. A. (1998). Price dynamics in limit order markets. *Review of Financial Studies*, 11(4):789–816.
- Rindi, B. and Werner, I. M. (2020). U.S. tick size pilot. Working paper, Fisher College of Business.
- Wood, R. A., McInish, T. H., and Ord, J. K. (1985). An investigation of transactions data for NYSE stocks. *The Journal of Finance*, 40(3):723–739.
- Yao, C. and Ye, M. (2018). Why trading speed matters: A tale of queue rationing under price controls. *Review of Financial Studies*, 31(6):2157–2193.