Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Market Informedness and Market-Maker Profitability: The Trade-Off Between Adverse Selection and Price Discovery

Konrad Jerzy Ochedzan

kochedzan@student.ethz.ch

ETH Zurich

Nino Antulov-Fantulin

ETH Zurich

Research Article

Keywords: market microstructure, market making, limit order book, informed trading, price discovery, reinforcement learning, Hawkes processes

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10076901/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Market Informedness and Market-Maker Profitability: The Trade-Off Between Adverse Selection and Price Discovery

Konrad Ochedzan[1*] and Nino Antulov-Fantulin[2,3]

[1]Master of Science in Quantitative Finance, University of Zurich and ETH Zurich, Zurich, Switzerland.
[2]Computational Social Science, Department of Humanities, Social and Political Sciences, ETH Zurich, Zurich, Switzerland.
[3]Aisot Technologies, Zurich, Switzerland.

*Corresponding author(s). E-mail(s): konrad.ochedzan@gmail.com;

Abstract

This paper studies how market informedness affects market makers' profitability in a computational market environment with heterogeneous learning agents. We develop an agent-based market model in which market makers differ in their information sets and inventory-risk aversion, prices form endogenously, fundamental values evolve exogenously, and market-taker order flow follows a state-dependent self-exciting process. The model provides a controlled computational laboratory for analyzing the interaction between informed trading, adverse selection, price discovery, and liquidity provision. We establish finite-horizon stability properties of the market-taker order-flow process and solve the market-making problem using multi-agent reinforcement learning with centralized training and decentralized execution. The results show that informed market order flow is particularly harmful when aggregate market informedness is low, exposing market makers to severe adverse-selection risk. However, as market informedness increases, market-maker profitability displays an overall upward trend despite local non-monotonicities arising from complex market dynamics and stochastic learning. This suggests that the price-discovery benefits of informed trading can offset its adverse-selection costs. The findings contribute to computational economics by showing how agent heterogeneity, endogenous price formation, and learning-based liquidity provision jointly shape market outcomes.

Keywords: market microstructure, market making, limit order book, informed trading, price discovery, reinforcement learning, Hawkes processes

1 Introduction

Modern financial markets are dynamic systems in which liquidity demand, liquidity supply, information, and strategic behavior interact at high frequency. At coarse time scales, asset prices are often represented by diffusion-type models, with the Black–Scholes framework providing a canonical benchmark for continuous-time price dynamics (Black and Scholes (1973)). At the transaction level, however, the relevant object is the market microstructure: trades and quotes are discrete, prices move in ticks, and the state of the market is summarized by the limit order book. This micro-level structure can generate volatility, clustering, and temporary deviations between transaction prices and latent economic value that are not visible in aggregate price models.

A central difficulty in modeling such markets is that order flow is both strategically and statistically heterogeneous. Behavioral feedback effects such as herding and belief-driven trading can generate bursts of activity and short-lived price dislocations (Gennaioli and Shleifer (2018); Shleifer (2000)). Informational heterogeneity creates a separate but related mechanism: some traders possess, infer, or react more quickly to information about the asset's fundamental value, while others primarily condition on observable market data. This gives rise to adverse selection because liquidity providers are more likely to trade against informed market orders when quoted prices are temporarily misaligned with fundamentals (Barucci, Mathieu, and Sanchez-Betancourt (2025); Campi and Zabaljauregui (2020); Glosten and Milgrom (1985)).

These mechanisms are especially important for market makers. Market makers supply liquidity by posting bid and ask limit orders, earn compensation through the bid–ask spread and exchange rebates, and face losses from inventory accumulation and adverse selection. Their profitability therefore depends not only on their own quoting policy, but also on the informational content of order flow, the endogenous response of prices, and the behavior of competing liquidity providers. This interaction is difficult to capture analytically because realistic market-making environments combine a discrete limit order book, clustered and size-heterogeneous market orders, partial information about fundamentals, and strategic competition among multiple agents.

This paper studies these interactions in a computational market environment. The model contains heterogeneous market makers with different information sets and inventory-risk aversion, a latent fundamental value, endogenous price formation through the limit order book, and market-taker order flow generated by a state-dependent marked Hawkes mechanism. The resulting environment is used as a controlled computational laboratory for analyzing how market informedness affects market-maker profitability through the joint channels of adverse selection and price discovery.

The learning component is based on reinforcement learning, which is well suited to sequential decision problems in large state spaces where closed-form optimal controls are unavailable (Schulman, Wolski, Dhariwal, Radford, and Klimov (2017); Sutton and Barto (2018); Szepesvári (2010)). Because the environment contains multiple strategic liquidity providers, the agents are trained using a multi-agent approach with centralized training and decentralized execution, a paradigm developed for interacting decision makers with non-stationary learning dynamics (Lowe et al. (2017); Yu et al.

(2022)). This allows market makers to learn quoting policies from repeated interaction with the simulated market rather than from an exogenously specified rule.

The main finding is that market informedness has a non-trivial effect on liquidity provision. Informed market order flow is most damaging when aggregate market informedness is low, because market makers are then exposed to adverse-selection losses without sufficient price discovery. As aggregate informedness increases, market-maker profitability exhibits an overall upward trend despite local non-monotonicities caused by complex market dynamics and stochastic learning. The results suggest that the price-discovery benefits of informed trading can offset, and eventually dominate, its adverse-selection costs in a sufficiently informative market.

2 Related Literature

This paper relates first to the market-microstructure literature on spreads, dealer behavior, and liquidity provision. Classic models explain the bid–ask spread as compensation for transaction costs, immediacy services, and inventory risk faced by dealers and liquidity suppliers (Amihud and Mendelson (1980); Demsetz (1968); Garman (1976); Ho and Stoll (1981); Stoll (1978); Tinic (1972)). Related work studies order-placement incentives and competition among liquidity suppliers, showing that the spread and execution outcomes depend on the strategic structure of the trading mechanism (Cohen, Maier, Schwartz, and Whitcomb (1981); Ho and Stoll (1983); Mildenstein and Schleef (1983)). Information-based models add an adverse-selection component to the spread by showing that market makers protect themselves against the possibility of trading with better-informed counterparties (Copeland and Galai (1983); Glosten and Milgrom (1985)).

A second strand studies optimal market making and trading under inventory risk, signals, and partial information. The Avellaneda–Stoikov framework formulates high-frequency market making as an inventory-sensitive quoting problem around a reference price (Avellaneda and Stoikov (2008)). Subsequent models incorporate order-flow signals, alpha signals, temporary and transient price impact, and speculative trading motives into optimal execution and market-making problems (Bank, Cartea, and Körber (2023); Cartea and Jaimungal (2016); Cartea and Wang (2020); Drissi (2022); Lehalle and Neuman (2019); Neuman and Voß (2022)). Partial-information formulations explicitly connect market-making decisions to latent states and stochastic intensities, while recent work with fads, informed traders, and uninformed traders studies how the composition of market takers affects market-maker performance (Barucci et al. (2025); Campi and Zabaljauregui (2020)). The present paper differs from this line by making transaction prices endogenous to the interaction of several heterogeneous market makers and a state-dependent stream of market orders.

The paper also builds on statistical models of order-flow dynamics. Hawkes processes provide a tractable representation of self-exciting arrivals, where one event temporarily increases the likelihood of subsequent events (Hawkes (1971)). Intensity-based point-process models have been used to describe security-market events in continuous time (Bowsher (2007)). Empirical studies of limit order books document clustered activity, dependence in order signs, and rich volume dynamics (Chakraborti,

Muni Toke, Patriarca, and Abergel (2011); Gould et al. (2013)). Hawkes models have been applied to high-frequency market data and systemic price co-jumps, capturing temporal clustering beyond Poisson order-arrival assumptions (Bormetti et al. (2015); Filimonov and Sornette (2015)). Marked and multivariate extensions incorporate order size and cross-excitation across event types, which is important because large market orders can walk the book and generate discontinuous quote changes (Jain, Firoozye, Kochems, and Treleaven (2024); Rambaldi, Bacry, and Lillo (2017)). Our model uses this tradition to generate clustered and size-heterogeneous market-taking events, but modulates their intensities by the relationship between the market price and a latent fundamental value.

Finally, the paper contributes to computational and agent-based approaches to financial markets. Agent-based market simulations have been used to study how heterogeneous traders and trading rules can generate aggregate market phenomena that are difficult to derive in closed form (Paddrik et al. (2012)). Algorithmic-trading references provide the broader market-design and execution context for these simulation-based approaches (Cartea, Jaimungal, and Penalva (2015)). Reinforcement learning provides a general framework for sequential decision-making and policy optimization in complex environments (Schulman et al. (2017); Sutton and Barto (2018); Szepesvári (2010)). Multi-agent reinforcement learning extends this framework to interacting agents whose learning problems are coupled through the environment and through each other's policies (Lowe et al. (2017); Yu et al. (2022)). Recent applications to market making and limit order books show that learning-based agents can be used to analyze quoting behavior, equilibrium formation, and endogenous price dynamics in simulated markets (Gašperov, Begušić, Posedel Šimović, and Kostanjčar (2021); He and Lin (2022)). Closely related computational market environments combine deep learning, queue-reactive order-book simulation, and multi-agent reinforcement learning to study realistic limit-order-book behavior (Bodor and Carlier (2025); Cheridito, Dupret, and Wu (2025)). Relative to these studies, the present paper focuses on how market informedness changes the balance between adverse selection and price discovery for heterogeneous learning market makers.

3 Contributions

This paper makes three contributions.

First, it develops a computational market environment with heterogeneous market-making agents, endogenous price formation, state-dependent self-exciting order flow, and learning-based decision-making.

Second, it provides stability guarantees for the resulting market-taker process, ensuring that the simulated market dynamics are well defined over finite horizons.

Third, it uses this environment to study how market informedness affects liquidity provision. The results show that informed order flow is most harmful in poorly informed markets, while higher aggregate informedness can improve market-maker profitability through better price discovery.

4 Computational Market Environment

This section describes the simulation environment used throughout the paper. The objective is to obtain an order-driven market in which (i) liquidity demand arrives in continuous time with empirically realistic clustering, (ii) trades have heterogeneous sizes, and (iii) arrival rates react to both microstructure conditions (spread, depth) and mispricing relative to an exogenous fundamental value.

4.1 Notations

Market makers act on a fixed decision grid and interact through a shared limit order book (LOB), while market takers generate trades in event time. Let $0 = t_0^{\text{dec}} < t_1^{\text{dec}} < \cdots < t_{N_{\text{steps}}}^{\text{dec}} = T$ denote agents' decision times. We assume the following notations.

Orders end up in a *queue* which accounts for the order of limit orders. That leads to a situation, where at each point of time there is a finite list of available prices at each side. We denote a list of open asks' price levels at time t as $\mathcal{A}_t$ and bids' as $\mathcal{B}_t$. In the LOB for each price p the book maintains a first-in-first-out (FIFO) queue (price–time priority) of resting orders:

$$\mathcal{Q}_t^b(p) = \big((j^b_{1,\ p}, q^b_{1,\ p}), (j^b_{2,\ p}, q^b_{2,\ p}), \ldots, (j^b_{n_p,\ p}, q^b_{n_p,\ p})\big)_t, \tag{1}$$

$$\mathcal{Q}_t^a(p) = \big((j^a_{1,\ p}, q^a_{1,\ p}), (j^a_{2,\ p}, q^a_{2,\ p}), \ldots, (j^a_{n_p,\ p}, q^a_{n_p,\ p})\big)_t, \tag{2}$$

where $(j^x_{m,\ p}, q^x_{m,\ p})$ denotes an order of size $q^x_{m,\ p}$ posted by entity $j^x_{m,\ p}$ on side $x \in \{\text{bid}, \text{ask}\}$ and price level p. Importantly, all prices live on a fixed tick grid with tick size Δp. For the simulation purposes to make sense out of agents actions and prevent floating point drift, every price that enters the book is first *canonicalized* by rounding to the appropriate number of decimals implied by Δp. This guarantees that identical economic price levels map to the same dictionary key. Throughout, we denote by $\text{canon}(\cdot)$ this rounding operation. Thanks to that, in cases of odd length spread, the offers stay on the tick grid.

Whenever a market order fills a limit order, the market maker who quoted that offer may (in the implementation does) receive a rebate $R_{\text{rebate}} > 0$. That serves as additional motivation to take the risks of market making.

The *best bid* is the highest available buy price; the *best ask* is the lowest available sell price. Their difference is the *bid–ask spread*, and their average is the *mid-price*. The total available quantity at each level constitutes the *depth*. In this paper the best bid is denoted as $B_t = \max\{p : p \in \mathcal{B}_t\}$ and the best ask as $A_t = \min\{p : p \in \mathcal{A}_t\}$. The liquidity at best ask and liquidity at best bid are denoted respectively as $Q_t^a = \sum_{(j_i, q_i) \in Q_t^a(A_t)} q_i$ and $Q_t^b = \sum_{(j_i, q_i) \in Q_t^b(B_t)} q_i$.

We model the market price defined as last transaction price P_t^{last} as the last price at which a trade occurred before or at time t. Market mid is defined as:

$$M_t = \frac{B_t + A_t}{2}, \tag{3}$$

and market micro-price as:

$$P_t^{\text{micro}} = \frac{A_t\, Q_t^b + B_t\, Q_t^a}{Q_t^a + Q_t^b}, \tag{4}$$

whenever both sides of the book are non-empty. If either side is empty, we fallback to the last transaction price P_t^{last} which is the last price at which a trade occurred before or at time t.

Each market maker aggregates both inventory and cash which for agent i at time t we denote as respectively $q_t^{(i)}$ and $c_t^{(i)}$.

4.2 Fundamental value

The fundamental value is modeled as an exogenous process. Historically, asset prices have often been modeled by geometric Brownian motion or by richer specifications designed to reproduce stylized volatility patterns. The present paper, however, focuses on intraday market microstructure rather than on long-horizon asset pricing. For this reason, and similarly to the role played by the reference-price process in Avellaneda and Stoikov (2008), we model the fundamental value by a standard Brownian motion. This choice keeps the specification parsimonious and is appropriate for the short horizons considered here, over which both drift and the risk-free rate can be treated as negligible. In addition, introducing a process with an average one-sided trend would create stronger incentives for agents to take directional bets, thereby generating inventory and wealth distributions that are less representative of intraday market making. Therefore, we model the fundamental as a Brownian motion with zero drift:

$$S_t = S_0 + \sigma W_t,$$

where W_t is a Brownian motion, S_0 is the initial price, and $\sigma > 0$ is the volatility. Since an arithmetic Brownian motion can in principle cross zero, the simulator imposes an economically natural lower bound at zero: in the implementation, zero is treated as a hard floor and the fundamental price is not allowed to fall below it.

We also introduce a notion of a "noisy estimator" of the fundamental value so that some agents can be given additional information about the market. This noisy estimator is denoted by $\widetilde{S}_t^{(i)}$ and defined as $\widetilde{S}_t^{(i)} = S_t + \varepsilon_t^{(i)}$, where $\varepsilon_t^{(i)} \sim \mathcal{N}(0, \sigma_{signal}^{(i)})$ is an exogenous, agent-specific noise term.

4.3 Market takers

We model market taking events as a multivariate marked point process. Following Rambaldi et al. (2017), the mark jointly specifies the trade direction (consuming the ask side or the bid side) and a discrete size bin. Let K denote the number of size bins on each side and define $C = 2K$. We index event types by

$$i \in \{1, \dots, K\} \quad \text{(market buy events consuming asks)},$$

$$i \in \{K+1, \dots, 2K\} \quad \text{(market sell events consuming bids)}.$$

For each type i we introduce a counting process $N_i(t)$ and an $\mathcal{F}_t$-predictable intensity $\lambda_i(t)$, where $\mathcal{F}_t$ is the natural filtration of the process.

Raw Hawkes dynamics.

To capture self-excitation, we first define a *raw* Hawkes intensity vector $\boldsymbol{\lambda}^H(t) = (\lambda_1^H(t), \dots, \lambda_C^H(t))^\top$ as in Hawkes (1971):

$$\lambda_i^H(t) = \mu_i + \sum_{j=1}^{C} \int_0^t \phi_{i \leftarrow j}(t-s)\, \mathrm{d}N_j(s), \qquad i = 1, \dots, C, \tag{5}$$

with baseline rates $\mu_i > 0$ and excitation kernels $\phi_{i \leftarrow j}(\cdot) \geq 0$. As standard in microstructure Hawkes modeling and consistent with Rambaldi et al. (2017), we assume a sum-of-exponentials parameterization:

$$\phi_{i \leftarrow j}(u) = \sum_{r=1}^{R} \alpha_{ijr}\, e^{-\beta_{ijr} u}, \qquad u \geq 0, \tag{6}$$

where $\alpha_{ijr} \geq 0$ and $\beta_{ijr} > 0$. This specification yields a piecewise-deterministic representation via auxiliary excitation states

$$Z_{ijr}(t) = \int_0^t \alpha_{ijr} e^{-\beta_{ijr}(t-s)}\, \mathrm{d}N_j(s), \qquad \lambda_i^H(t) = \mu_i + \sum_{j=1}^{C} \sum_{r=1}^{R} Z_{ijr}(t),$$

with exponential decay between jumps and instantaneous increments at event times.

Microstructure and information modulation.

The raw Hawkes dynamics account for clustering and cross-excitation across sizes and sides, but do not incorporate the dependence of trade arrivals on current market conditions or on mispricing relative to fundamentals. We introduce this dependence through a multiplicative gain vector $G(x(t)) \in \mathbb{R}_+^C$, where $x(t)$ is a collection of the LOB and fundamental-value features. We then construct:

$$\delta_t^a := A_t - P_t^{\text{micro}} \geq 0, \qquad \delta_t^b := P_t^{\text{micro}} - B_t \geq 0, \qquad I_t := S_t - P_t^{\text{micro}}. \tag{7}$$

The gain G reflects two effects: (i) sensitivity to instantaneous spread conditions captured by δ_t^a, δ_t^b and (ii) information-driven trading pressure when the observed market price deviates from the fundamental value via I_t. Following the mixture-of-informed approach in Barucci et al. (2025), we assume that a fraction $\psi \in [0,1]$ controls the market informedness. We define spread sensitivity $\eta > 0$ and mispricing sensitivity $\zeta > 0$ equal for both sides and all size bins. We then set the uninformed and informed gains as

$$g^{U,a}(x(t)) = \exp\big(-\eta\, \delta_t^a\big), \qquad g^{I,a}(x(t)) = \exp\big(-\eta\, \delta_t^a + \zeta\, I_t\big), \tag{8}$$

$$g^{U,b}(x(t)) = \exp\big(-\eta\,\delta_t^b\big), \qquad g^{I,b}(x(t)) = \exp\big(-\eta\,\delta_t^b - \zeta\,I_t\big). \tag{9}$$

The effective gains (mixture over trader types) are

$$G^a(x(t)) = \psi\,g^{I,a}(x(t)) + (1-\psi)\,g^{U,a}(x(t)), \qquad G^b(x(t)) = \psi\,g^{I,b}(x(t)) + (1-\psi)\,g^{U,b}(x(t)), \tag{10}$$

and we construct $G(x(t)) \in \mathbb{R}^C$ by concatenation:

$$G_i(x(t)) = \begin{cases} G^a(x(t)), & i = 1, \dots, K, \\ G^b(x(t)), & i = K+1, \dots, 2K. \end{cases}$$

If one side of the book is empty, then the corresponding gains are set to zero.
The *effective* intensity used for simulation is:

$$\lambda_i(t) = G_i(x(t))\,\lambda_i^H(t), \qquad i = 1, \dots, C, \tag{11}$$

and the total event rate is

$$\Lambda_{\text{eff}}(t) = \sum_{i=1}^{C} \lambda_i(t) = \sum_{i=1}^{C} G_i(x(t))\,\lambda_i^H(t). \tag{12}$$

That notion is a natural extension of Barucci et al. (2025) where we would take $C = 2$ for only unitary order volumes at both sides, null Hawkes kernels and constant baseline Hakwes intensities μ_i. Therefore, this research serves as an extension in a more general setup to the aforementioned paper.

Size realization within bins.

An event type determines only the side and the size bin. Conditional on an event in bin k, the executed quantity is sampled as an integer uniformly from a pre-specified range associated with that bin. This binning is the mechanism that produces heterogeneous order sizes while keeping the Hawkes dimension finite and relatively condensed, as in Rambaldi et al. (2017).

Warm start

To avoid an artificial burn-in period at the beginning of each episode, we initialize the raw Hawkes state in a stationary regime of the linear Hawkes model (5)–(6). Let $A \in \mathbb{R}^{C\times C}$ be the integrated kernel matrix with entries

$$A_{ij} = \int_0^\infty \phi_{i\leftarrow j}(u)\,\mathrm{d}u = \sum_{r=1}^{R} \frac{\alpha_{ijr}}{\beta_{ijr}}. \tag{13}$$

If the spectral radius $\rho(A) < 1$, the stationary mean intensity $\boldsymbol{\lambda}^* = \mathbb{E}[\boldsymbol{\lambda}^H(t)]$ exists and satisfies

$$\boldsymbol{\lambda}^* = \boldsymbol{\mu} + A\,\boldsymbol{\lambda}^*, \qquad \text{hence} \qquad \boldsymbol{\lambda}^* = (I - A)^{-1}\boldsymbol{\mu}.$$

In the simulation, we set $\boldsymbol{\lambda}^H(0) := \boldsymbol{\lambda}^*$ and initialize the excitation variables $Z_{ijr}(0)$ so that $\lambda_i^H(0) = \mu_i + \sum_{j,r} Z_{ijr}(0)$. Concretely, we place the "excess over baseline" $\lambda_i^* - \mu_i$ on the diagonal $j = i$ and split it evenly across the R exponential components:

$$Z_{ijr}(0) = \begin{cases} \dfrac{\lambda_i^* - \mu_i}{R}, & i = j, \\ 0, & i \neq j. \end{cases}$$

This choice ensures $\lambda_i^H(0) = \lambda_i^*$ while keeping the initialization simple and reproducible.

Continuous time simulation

Market-taker arrivals are simulated in continuous time using an event-driven procedure based on Ogata's thinning / time-change methodology (Daley and Vere-Jones (2003); Ogata (1981)). The market makers, in contrast, act on a fixed decision grid. Over each interval $(t_n^{\text{dec}}, t_{n+1}^{\text{dec}})$, where t_n^{dec} is agents' n-th decision time, agents do not submit new quotes, so the control is held fixed. However, taker trades occurring within the interval consume resting depth and may change the best quotes and queue sizes. Accordingly, the effective intensities are piecewise defined between event times: between consecutive events the raw Hawkes state decays deterministically, while after each executed trade the LOB-derived multiplier $G(x(t))$ is recomputed from the updated book.

after each event we may draw $S \sim \text{Exp}(1)$ and define the next event time as the solution of

$$\int_t^{t_{\text{next}}} \Lambda_{\text{eff}}(s)\, \mathrm{d}s = S. \tag{14}$$

Between events, the Hawkes excitation variables decay exponentially:

$$Z_{ijr}(t+u) = Z_{ijr}(t)\, e^{-\beta_{ijr} u}, \qquad u \geq 0,$$

so the raw intensities evolve as

$$\lambda_i^H(t+u) = \mu_i + \sum_{j=1}^{C} \sum_{r=1}^{R} Z_{ijr}(t)\, e^{-\beta_{ijr} u}.$$

Within a time segment over which $G(x(\cdot))$ is held fixed, the integrated effective intensity admits a closed form. Denoting the segment length by Δ, we obtain

$$\int_t^{t+\Delta} \Lambda_{\text{eff}}(s)\, \mathrm{d}s = \sum_{i=1}^{C} G_i(x(t)) \left[\mu_i\, \Delta + \sum_{j=1}^{C} \sum_{r=1}^{R} Z_{ijr}(t)\, \frac{1 - e^{-\beta_{ijr}\Delta}}{\beta_{ijr}} \right]. \tag{15}$$

Therefore, on each segment the mapping $\Delta \mapsto \int_t^{t+\Delta} \Lambda_{\text{eff}}(s)\, \mathrm{d}s$ is continuous and strictly increasing, which makes (14) well-posed. In the implementation, the root-finding step is performed by bisection on Δ using (15).

We do not redraw the exponential threshold at every market-maker decision time. Instead, we maintain a remaining "area to the next event" S^{rem} corresponding to the left-hand side of (14). On each time segment $[t, t_{\mathrm{end}}]$ over which the multiplier $G(x(\cdot))$ is held fixed and where t_{end} stands for the next agents' decision step, we define the integrated hazard by

$$H(t, t_{\mathrm{end}}) := \int_t^{t_{\mathrm{end}}} \Lambda_{\mathrm{eff}}(s)\, \mathrm{d}s.$$

Thus, $H(t, t_{\mathrm{end}})$ is precisely the cumulative compensator accumulated over the current segment.

- If $H(t, t_{\mathrm{end}}) < S^{\mathrm{rem}}$, then no trade occurs before t_{end}. We advance the clock to t_{end}, update G according to market-maker actions at t_{end}, set $S^{\mathrm{rem}} \leftarrow S^{\mathrm{rem}} - H(t, t_{\mathrm{end}})$, and proceed.
- If $H(t, t_{\mathrm{end}}) \geq S^{\mathrm{rem}}$, then a trade occurs inside $(t, t_{\mathrm{end}}]$. We solve for t_{next} such that $\int_t^{t_{\mathrm{next}}} \Lambda_{\mathrm{eff}}(s)\, \mathrm{d}s = S^{\mathrm{rem}}$, advance to t_{next}, and sample the event type. The Hawkes state is then updated by applying a jump in the realized dimension:

  $$Z_{ijr}(t_{\mathrm{next}}^+) = Z_{ijr}(t_{\mathrm{next}}) + \alpha_{ijr}\, \mathbf{1}\{j = \text{realized type}\}.$$

  After executing the corresponding market order in the LOB, we draw a fresh $S^{\mathrm{rem}} \sim \mathrm{Exp}(1)$ representing the waiting time for the next taker event and continue.

If the next market-maker decision time lies before the unconstrained solution to (14), the procedure above naturally splits the integrated intensity across segments and carries S^{rem} forward. This coupling ensures that market-taker trades occur in true continuous time, while market-maker actions remain on the fixed decision grid, and the exponential time-change mechanism stays consistent across all regime and state updates.

5 Stability of the modulated Hawkes mechanism

The multiplicative gains in (10) make the market-taker process state dependent and potentially much more active than a standard linear Hawkes process. In particular, when the microprice is below the fundamental value, buy-side gains are amplified, and when the microprice is above the fundamental value, sell-side gains are amplified. This raises a stability question: can the exponential gain in $|I_t|$ lead to explosive order arrivals or uncontrolled mispricing within one episode? The following results show that, on finite horizons, the finite depth of the LOB provides a stabilizing feedback mechanism. The amplified side consumes finite resting depth and therefore cannot generate infinitely many trades before the next decision time. Detailed proofs and auxiliary lemmas are given in Appendix A.

For the stability statements, call an event type *corrective* at time t if it trades in the direction that is favored by the sign of the pre-jump mispricing: buy types $i \leq K$

are corrective when $I_{t^-} \geq 0$, and sell types $i > K$ are corrective when $I_{t^-} < 0$. The other types are called *destabilizing*. Let T_{ecs} denote the total time during $[0, T]$ at which the corrective side of the book is empty:

$$T_{\text{ecs}} := \int_0^T \mathbf{1}_{\{\text{corrective side empty at } t^-\}} \, \mathrm{d}t. \tag{16}$$

Assumptions for the stability analysis.

For the stability results in this subsection, we work on a fixed finite horizon $[0, T]$ and use the following assumptions. These assumptions are in accordance with the market environment described above: they formalize the depth, execution, quote-range, and fundamental-value specifications imposed by the simulator, rather than additional restrictions on the learned quoting policies. Assumptions (S1)–(S6) are sufficient for non-explosion and the exponential mispricing bounds; the pathwise tail bound additionally uses the Brownian fundamental specification in (S7).

(S1) Positive gain parameters.
The gain parameters satisfy $\eta > 0$, $\zeta > 0$, and $\psi \in (0, 1]$.

(S2) Finite posted depth.
At every decision time t_n^{dec}, the total resting quantity on each side of the book is bounded by a deterministic constant $D_{\text{tot}} < \infty$. This bound is on aggregate depth across all market makers.

(S3) Minimum executable size.
There is a constant $q_{\min} > 0$ such that every executed market order removes at least $q_{\min}$ units from the corresponding side. If less than $q_{\min}$ units are available on a side, then no market order against that side is executed.

(S4) Empty-side shutdown.
An empty side cannot be hit by further market orders: if the ask side is empty, then $G^a = 0$, and if the bid side is empty, then $G^b = 0$.

(S5) Finite decision grid.
The number of market-maker decision times is finite and denoted by N_{steps}. When a rate per unit time is required, the grid is assumed to satisfy $t_{n+1}^{\text{dec}} - t_n^{\text{dec}} \geq \Delta_{\min}^{\text{dec}} > 0$.

(S6) Bounded quote distances.
There exists $\delta_{\max} < \infty$ such that, whenever the relevant side of the book is non-empty,

$$0 \leq \delta_t^a \leq \delta_{\max}, \qquad 0 \leq \delta_t^b \leq \delta_{\max}.$$

In the implementation, this bound is maintained by restricting the admissible posting range and cancelling quotes that move outside it after continuous-time taker trading.

(S7) Brownian fundamental value.
The fundamental value follows the Brownian specification introduced in Section 4.2,

$$S_t = S_0 + \sigma W_t, \qquad \sigma > 0,$$

where W is a standard Brownian motion.

Proposition 1 (Non-explosion) *Set*

$$N_{\max} := \frac{2D_{\text{tot}}N_{\text{steps}}}{q_{\min}}. \tag{17}$$

Under Assumptions (S1)–(S6)*, the total number of market-taker events satisfies*

$$N([0,T]) := \sum_{i=1}^{C} N_i([0,T]) \le N_{\max} \qquad a.s. \tag{18}$$

In particular, $\mathbb{E}[N([0,T])] \le N_{\max}$.

Economically, finite posted depth, minimum trade sizes, and empty-side shutdowns prevent an infinite cascade of market orders over a finite trading horizon. This is usefull in mathematical modelling of the market as well as grants a decent bound on what maximal level of activity can be expected under given market conditions. Proposition 1 is an extension of standard non-explosion results for Hawkes processes to the environment presented in this paper. Let $\mu_{\min} := \min_{1\le i\le C} \mu_i$ and

$$\gamma := \psi\, e^{-\eta\delta_{\max}}\, \mu_{\min}. \tag{19}$$

This is the uniform coefficient in the lower bound for corrective intensity during periods in which the corrective side of the book is non-empty.

Proposition 2 (Exponential mispricing integrability) *Under Assumptions* (S1)–(S6)*,*

$$\mathbb{E}\left[\int_0^T e^{\zeta|I_{t^-}|}\mathbf{1}_{\{\textit{corrective side non-empty at } t^-\}}\,\mathrm{d}t\right] \le \frac{N_{\max}}{\gamma}. \tag{20}$$

Moreover, on the full interval,

$$\mathbb{E}\left[\int_0^T e^{\zeta|I_{t^-}|}\,\mathrm{d}t\right] \le \frac{N_{\max}}{\gamma} + \mathbb{E}\Big[e^{\zeta\sup_{t\le T}|I_{t^-}|}T_{\text{ecs}}\Big]. \tag{21}$$

The second term is the contribution of periods during which the corrective side is empty.

Economically, the result captures the model's price-discovery force: when the microprice deviates from fundamentals and corrective liquidity is available, informed market orders discipline the deviation, making adverse selection for market makers part of the same mechanism that stabilizes prices.

Corollary 3 (Occupation-time bound) *For every* $M > 0$*,*

$$\mathbb{E}\left[\int_0^T \mathbf{1}_{\{|I_{t^-}|>M\}}\mathbf{1}_{\{\textit{corrective side non-empty at } t^-\}}\,\mathrm{d}t\right] \le \frac{N_{\max}}{\gamma}e^{-\zeta M}. \tag{22}$$

Thus, outside corrective-empty periods, the expected occupation time of mispricing above level M *decays exponentially in* M*. On the full interval,*

$$\mathbb{E}\left[\int_0^T \mathbf{1}_{\{|I_{t^-}|>M\}}\,\mathrm{d}t\right] \le \frac{N_{\max}}{\gamma}e^{-\zeta M} + \mathbb{E}[T_{\text{ecs}}], \tag{23}$$

where heuristically under competition and optimality $\mathbb{E}[T_{\text{ecs}}] \approx 0$.

This corollary means that large mispricing are expected to be short, with durations that decrease exponentially in the threshold except during temporary liquidity dry-ups when the corrective side of the book is absent then any bound requires some prior knowledge about the posting strategy of agents. Nevertheless, in any sensible market the time spent with one side of the book empty is negligible leading to an approximate bound on the total time mispricing spends being large.

The next statement controls the entire path, including corrective-empty periods.

Proposition 4 (Pathwise mispricing tail bound) *Under Assumptions* (S1)–(S7)*, let*

$$\Delta_P := 2\delta_{\max}(N_{\max}+N_{\text{steps}}), \qquad K_0 := |I_0|+\Delta_P = \left|S_0 - P_0^{\text{micro}}\right| + 2\delta_{\max}(N_{\max}+N_{\text{steps}}). \tag{24}$$

Then, for every $K > K_0$*,*

$$\mathbb{P}\left(\sup_{t\in[0,T]} |I_t| \le K\right) \ge 1 - 2\exp\left(-\frac{(K-K_0)^2}{2\sigma^2 T}\right). \tag{25}$$

In particular, $\mathbb{P}(\sup_{t\le T}|I_t| < \infty) = 1$.

This theorem means that the endogenous price process cannot generate unbounded dislocations over the simulation horizon: the worst deviation is controlled by finite LOB movements and by the volatility of the exogenous fundamental value. It allows to access the tail behaviour of the mispricing.

Remark 5 (Moments of the supremum) *As a standard result, the Gaussian tail bound* (25) *implies finiteness of all polynomial moments of* $\sup_{t\le T}|I_t|$*. In particular, for every* $p \ge 1$*,*

$$\mathbb{E}\left[\sup_{t\le T}|I_t|^p\right] \le K_0^p + p\int_{K_0}^{\infty} K^{p-1} 2\exp\left(-\frac{(K-K_0)^2}{2\sigma^2 T}\right) \mathrm{d}K < \infty.$$

Moreover, for every $\lambda < 1/(2\sigma^2 T)$*,*

$$\mathbb{E}\left[\exp\left(\lambda\big(\sup_{t\le T}|I_t| - K_0\big)^2\right)\right] < \infty.$$

These bounds formalize the stabilizing mechanism of the extended Hawkes model. Large mispricing increases the intensity of corrective flow exponentially, but corrective flow consumes finite depth. Consequently, periods with large active corrective pressure are short in compensator time, and even periods without an active corrective side remain pathwise controlled on finite horizons by the bounded number of LOB jumps together with the Brownian fluctuations of the fundamental value.

6 Learning Framework

We model market makers as learning agents trained with multi-agent proximal policy optimization (MAPPO). All experiments were implemented in Python using the RLlib

reinforcement-learning library (Liang et al. (2018)), while the simulator follows the PettingZoo Terry et al. (2021) `ParallelEnv` interface. In particular, the environment is implemented as a parallel multi-agent market and interfaced with RLlib through the PettingZoo adapter, which yields a standard multi-agent training loop.

The learning setup follows the centralized training with decentralized execution (CTDE) paradigm. During execution, each agent acts only on its local observation. During training, a centralized critic additionally receives a privileged global state and an agent identifier.

6.1 Observation space

At each decision step, agent i observes a local state composed of three blocks: (i) scalar features, (ii) a public summary of the limit order book, and (iii) a history window.

The scalar block contains the agent's own inventory and cash, time elapsed measure, market price, last transaction price and any private signal available to informed agents. The public LOB block summarizes liquidity around the mid-price while preserving queue-priority information: instead of collapsing each price level to a single depth number, the observation distinguishes between the agent's own resting liquidity and competing liquidity at tracked price levels. Finally, the history block provides a compact recent trajectory of market and agent-specific variables.

This decomposition was chosen to combine portfolio state, current market conditions, and short-term temporal context, while keeping the actor observation decentralized. The full construction of all observation components is provided in Appendix C.

6.2 Action space

At each step, the market maker chooses quote placement and quote size independently on the bid and ask sides. For each side, the policy outputs a discrete level from mid to determine the posted price and a discrete size bin.

The size bin determines whether a new order is submitted and, if so, maps deterministically to an explicit order volume. This discretization makes actions easier to interpret economically and reduces the number of behaviorally redundant choices. Order cancellations are not learned directly: they follow deterministic rules designed to stabilize training and to mimic realistic order-management behavior. The exact cancellation mechanism is described in Appendix B.

6.3 Policy network

The actor network combines three complementary encoders. First, scalar portfolio and market features are processed by a multilayer perceptron (MLP). Second, the agent's outstanding orders are fused with surrounding market liquidity through a cross-attention block, allowing the policy to condition on queue position and local competitive pressure. Third, a gated recurrent unit (GRU) encodes the history window to provide short-term temporal context. The resulting embeddings are concatenated and passed to the policy head.

To capture regime-dependent behavior, the actor uses a mixture-of-experts (MoE) head. Expert selection is based on the sign of the agent's inventory and, for agents with access to a private value signal, the sign of estimated mispricing $M_t - \widetilde{S}_t$. This design encourages specialization across economically meaningful regimes while keeping execution fully decentralized.

The precise architectural specification is deferred to Appendix D.

6.4 Centralized critic

The centralized critic provides the CTDE value baseline used in MAPPO. It receives a privileged global state containing market-wide variables, per-agent scalar states, and market-wide queue information, together with the identity of the agent whose value is being evaluated. The actor never accesses this privileged information.

Architecturally, the critic mirrors the main design principles of the actor. It combines an MLP-based scalar encoder, which uses a history embedding produced by the same GRU architecture as the actor, with a cross-attention module operating on the market-wide queue state, and maps the resulting representation to a value estimate. As in the actor, the value head uses regime-dependent experts. Here the gating depends on the sign of the current agent's inventory and the sign of the actual mispricing $M_t - S_t$, allowing the critic to distinguish between economically different market states during training.

A full mathematical description of the critic is provided in Appendix E.

6.5 Rewards

A central part of the environment specification is the objective of the market makers' training. In the literature, the objective is typically formulated as a terminal wealth criterion penalized by inventory risk; see for instance Avellaneda and Stoikov (2008); Bank et al. (2023); Barucci et al. (2025); Cartea and Jaimungal (2016); Lehalle and Neuman (2019); Neuman and Voß (2022).

In the environment, we decompose that reward into terminal and running components. This gives a denser learning signal to the RL algorithm while preserving the same economic trade-off between profitability and inventory control.

At the same time, we further enhance the reward by introducing liquidation prices computed from executable market quotes. This is a natural extension in a fully observable LOB. For brevity, denote $c_n^{(i)} := c_{t_n^{\mathrm{dec}}}^{(i)}$ and $q_n^{(i)} := q_{t_n^{\mathrm{dec}}}^{(i)}$.

For each market maker i, let B_n^{-i} and A_n^{-i} denote the best bid and best ask after excluding agent i's own resting orders. We then define the liquidation price by

$$P_{n,\mathrm{liq}}^{(i)} = \begin{cases} B_n^{-i}, & q_n^{(i)} > 0, \\ A_n^{-i}, & q_n^{(i)} < 0, \\ 0, & q_n^{(i)} = 0. \end{cases} \tag{26}$$

Accordingly, the liquidation wealth of agent i at decision time t_n^{dec} is

$$W_{n,\text{liq}}^{(i)} = c_n^{(i)} + q_n^{(i)} P_{n,\text{liq}}^{(i)}. \tag{27}$$

The one-step reward used by the simulator is the increment in liquidation wealth penalized by inventory holding. Writing

$$\Delta W_n^{(i)} := W_{n,\text{liq}}^{(i)} - W_{n-1,\text{liq}}^{(i)},$$

the implemented (unscaled) reward is

$$\tilde{r}_n^{(i)} = \begin{cases} \Delta W_n^{(i)} - \phi_i \big(q_n^{(i)}\big)^2 \Delta t, & n = 1, \dots, N_{\text{steps}} - 1, \\ \Delta W_{N_{\text{steps}}}^{(i)} - \alpha_i \big(q_{N_{\text{steps}}}^{(i)}\big)^2, & n = N_{\text{steps}}, \end{cases} \tag{28}$$

where N_{steps} is the number of agents' decision steps and (α_i, ϕ_i) are agent-specific inventory-aversion parameters. Hence, for all non-terminal decision steps the agent receives a running reward equal to the change in liquidation wealth minus the running inventory penalty, while at the terminal step the running penalty is replaced by a terminal quadratic penalty on the residual inventory.

Finally, to keep the numerical scale of rewards moderate during training, the reward passed to the learning algorithm is normalized by a constant factor. Denoting the unscaled reward in (28) by $\tilde{r}_n^{(i)}$, the simulator uses the scaled reward

$$r_n^{(i)} = \frac{1}{N_{\text{steps}}} \tilde{r}_n^{(i)}.$$

This scaling is introduced purely for numerical stability. This rewards then enters the computations of total reward in PPO algorithm.

7 Results

In this section, we analyze how market informedness affects the profitability of market makers. We begin with a stationary setting and then examine the robustness of the results by extending the duration of the experiment and introducing a non-stationary environment.

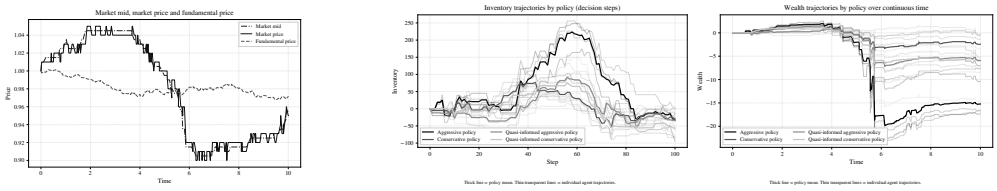
7.1 Baseline experiments

The baseline experiments show that adverse selection is particularly costly when market prices deviate substantially from fundamental value. Within the stationary informedness regime, the overall relationship is upward: market makers' profitability tends to improve as the level of market informedness increases.

Figure 1 presents the evaluation results for a single episode consisting of 100 simulation steps over the time horizon $T = 10$, under conditions of low market volatility and low market informedness. Subfigure 1a illustrates the evolution of the fundamental

value together with the market mid-price and transaction prices. During approximately the first half of the experiment, the market overprices the asset relative to its fundamental value. Around time $t = 4$, the market experiences a shock that drives the price back toward fundamentals, resulting in moderate losses distributed across all agents. Following this reversal, the sign of the mispricing changes, and, through the self-exciting dynamics of the Hawkes process, the market eventually begins to underprice the asset.

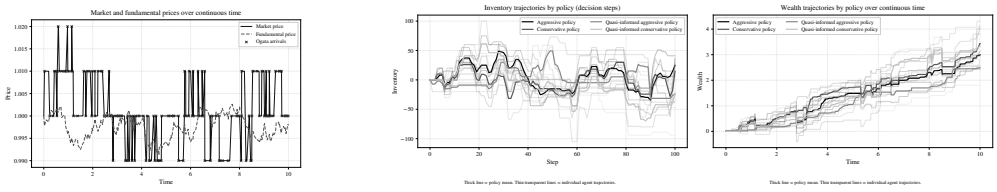
Subfigure 1b suggests that this price movement is driven primarily by the aggressive policy. Aggressive agents accumulate the largest inventory positions during the trend reversal and subsequently unwind them at less favorable prices during the period of market stagnation between $t = 6$ and $t = 8$. As a result, they incur substantial losses, which are shown in Subfigure 1c.



(a) Price evolution. (b) Inventory evolution. (c) Wealth evolution.

Fig. 1: Simulation under low volatility ($\sigma = 0.005$) and low informedness ($\psi = 0.01$). Seed = 42.

Figure 2 illustrates the results of a single episode conducted under the same setup, with the only difference being a higher level of market informedness, set to $\psi = 0.95$. Under this specification, market price dynamics differ markedly from those observed in the low-informedness case. As shown in Subfigure 2a, mispricing remains relatively limited, typically amounting to only 1–2 ticks and often falling below the tick size. Subfigure 2b shows that agents maintain relatively low inventory levels, reflecting the greater predictability of market movements. This more effective inventory management translates into a controlled and steady increase in terminal wealth across all policies.



(a) Price evolution. (b) Inventory evolution. (c) Wealth evolution.

Fig. 2: Simulation under low volatility ($\sigma = 0.005$) and high informedness ($\psi = 0.95$). Seed = 42.

These two experiments provide qualitative evidence that low market informedness is generally unfavorable for market makers, as it limits their ability to predict future price movements on the basis of past market states. In such an environment, the relatively small informed segment of the market generates adverse selection, which may trigger substantial reversals in market regimes and, consequently, losses for market makers who are unable to unwind their positions sufficiently quickly.

This interpretation is supported by Table 1, which reports the average terminal wealth for each policy and each informedness ratio ψ. For the terminal wealth associated with each policy, as well as for the aggregate terminal wealth corresponding to each value of ψ, an ordinary least squares (OLS) regression has been estimated in order to identify the general trend. In all cases, the estimated slope coefficient is positive, which is consistent with the conclusions drawn from the single-episode analysis. Repeating the same experiments under higher market volatility, $\sigma = 0.02$, leads to the same qualitative conclusion: market makers' profitability exhibits an overall upward trend as a function of market informedness.

Policy	$\psi = 0,01$	$\psi = 0,05$	$\psi = 0,10$	$\psi = 0,30$	$\psi = 0,50$	$\psi = 0,70$	$\psi = 0,90$	$\psi = 0,95$	$\psi = 0,99$	$\psi = 1,00$	$\hat{\alpha}$	$\hat{\beta}$	R^2
Low volatility ($\sigma = 0.005$)													
Aggressive	-0,73 ± 5,95	1,68 ± 2,57	1,56 ± 2,10	2,88 ± 1,54	4,36 ± 1,60	2,90 ± 1,11	4,62 ± 1,12	3,61 ± 0,98	3,38 ± 0,97	3,93 ± 1,23	1,12	3,09	0,62
Conservative	0,96 ± 3,40	1,25 ± 2,58	2,44 ± 1,27	3,05 ± 1,32	1,96 ± 0,83	2,56 ± 1,12	3,06 ± 0,77	3,34 ± 0,84	3,40 ± 0,90	3,09 ± 0,97	1,57	1,71	0,65
Quasi-informed aggressive	-0,50 ± 3,24	0,85 ± 3,40	3,04 ± 1,62	2,58 ± 1,01	3,43 ± 1,27	3,53 ± 0,89	4,70 ± 1,39	3,58 ± 0,97	4,11 ± 1,00	3,20 ± 1,04	1,21	2,99	0,62
Quasi-informed conservative	-0,91 ± 3,16	0,71 ± 1,79	2,88 ± 1,78	2,55 ± 1,38	2,91 ± 1,06	3,40 ± 1,34	2,54 ± 0,69	2,80 ± 0,84	3,14 ± 0,90	2,24 ± 0,75	1,15	1,96	0,37
Sum	**-1,18**	**4,49**	**9,92**	**11,06**	**12,66**	**12,39**	**14,92**	**13,33**	**14,03**	**12,46**	**5,05**	**9,75**	**0,64**
High volatility ($\sigma = 0.02$)													
Aggressive	-3,12 ± 6,14	-1,47 ± 2,76	0,60 ± 3,11	-0,08 ± 3,99	0,19 ± 3,22	-0,07 ± 3,23	0,51 ± 3,27	2,36 ± 2,93	0,44 ± 4,43	2,73 ± 5,01	-1,50	3,11	0,58
Conservative	-2,13 ± 6,20	-0,26 ± 2,96	0,65 ± 2,26	0,66 ± 1,66	0,09 ± 2,18	0,68 ± 2,26	0,68 ± 2,96	2,36 ± 2,05	1,04 ± 2,20	1,89 ± 2,75	-0,68	2,26	0,57
Quasi-informed aggressive	-0,26 ± 4,41	2,67 ± 3,59	6,33 ± 3,55	4,27 ± 2,39	2,11 ± 2,79	2,58 ± 2,46	2,39 ± 2,30	3,67 ± 2,66	1,77 ± 3,16	4,54 ± 2,13	2,90	0,20	0,00
Quasi-informed conservative	-2,44 ± 4,61	0,11 ± 4,68	-0,82 ± 4,18	1,62 ± 3,25	2,54 ± 2,31	-0,48 ± 2,42	0,41 ± 2,34	3,08 ± 1,98	-0,56 ± 2,37	3,66 ± 2,43	-0,68	2,52	0,28
Sum	**-7,95**	**1,05**	**6,76**	**6,47**	**4,93**	**2,71**	**3,99**	**11,47**	**2,69**	**12,82**	**0,05**	**8,09**	**0,33**

Table 1: Final wealth distribution across informed ratios ψ under low volatility ($\sigma = 0.005$) and high volatility ($\sigma = 0.02$). Policy-level entries are mean ± standard deviation (STD), rounded to two decimals. The **Sum** row reports the sum of policy means. The last three columns report $(\hat{\alpha}, \hat{\beta}, R^2)$ from the OLS regression of mean final wealth on the informed ratio ψ. Training for 50 iterations evaluated on 400 distinct episodes.

7.2 Quoting

Quoting behavior differs systematically across policies, but it does not display a clear monotonic relationship with market informedness. The evidence instead suggests that profitability is driven primarily by overall market conditions, even though agents learn to adapt their quoting decisions to inventory positions and mispricing regimes.

Another important aspect of the baseline analysis concerns quoting behavior and its dependence on policy type and market informedness.

Averaged across all informedness ratios ψ, posted spreads differ across policies in an economically intuitive way. In the low-volatility regime, $\sigma = 0.005$, the Aggressive policy posts the tightest quotes on average, with a mean spread of 2.41 ticks, whereas the Conservative policy posts an average spread of 2.64 ticks, approximately 9.5% higher. The quasi-informed policies quote more widely still, with average spreads of 2.79 and 2.80 ticks for the aggressive and conservative variants, respectively. This cross-policy ordering is consistent with the intended distinction between more and less aggressive quoting styles.

At the same time, the relationship between the average posted spread and the informedness ratio is weak and irregular. Regressing average spread on ψ yields very low explanatory power for all policies, with $R^2 < 0.05$ throughout. This suggests that the previously documented increase in terminal wealth with market informedness is not driven by a simple monotonic adjustment in average quoting width. Rather, the profitability effect appears to stem primarily from changes in the market environment itself, such as the reduction in persistent mispricing and adverse selection at higher values of ψ.

A similar conclusion follows from the analysis of posted volume across spread levels. Across policies, agents generally post larger volumes closer to the mid-price and smaller volumes further away. The decline is moderate: mean posted volume typically decreases by about one unit between the closest and farthest quoted levels, that is, from slightly above 3 to slightly above 2. Aggressive policies also tend to post somewhat larger volumes near the mid-price than their conservative counterparts.

However, as in the case of average spread, posted volume at a given level from mid generally exhibits no strong linear relation with ψ, and in most cases the regression fit remains weak, with $R^2 < 0.10$. Two exceptions are the Quasi-informed aggressive policy at level 1, for which $R^2 = 0.484$ with a positive slope, and the Conservative policy at level 0, for which $R^2 = 0.30$ with a negative slope. Aside from these isolated cases, no common economically meaningful pattern is visible across policies and levels. Taken together, these findings support the view that the increase in profitability associated with higher market informedness is driven primarily by improvements in the market environment rather than by systematic changes in average quoting width or posted size.

At first glance, these results might suggest that the policies do not learn meaningful behavior. However, an analysis conditional on inventory and mispricing regimes indicates otherwise. Averaged across all informedness ratios ψ, and for both low and high market volatility, all policies tend to post tighter quotes and larger volumes on the ask side when inventory is positive, and analogously on the bid side when inventory is negative. Quantitatively, the relevant half-spreads are typically narrower by a

factor of approximately 1.6–3.2, while the corresponding posted volumes are larger by about 35%–120%, with an average increase of roughly 70%. This pattern is consistent with meaningful inventory-aware specialization.

Moreover, the uninformed policies remain nearly symmetric across underpricing and overpricing regimes once inventory is held fixed, which is expected given that they do not observe the fundamental value. By contrast, the quasi-informed policies do differentiate their behavior across mispricing regimes. This distinction is especially clear for the Quasi-informed aggressive policy and less regular for the Quasi-informed conservative one. In particular, both policies tend to post higher ask-side volumes in overpricing regimes and higher bid-side volumes in underpricing regimes, suggesting that informed agents attempt to unwind positions when prices are favorable. When this pattern is analyzed separately for each value of ψ, it remains persistent for the Quasi-informed aggressive policy, but becomes more ambiguous for its conservative counterpart. This suggests that the Quasi-informed conservative policy faces a stronger trade-off between exploiting informational advantages and controlling inventory risk.

No equally clear pattern emerges for bid and ask half-spreads. This is likely due to the role of stochastic market order arrivals and random queueing in determining the effective importance of quoted spreads. Because market orders arriving between decision steps may consume liquidity beyond the first level, posted volume appears to be more consequential than the precise choice of quoting level, making the effect of half-spread adjustments less distinct.

To further assess the robustness of these findings and verify that the learned policies are not merely overfitting the specific market structure considered so far, we extend the analysis in two directions. First, we evaluate the already trained policies in substantially longer episodes than those used during training. Second, we introduce non-stationarity into the market regime in order to examine whether the main conclusions remain valid in a more dynamic environment.

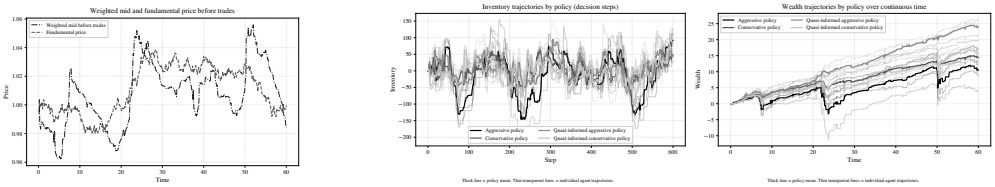
7.3 Expanded experiment

An important robustness question is whether policies trained on episodes of 100 decision steps retain their performance in substantially longer runs. The expanded experiment addresses this issue by evaluating the same policies in episodes of 600 decision steps.

The results of the extended experiment are presented in Figure 3. In this specification, market informedness is set to $\psi = 0.10$. In such an environment, mispricings occur frequently and often reach a magnitude of several ticks. The resulting trend reversals are qualitatively similar to those observed in the shorter experiments conducted in less informed markets.

The analysis of inventory dynamics shows that inventories, particularly for the Aggressive policy, can become large in absolute value. However, over a longer horizon they continue to fluctuate around zero rather than drift persistently in one direction. The wealth trajectories display a generally positive trend over time, interrupted by occasional sharp declines, again most visibly for the Aggressive policy. Overall, the behavior of the policies appears stable throughout the full episode. In particular, there is no visible change in behavior after time $t = 10$ (or step 100), which corresponds to

the original time horizon used during training. This indicates that the learned policies generalize well to longer experiments. Apart from scale, no clear quantitative change is visible in the effect of market informedness on market makers' profitability in this longer-horizon setting.



(a) Weighted market mid and fundamental value.

(b) Inventory evolution.

(c) Wealth evolution.

Fig. 3: Simulation under low volatility ($\sigma = 0.005$) and medium-low informedness ($\psi = 0.10$). Seed = 37.

Quantitatively, the results remain similar to those of the baseline experiment.

7.4 Non-stationary experiments

The non-stationary experiments indicate that the positive relationship between market informedness and agents' profitability is not merely an artifact of fitting a stationary environment. Instead, it remains visible across a broad range of non-stationary market regimes.

While the expanded-horizon experiment shows that the learned policies generalize over time, an additional question is whether they also remain effective when the underlying market environment itself changes. To address this issue, we consider a range of non-stationary regime specifications for the informed ratio ψ and market volatility σ, covering both cases in which agents are aware of the regime and cases in which they are not. The results are reported in Table 2, which presents the mean per-step wealth change for each policy and informedness level, together with OLS estimates summarizing the relationship between profitability and ψ under each regime specification.

Overall, the evidence is consistent with the main conclusion of the stationary analysis. For every regime specification, the aggregate measure of profitability, reported in the **Sum** row, exhibits a positive slope coefficient with respect to market informedness. This indicates that the previously documented upward trend in market makers' wealth is not driven solely by the stationary benchmark, but persists even when the market environment changes over time. In several cases, this relationship is also quantitatively strong, as reflected in relatively high values of R^2, particularly for regime models 4, 5, and 6. This suggests that the positive association between informedness and profitability remains stable across a broad set of structured regime-switching environments.

At the policy level, the same pattern largely continues to hold. The Aggressive, Quasi-informed aggressive, and Quasi-informed conservative policies display positive slope coefficients in all reported regime specifications. The only systematic exception is the Conservative policy, for which the estimated slope becomes negative in regime models 5 and 7. This suggests that the most inventory-protective policy may be less capable of benefiting from higher market informedness in environments with more frequent or irregular regime changes. By contrast, the quasi-informed policies remain positively related to ψ throughout, which is consistent with the broader interpretation of this study: access to informational signals becomes increasingly valuable as market informedness rises, even when the surrounding market structure is non-stationary.

At the same time, the table also shows that the strength of this relationship depends on the type of non-stationarity considered. In particular, the most irregular specifications, especially regime model 7 in the unknown setting, produce a much weaker fit. Thus, non-stationarity does not eliminate the positive profitability effect of informedness, but it does make the relationship noisier and less uniform, possibly due to limitations of machine learning. Taken together, these results support the view that the main finding of the paper is robust: higher market informedness is generally associated with better market-making performance, and this conclusion survives both temporal extension of the experiment and substantial departures from stationarity.

Regime mode	Policy	ψ (informed ratio)															OLS on ψ		
		0.00	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	1.00	$\hat{\beta}$	$\hat{\alpha}$	R^2
	agg	-0.64	-0.19	-0.04	0.04	0.15	0.03	0.26	0.26	0.24	0.30	0.29	0.34	0.31	0.26	0.33	0.54	-0.14	0.59
	cons	-0.01	-0.05	0.03	0.00	0.07	0.03	0.14	0.13	0.13	0.13	0.09	0.12	0.16	0.09	0.17	0.15	0.01	0.67
`Regime model 2` (known)	`quasi-inf-agg`	-0.67	0.02	0.14	0.19	0.28	0.22	0.33	0.35	0.30	0.31	0.35	0.34	0.36	0.32	0.36	0.44	-0.00	0.39
	`quasi-inf-cons`	-1.36	0.02	0.16	0.23	0.33	0.17	0.34	0.33	0.30	0.39	0.36	0.38	0.37	0.28	0.39	0.61	-0.12	0.27
	Sum	**-2.67**	**-0.19**	**0.29**	**0.47**	**0.84**	**0.44**	**1.07**	**1.06**	**0.97**	**1.33**	**1.09**	**1.18**	**1.20**	**0.95**	**1.24**	**1.74**	**-0.25**	**0.40**
	agg	-0.52	-0.14	-0.03	-0.02	0.10	-0.05	0.01	0.12	0.09	0.23	0.18	0.25	0.21	0.35	0.25	0.47	-0.17	0.69
	cons	-0.20	0.16	0.13	0.24	0.18	0.07	0.13	0.26	0.22	0.26	0.20	0.23	0.25	0.23	0.27	0.19	0.08	0.35
`Regime model 2` (unknown)	`quasi-inf-agg`	-0.90	0.10	0.23	0.30	0.40	0.22	0.27	0.40	0.38	0.39	0.31	0.43	0.37	0.41	0.42	0.47	0.01	0.29
	`quasi-inf-cons`	-0.49	0.14	0.25	0.23	0.30	0.16	0.23	0.31	0.30	0.32	0.26	0.30	0.30	0.29	0.30	0.27	0.07	0.26
	Sum	**-2.11**	**0.26**	**0.58**	**0.76**	**0.97**	**0.39**	**0.63**	**1.09**	**0.99**	**1.20**	**0.95**	**1.21**	**1.13**	**1.29**	**1.24**	**1.40**	**-0.00**	**0.39**
	agg	-0.38	-0.16	-0.11	-0.12	0.02	0.04	0.05	0.00	0.08	0.07	0.07	0.17	0.18	0.19	0.12	0.29	-0.13	0.65
	cons	-0.03	-0.10	-0.06	-0.10	-0.01	0.02	0.02	0.08	0.09	0.03	-0.02	0.04	0.01	0.09	0.08	0.12	-0.05	0.58
`Regime model 4` (known)	`quasi-inf-agg`	-0.85	-0.05	0.06	0.16	0.19	0.16	0.22	0.26	0.24	0.27	0.24	0.32	0.30	0.32	0.31	0.45	-0.10	0.64
	`quasi-inf-cons`	-0.74	-0.00	0.07	0.15	0.15	0.11	0.16	0.20	0.22	0.19	0.12	0.30	0.27	0.29	0.22	0.41	-0.10	0.62
	Sum	**-2.00**	**-0.31**	**-0.04**	**0.09**	**0.35**	**0.33**	**0.45**	**0.55**	**0.63**	**0.56**	**0.41**	**0.82**	**0.76**	**0.89**	**0.73**	**1.27**	**-0.38**	**0.70**
	agg	-0.34	-0.11	-0.10	-0.06	0.04	0.05	0.04	0.06	0.12	0.08	0.04	0.13	0.17	0.14	0.14	0.25	-0.12	0.70
	cons	0.02	-0.02	0.06	0.02	0.10	0.04	0.06	0.09	0.11	0.08	0.04	0.10	0.17	0.12	0.15	0.14	-0.01	0.74
`Regime model 4` (unknown)	`quasi-inf-agg`	-0.42	0.04	0.08	0.14	0.19	0.14	0.18	0.25	0.25	0.22	0.15	0.26	0.31	0.30	0.30	0.34	-0.02	0.79
	`quasi-inf-cons`	-0.43	0.06	0.10	0.16	0.21	0.15	0.16	0.25	0.25	0.23	0.18	0.26	0.30	0.28	0.30	0.33	-0.01	0.77
	Sum	**-1.17**	**-0.03**	**0.13**	**0.26**	**0.54**	**0.37**	**0.44**	**0.65**	**0.73**	**0.61**	**0.41**	**0.75**	**0.96**	**0.84**	**0.89**	**1.05**	**-0.15**	**0.79**
	agg	-0.38	-0.32	-0.09	-0.02	0.03	0.02	0.09	0.09	0.13	0.11	0.06	0.02	0.17	0.14	0.03	0.28	-0.06	0.49
	cons	0.17	0.10	0.09	0.09	0.08	0.08	0.05	0.06	0.03	0.04	-0.03	-0.05	-0.06	0.00	-0.00	-0.12	0.12	0.57
`Regime model 5` (known)	`quasi-inf-agg`	-0.20	-0.04	0.08	0.12	0.14	0.14	0.20	0.21	0.27	0.32	0.25	0.23	0.27	0.28	0.20	0.34	0.02	0.84
	`quasi-inf-cons`	-0.15	-0.03	0.08	0.11	0.17	0.16	0.21	0.20	0.23	0.21	0.19	0.09	0.06	0.13	0.08	0.26	0.05	0.74
	Sum	**-0.55**	**-0.29**	**0.16**	**0.30**	**0.42**	**0.40**	**0.56**	**0.57**	**0.67**	**0.69**	**0.46**	**0.29**	**0.44**	**0.55**	**0.31**	**0.76**	**0.13**	**0.73**
	agg	-0.32	-0.22	-0.09	-0.05	0.01	0.05	0.08	0.06	0.10	0.13	0.12	0.06	0.12	0.13	0.06	0.25	-0.11	0.78
	cons	0.14	0.04	0.06	0.03	0.04	0.05	0.04	0.01	0.06	0.05	-0.02	-0.06	-0.05	0.02	-0.01	-0.10	0.09	0.65
`Regime model 5` (unknown)	`quasi-inf-agg`	-0.29	0.01	0.06	0.09	0.12	0.15	0.21	0.18	0.24	0.28	0.25	0.18	0.22	0.23	0.19	0.33	0.00	0.88
	`quasi-inf-cons`	-0.20	-0.01	0.06	0.08	0.11	0.13	0.18	0.16	0.20	0.23	0.21	0.15	0.15	0.17	0.11	0.28	0.01	0.88
	Sum	**-0.67**	**-0.18**	**0.09**	**0.15**	**0.28**	**0.37**	**0.52**	**0.41**	**0.60**	**0.69**	**0.56**	**0.33**	**0.45**	**0.55**	**0.35**	**0.75**	**-0.01**	**0.92**
	agg	-0.31	-0.17	-0.11	-0.10	0.03	0.06	0.04	0.05	0.11	0.11	0.10	0.17	0.19	0.19	0.15	0.29	-0.11	0.76
	cons	-0.04	-0.08	-0.04	-0.05	0.06	0.04	0.02	0.07	0.11	0.06	0.02	0.10	0.12	0.13	0.12	0.21	-0.08	0.79
`Regime model 6` (known)	`quasi-inf-agg`	-0.70	-0.04	0.07	0.13	0.16	0.16	0.22	0.28	0.24	0.30	0.26	0.33	0.33	0.34	0.34	0.46	-0.09	0.75
	`quasi-inf-cons`	-0.66	-0.02	0.08	0.14	0.19	0.17	0.21	0.28	0.27	0.28	0.20	0.34	0.33	0.33	0.30	0.44	-0.09	0.73
	Sum	**-1.71**	**-0.31**	**0.00**	**0.12**	**0.44**	**0.43**	**0.49**	**0.67**	**0.73**	**0.75**	**0.58**	**0.95**	**0.98**	**0.99**	**0.91**	**1.40**	**-0.36**	**0.82**
	agg	-0.29	-0.10	-0.09	-0.05	0.04	0.05	0.04	0.05	0.08	0.09	0.06	0.12	0.15	0.15	0.12	0.22	-0.12	0.83
	cons	0.00	-0.01	0.05	0.03	0.10	0.06	0.05	0.08	0.11	0.08	0.06	0.10	0.17	0.15	0.15	0.16	-0.00	0.86
`Regime model 6` (unknown)	`quasi-inf-agg`	-0.41	0.03	0.08	0.12	0.18	0.16	0.18	0.24	0.23	0.22	0.14	0.26	0.30	0.30	0.30	0.34	-0.03	0.85
	`quasi-inf-cons`	-0.41	0.05	0.10	0.14	0.20	0.16	0.17	0.24	0.25	0.23	0.18	0.25	0.29	0.28	0.28	0.33	-0.02	0.82
	Sum	**-1.11**	**-0.03**	**0.13**	**0.24**	**0.52**	**0.43**	**0.44**	**0.61**	**0.67**	**0.62**	**0.44**	**0.74**	**0.91**	**0.87**	**0.85**	**1.04**	**-0.17**	**0.87**
	agg	-0.40	-0.29	-0.04	0.02	0.08	0.08	0.13	0.21	0.25	0.13	-0.03	-0.12	0.09	0.30	0.10	0.26	-0.10	0.28
	cons	0.14	-0.00	0.03	-0.09	-0.02	0.04	-0.04	0.02	0.05	0.11	-0.18	-0.15	-0.12	-0.16	0.06	-0.17	0.07	0.40
`Regime model 7` (known)	`quasi-inf-agg`	-0.39	-0.07	0.04	0.27	0.25	0.22	0.28	0.41	0.51	0.27	0.24	0.35	0.28	1.08	0.43	0.88	-0.09	0.53
	`quasi-inf-cons`	-0.19	-0.12	0.05	0.04	0.04	0.10	0.04	0.03	0.17	0.06	-0.07	-0.10	-0.03	-0.05	0.05	0.13	-0.03	0.21
	Sum	**-0.85**	**-0.48**	**0.09**	**0.24**	**0.35**	**0.44**	**0.42**	**0.67**	**0.97**	**0.58**	**-0.03**	**-0.01**	**0.21**	**1.17**	**0.63**	**1.10**	**-0.15**	**0.34**
	agg	-0.26	-0.07	-0.09	-0.04	0.01	0.10	0.06	-0.21	0.07	-0.12	-0.31	0.12	-0.01	-0.13	0.04	0.05	-0.08	0.02
	cons	-0.01	0.06	0.10	0.13	0.09	0.07	0.04	-0.12	0.08	-0.00	-0.30	0.01	0.04	-0.13	0.05	-0.13	0.07	0.20
`Regime model 7` (unknown)	`quasi-inf-agg`	-0.10	0.04	0.12	0.18	0.14	0.19	0.22	0.17	0.30	0.22	0.06	0.22	0.33	0.18	0.28	0.18	0.08	0.39
	`quasi-inf-cons`	-0.24	0.06	0.06	0.10	0.12	0.14	0.14	0.12	0.24	0.06	-0.09	0.21	0.21	0.07	0.11	0.11	0.03	0.12
	Sum	**-0.62**	**0.08**	**0.18**	**0.36**	**0.37**	**0.50**	**0.46**	**-0.04**	**0.70**	**0.15**	**-0.64**	**0.55**	**0.57**	**-0.01**	**0.48**	**0.20**	**0.10**	**0.04**

Table 2: Mean per-step wealth change by regime model, policy, and informed ratio ψ. The **Sum** row reports the sum across the four policies within each regime specification. The final three columns give OLS estimates from regressing mean per-step wealth change on the informed ratio ψ. All values are rounded to two decimals. Here `agg`, `cons`, `quasi-inf-agg`, and `quasi-inf-cons` denote, respectively, the Aggressive, Conservative, Quasi-informed aggressive, and Quasi-informed conservative strategies.

Regime mapping (table order):

Regime model 1	`One change at a fixed time to a fixed regime`
Regime model 2	`One change at a fixed time to a random regime`
Regime model 3	`One change at a random time to a fixed regime`
Regime model 4	`One change at a random time to a random regime`
Regime model 5	`Fixed number of changes on a fixed grid to random regimes`
Regime model 6	`Fixed number of changes on a random grid to random regimes`
Regime model 7	`Exponential waiting times for random regimes`

8 Conclusions

The main result of this paper is that market-making profitability is, on average, increasing in market informedness. In the simulated environment, higher values of

the informed ratio ψ are generally associated with higher terminal wealth, with the strongest improvement occurring when moving away from the least informed market regimes. At the same time, the relationship is not strictly monotonic: at high levels of informedness it becomes flatter and occasionally locally non-monotonic. The non-stationary experiments show that this pattern is not an artifact of the stationary baseline. The positive association remains visible across a broad class of regime-switching specifications, although it becomes weaker and less regular in the most irregular environments.

At first sight, this conclusion appears to differ from the findings of Barucci et al. (2025), where a larger share of informed traders worsens market-maker performance. However, the difference is naturally explained by the underlying market structure. In Barucci et al. (2025), the market maker quotes around an exogenous reference price under simpler order-flow assumptions. In the present setting, by contrast, prices are formed endogenously through the interaction of market and limit orders, order flow is self-exciting, and trade sizes are heterogeneous. As a result, informed trading has two opposing effects: it increases adverse-selection risk, but it also improves price discovery by reducing the persistence of mispricing between transaction prices and fundamentals. The results suggest that, outside the least informed environments, the price-discovery channel can dominate the toxic-flow channel.

More broadly, the paper shows that in richer order-book environments, informed trading need not reduce profitability; instead, it may improve market makers' performance by limiting prolonged deviations from fundamental value.

8.1 Limitations and future work

The analysis remains subject to several limitations that also point to natural directions for future research. First, although the model captures endogenous price formation, self-exciting order flow, and strategic competition, it is not yet tightly calibrated to high-frequency order-book data. A closer empirical calibration of queue dynamics, latency, participant heterogeneity, and other market frictions would help assess the external validity of the results more precisely. Second, the informational environment remains stylized, relying on a binary informed–uninformed decomposition. Extending the model to richer information structures and allowing agents to infer market regimes from partially observed signals would bring the setup closer to actual trading environments. Finally, an important theoretical extension would be to characterize more sharply the threshold at which the price-discovery effect of informed trading ceases to dominate the toxic-flow effect. Such an analysis could connect the reinforcement-learning evidence developed here more tightly to the comparative statics of the analytical market-microstructure literature.

9 Declarations

9.1 Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

9.2 Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

9.3 Author Contributions

Konrad Ochedzan contributed to the study conception and design, methodology, software implementation, numerical experiments, formal analysis, visualization, and writing of the original draft. Nino Antulov-Fantulin contributed to the study conception and design, methodology, and supervision. Both authors read and approved the final manuscript.

9.4 Code Availability

The code used to generate the simulation results and figures is available at Github.

9.5 Gen AI Use

Generative Artificial Intelligence, namely ChatGPT from OpenAI, has been used as an editorial and programming aid for selected code and text revisions. The research design, mathematical formulation, implementation decisions, validation of outputs, and final responsibility for the code and manuscript content remain with Konrad Ochedzan. ChatGPT is not an author or contributor in the scholarly-authorship sense, and its use should be understood as assisted editing and development support under human review.

Appendix A Proofs for finite-horizon stability

This appendix proves the stability statements in Section 5. Throughout, $N([0,T]) = \sum_{i=1}^{C} N_i([0,T])$, $C = 2K$, and $\lambda_i(t) = G_i(x(t^-))\lambda_i^H(t)$ denotes the predictable effective intensity. The convention $I_{t^-} = S_{t^-} - P_{t^-}^{\text{micro}}$ is used at possible jump times of the book. Since the fundamental process is continuous in the model, $S_{t^-} = S_t$.

A.1 Corrective and destabilizing events

At any time t, define the predictable sets of corrective and destabilizing types by

$$\text{corr}(t) := \begin{cases} \{1,\dots,K\}, & \text{if } I_{t^-} \geq 0, \\ \{K+1,\dots,2K\}, & \text{if } I_{t^-} < 0, \end{cases} \qquad \text{dest}(t) := \{1,\dots,2K\} \setminus \text{corr}(t). \quad \text{(A1)}$$

Thus, if the microprice is below the fundamental value, buys are corrective; if the microprice is above the fundamental value, sells are corrective. The convention at $I_{t^-} = 0$ is immaterial for the bounds.

Lemma 6 (Destabilizing gain is at most one) *Under Assumptions* (S1)–(S6), $G_i(x(t^-)) \leq 1$ *for every* $t \in [0,T]$ *and every* $i \in \text{dest}(t)$. *Economically, order flow that would push prices*

farther away from fundamentals is not amplified by the informedness mechanism; it remains at its baseline Hawkes level or is dampened.

Proof Suppose first that $I_{t^-} > 0$. Then sells are destabilizing and

$$G^b(x(t^-)) = \psi e^{-\eta\delta^b_{t^-} - \zeta I_{t^-}} + (1-\psi)e^{-\eta\delta^b_{t^-}}.$$

Both exponents are non-positive because $\eta, \zeta > 0$, $\delta^b_{t^-} \geq 0$, and $I_{t^-} > 0$. Hence $G^b(x(t^-)) \leq 1$. The case $I_{t^-} < 0$ is symmetric. When $I_{t^-} = 0$, the destabilizing side under the convention in (A1) has gain $e^{-\eta\delta^b_{t^-}} \leq 1$. □

Lemma 7 (Corrective intensity lower bound) *Let γ be defined by (19). At every time t such that the corrective side of the book is non-empty at t^-,*

$$\sum_{i \in \mathrm{corr}(t)} \lambda_i(t) \geq \gamma e^{\zeta |I_{t^-}|}. \tag{A2}$$

Economically, this lower bound is the model's correction channel: whenever liquidity is available on the side that would move prices toward fundamentals, informed traders become more aggressive as the asset becomes more underpriced or overpriced.

Proof Suppose $I_{t^-} > 0$ and the ask side is non-empty. Then corrective types are buys. For every $i \leq K$,

$$\begin{aligned}
\lambda_i(t) &= G^a(x(t^-))\lambda^H_i(t) \\
&\geq \psi e^{-\eta\delta^a_{t^-} + \zeta I_{t^-}} \lambda^H_i(t) \\
&\geq \psi e^{-\eta\delta_{\max}} \mu_i e^{\zeta I_{t^-}} \geq \gamma e^{\zeta I_{t^-}},
\end{aligned}$$

where we used $\delta^a_{t^-} \leq \delta_{\max}$ and $\lambda^H_i(t) \geq \mu_i$. Summing over the corrective types gives (A2). The case $I_{t^-} < 0$ is symmetric, with sells as the corrective types and $|I_{t^-}| = -I_{t^-}$. At $I_{t^-} = 0$, the same argument gives the lower bound γ. □

A.2 Event-count bounds

Proof of Proposition 1 Fix an inter-decision interval $[t^{\mathrm{dec}}_n, t^{\mathrm{dec}}_{n+1})$. By Assumptions (S1)–(S6), the total resting depth on each side at t^{dec}_n is at most D_{tot}, and no new limit orders are posted before t^{dec}_{n+1}. Each market buy consumes at least $q_{\min}$ units from the ask side, and each market sell consumes at least $q_{\min}$ units from the bid side. Once a side is depleted, the corresponding gain is zero, so no further events of that type can execute before the next decision time. Therefore,

$$N([t^{\mathrm{dec}}_n, t^{\mathrm{dec}}_{n+1})) \leq \frac{D_{\mathrm{tot}}}{q_{\min}} + \frac{D_{\mathrm{tot}}}{q_{\min}} = \frac{2D_{\mathrm{tot}}}{q_{\min}}.$$

Summing over the N_{steps} inter-decision intervals yields $N([0,T]) \leq 2D_{\mathrm{tot}}N_{\mathrm{steps}}/q_{\min} = N_{\max}$ almost surely. Taking expectations gives the final claim. □

The next auxiliary proposition is not needed for the pathwise non-explosion bound, but it records how the destabilizing part of the process is controlled by the Hawkes branching structure. Let A be the integrated kernel matrix in (13) and $\|A\|_1 := \max_{1 \leq j \leq C} \sum_{i=1}^{C} A_{ij}$.

Proposition 8 (Auxiliary moment bound) *If, in addition to Assumptions* (S1)–(S6), $\|A\|_1 < 1$, *then*

$$\mathbb{E}[N([0,T])] \le \frac{N_{\max} + \|\boldsymbol{\mu}\|_1 T}{1 - \|A\|_1}, \tag{A3}$$

where $\|\boldsymbol{\mu}\|_1 = \sum_{i=1}^{C} \mu_i$. *If* $t_{n+1}^{\mathrm{dec}} - t_n^{\mathrm{dec}} \ge \Delta_{\min}^{\mathrm{dec}} > 0$, *then*

$$\mathbb{E}[N([0,T])] \le \frac{2D_{\mathrm{tot}}/(q_{\min}\Delta_{\min}^{\mathrm{dec}}) + \|\boldsymbol{\mu}\|_1}{1 - \|A\|_1}\, T. \tag{A4}$$

Economically, this bound allows the model to generate clustered high-frequency order flow without making total trading volume explode, which is essential for comparing market-maker profitability across informedness regimes.

Proof Let N^{corr} and N^{dest} count corrective and destabilizing events according to (A1). Corrective events consume finite depth, so the same depth argument used above gives

$$N^{\mathrm{corr}}([0,T]) \le N_{\max} \qquad \text{a.s.} \tag{A5}$$

By Lemma 6, destabilizing gains are bounded by one. Hence, using the compensator identity and non-negativity of the raw Hawkes intensities,

$$\mathbb{E}[N^{\mathrm{dest}}([0,T])] \le \mathbb{E}\left[\int_0^T \sum_{i=1}^{C} \lambda_i^H(t)\,\mathrm{d}t\right]. \tag{A6}$$

For the raw Hawkes process,

$$\int_0^T \sum_{i=1}^{C} \lambda_i^H(t)\,\mathrm{d}t = \|\boldsymbol{\mu}\|_1 T + \sum_{i,j} \int_0^T \int_0^{t^-} \phi_{i\leftarrow j}(t-s)\,\mathrm{d}N_j(s)\,\mathrm{d}t.$$

Fubini–Tonelli and the definition of A_{ij} give

$$\int_0^T \int_0^{t^-} \phi_{i\leftarrow j}(t-s)\,\mathrm{d}N_j(s)\,\mathrm{d}t \le A_{ij} N_j([0,T]).$$

Therefore,

$$\mathbb{E}\left[\int_0^T \sum_{i=1}^{C} \lambda_i^H(t)\,\mathrm{d}t\right] \le \|\boldsymbol{\mu}\|_1 T + \sum_{i,j} A_{ij}\mathbb{E}[N_j([0,T])] \le \|\boldsymbol{\mu}\|_1 T + \|A\|_1 \mathbb{E}[N([0,T])].$$

Combining this inequality with (A5) and (A6) yields

$$\mathbb{E}[N([0,T])] \le N_{\max} + \|\boldsymbol{\mu}\|_1 T + \|A\|_1 \mathbb{E}[N([0,T])].$$

Rearranging gives (A3). The linear bound (A4) follows from $N_{\mathrm{steps}} \le T/\Delta_{\min}^{\mathrm{dec}}$. □

A.3 Exponential integrability and occupation time

Proof of Proposition 2 Define the aggregate corrective intensity

$$\Lambda^{\mathrm{corr}}(t) := \sum_{i \in \mathrm{corr}(t)} \lambda_i(t).$$

It is predictable, and by the compensator identity,

$$\mathbb{E}\left[\int_0^T \Lambda^{\mathrm{corr}}(t)\,\mathrm{d}t\right] = \mathbb{E}[N^{\mathrm{corr}}([0,T])] \le N_{\max},$$

where the final inequality is (A5). On the set where the corrective side is non-empty, Lemma 7 gives $\Lambda^{\mathrm{corr}}(t) \geq \gamma e^{\zeta|I_{t^-}|}$. When the corrective side is empty, the corresponding gain is zero. Hence

$$\gamma \mathbb{E}\left[\int_0^T e^{\zeta|I_{t^-}|}\mathbf{1}_{\{\text{corrective side non-empty at } t^-\}}\,\mathrm{d}t\right] \leq \mathbb{E}\left[\int_0^T \Lambda^{\mathrm{corr}}(t)\,\mathrm{d}t\right] \leq N_{\max},$$

which proves (20).

For the full-interval bound, decompose

$$\begin{aligned}\int_0^T e^{\zeta|I_{t^-}|}\,\mathrm{d}t = &\int_0^T e^{\zeta|I_{t^-}|}\mathbf{1}_{\{\text{corrective side non-empty at } t^-\}}\,\mathrm{d}t \\ &+ \int_0^T e^{\zeta|I_{t^-}|}\mathbf{1}_{\{\text{corrective side empty at } t^-\}}\,\mathrm{d}t.\end{aligned}$$

The first term is controlled by (20); the second is at most $e^{\zeta \sup_{t\leq T}|I_{t^-}|}T_{\mathrm{ecs}}$. Taking expectations proves (21). □

Proof of Corollary 3 For every $M > 0$,

$$\mathbf{1}_{\{|I_{t^-}|>M\}} \leq e^{-\zeta M}e^{\zeta|I_{t^-}|}.$$

Multiplying by the indicator that the corrective side is non-empty, integrating, and applying (20) gives (22). For the full-interval bound, split the occupation time according to whether the corrective side is empty:

$$\begin{aligned}\mathbb{E}\left[\int_0^T \mathbf{1}_{\{|I_{t^-}|>M\}}\,\mathrm{d}t\right] &\leq \mathbb{E}\left[\int_0^T \mathbf{1}_{\{|I_{t^-}|>M\}}\mathbf{1}_{\{\text{corrective side non-empty at } t^-\}}\,\mathrm{d}t\right] + \mathbb{E}[T_{\mathrm{ecs}}] \\ &\leq \frac{N_{\max}}{\gamma}e^{-\zeta M} + \mathbb{E}[T_{\mathrm{ecs}}].\end{aligned}$$

□

A.4 Pathwise mispricing control

Lemma 9 (Deterministic microprice bound) *Under Assumptions* (S1)–(S6),

$$\sup_{t\in[0,T]}\left|P_t^{\mathrm{micro}} - P_0^{\mathrm{micro}}\right| \leq 2\delta_{\max}(N_{\max} + N_{\mathrm{steps}}) = \Delta_P. \tag{A7}$$

Economically, the learned market makers operate in a discrete LOB with controlled quote distances, so endogenous price formation remains bounded even when liquidity demand is clustered.

Proof The microprice changes only at trade times and decision times. By Proposition 1, there are at most $N_{\max}$ trade times, and by construction there are at most N_{steps} decision times. Between such times, the book state is constant and so is P_t^{micro}.

Whenever both sides are non-empty,

$$P_t^{\mathrm{micro}} = \frac{A_tQ_t^b + B_tQ_t^a}{Q_t^a + Q_t^b} \in [B_t, A_t].$$

Assumptions (S1)–(S6) give $A_t - P_t^{\text{micro}} \le \delta_{\max}$ and $P_t^{\text{micro}} - B_t \le \delta_{\max}$, hence $A_t - B_t \le 2\delta_{\max}$. At a trade or decision time, the old and new microprices are therefore separated by at most $2\delta_{\max}$. If a side becomes empty, the fallback value P_t^{last} is an executed price at a level within the same admissible range, so the same one-step bound applies. Summing over at most $N_{\max} + N_{\text{steps}}$ changes proves (A7). □

Proof of Proposition 4 For every $t \in [0, T]$,

$$\begin{aligned} |I_t| &= \left| S_t - P_t^{\text{micro}} \right| \\ &= \left| (S_t - S_0) - (P_t^{\text{micro}} - P_0^{\text{micro}}) + (S_0 - P_0^{\text{micro}}) \right| \\ &\le |S_t - S_0| + \left| P_t^{\text{micro}} - P_0^{\text{micro}} \right| + |I_0| \,. \end{aligned}$$

Taking suprema and using Lemma 9 gives

$$\sup_{t \le T} |I_t| \le \sup_{t \le T} |S_t - S_0| + \Delta_P + |I_0| = \sup_{t \le T} |S_t - S_0| + K_0.$$

Thus, for $K > K_0$,

$$\mathbb{P}\left(\sup_{t \le T} |I_t| \le K \right) \ge \mathbb{P}\left(\sup_{t \le T} |S_t - S_0| \le K - K_0 \right).$$

Since $S_t - S_0 = \sigma W_t$, the reflection principle and the standard Gaussian tail bound imply, for $x > 0$,

$$\mathbb{P}\left(\sup_{t \le T} |W_t| > x \right) \le 2\mathbb{P}(|W_T| > x) \le 2 \exp\left(-\frac{x^2}{2T} \right).$$

Applying this with $x = (K - K_0)/\sigma$ yields (25). Sending $K \to \infty$ gives $\mathbb{P}(\sup_{t \le T} |I_t| < \infty) = 1$. □

The moment claim in the main-text remark follows from the standard tail-integration identity. For example, applying (25) to $X = \sup_{t \le T} |I_t|$ gives

$$\mathbb{E}[X^p] = p \int_0^\infty K^{p-1} \mathbb{P}(X > K) \, \mathrm{d}K \le K_0^p + p \int_{K_0}^\infty K^{p-1} 2 \exp\left(-\frac{(K - K_0)^2}{2\sigma^2 T} \right) \mathrm{d}K,$$

which is finite. The same Gaussian tail yields the stated sub-Gaussian-square integrability for every $\lambda < 1/(2\sigma^2 T)$.

For reference, the main constants appearing in the stability bounds are

$$\begin{aligned} N_{\max} &= \frac{2 D_{\text{tot}} N_{\text{steps}}}{q_{\min}} \le \frac{2 D_{\text{tot}} T}{q_{\min} \Delta_{\min}^{\text{dec}}}, \\ \gamma &= \psi e^{-\eta \delta_{\max}} \mu_{\min}, \\ \Delta_P &= 2\delta_{\max}(N_{\max} + N_{\text{steps}}), \\ K_0 &= |I_0| + \Delta_P. \end{aligned}$$

Appendix B Cancellation mechanisms

Each market maker stores its outstanding quotes by price-level and maintains an *age counter* per quote. The age is incremented once per decision step (i.e. on the discrete market-maker grid), so the time-to-live (TTL) is expressed in *steps*, not in continuous time. Deterministic cancellations occur at three points within each decision step:

1. **Time-to-live (TTL) at the beginning of the step.** At the beginning of each market-maker decision step, all existing quotes have their age incremented and any quote with age exceeding a fixed threshold is removed. To prevent pathological behavior where agents are forced to constantly repost their best quotes (thereby losing queue priority), the environment protects the closest quotes to the current mid on each side (closest in absolute tick distance). Thus, an old quote is removed only if it is both expired and not among the protected top-k levels.
2. **Enforcing per-side outstanding-order caps before posting.** Immediately before posting a new bid (respectively ask), the environment enforces two per-side caps: (i) a cap on the number of distinct active price levels on that side (denoted as $S_{\text{own}} > 0$) and (ii) a cap on the total outstanding quantity on that side. The environment never rejects the agent's posting action; instead, it makes room by canceling existing levels on that side until both caps are satisfied for the post-trade state including the new order. The default cancellation rule implemented in code cancels the level farthest from the current mid on that side (least competitive), with ties broken by age (oldest first). Formally, letting p denote a candidate price level and $\text{age}(p)$ its step-age, the environment ranks candidates by

$$d_{\text{bid}}(p) = M_t - p, \qquad d_{\text{ask}}(p) = p - M_t,$$

   and cancels the level maximizing the lexicographic key $(d_{\text{side}}(p), \text{age}(p))$. Importantly, the cap logic is applied only if the incoming order size is strictly positive, so an action corresponding to size 0 ("no new order") does not trigger cancellations and does not interact with the caps.
3. **Bounding quotes after continuous-time taker trading.** After the market evolves in continuous time over the interval $(t_n^{\text{dec}}, t_{n+1}^{\text{dec}}]$ due to market-taker arrivals, the mid-price is recomputed and any remaining quotes that drift beyond a fixed admissible distance from the updated mid are removed to keep the book bounded. Concretely, define L as maximum observable horizon. Then all bids satisfying $M_t - p > L \cdot \Delta p$ and all asks satisfying $p - M_t > L \cdot \Delta p$ are deleted from both the order book and the market maker's internal records.

Together, these rules remove cancellation degeneracy from the action space, reduce the frequency of extreme states caused by unbounded stale quotes, and make the effect of a placement action substantially more predictable from the agent's perspective.

Appendix C Observation space decomposition

The observation space can be divided into three parts: the actors' observable state, the global state used by the critic and an agent indicator. We start of by describing the agents' observation space.

Agents' observation space

Each agents' observation space can be further qualitatively divided into three sections - (i) scalars, (ii) LOB embedding, and (iii) history. This division matters greatly in the architecture. Below we describe the construction of each of these parts.

(i) Actor scalar features. The scalar block contains the agent's own inventory and cash, time elapsed measure, market price, last transaction price together with a (possibly absent) private signal:

$$u_t^{(i)} = \Big(q_t^{(i)},\ c_t^{(i)},\ M_t,\ P_t^{\text{last}},\ \tau_t,\ \widetilde{S}_t^{(i)},\ \mathbf{1}\{\text{agent } i \text{ informed}\} \Big) \in \mathbb{R}^7.$$

Here $\mathbf{1}\{\cdot\}$ is the signal-availability indicator and $\tau_t \in [0,1]$ is normalized time elapsed in the episode. Optionally, when regime information is exposed to the actor, this block is augmented by

$$r_t = \big(k_t, \psi_t^{\text{obs}}, \sigma_t^{\text{obs}}\big) \in \mathbb{R}^3,$$

where $k_t \in \{0,\ 1\}$ denotes whether the regime is known to agents and $\psi_t^{\text{obs}},\ \sigma_t^{\text{obs}}$ are respectively observed market volatility and informedness regime. Otherwise these variables are omitted from the actor input.

(ii) Public LOB features and liquidity decomposition. Let L denote the maximum observable distance from mid in ticks and let tick size be denoted as Δp. Everything farther than L ticks from the mid-price is canceled (see Appendix B), and agents observe the LOB only within L ticks of the mid-price after canonicalization. Using the level grid defined by:

$$p_{t,\ell}^b = \text{canon}(R_t - \ell\,\Delta p)\,, \qquad p_{t,\ell}^a = \text{canon}(R_t + \ell\,\Delta p)\,, \qquad \ell = 1, \ldots, L,$$

where the reference price $R_t = M_t$ if both sides exist and $R_t = P_t^{\text{last}}$ otherwise. Recall the FIFO queue structure defined in Equation 2. Thanks to this construction, rather than collapsing each level to a single depth number, the observation exposes a truncated queue representation that preserves time-priority structure up to a fixed depth. Fix an integer $S_{\text{own}} \geq 1$ stating how many different quotes are allowed per agent at one price level and define $S_{\text{oth}} = S_{\text{own}} + 1$. Importantly, this is not a cancellation mechanism but a posting constraint. If the agent already occupies the maximum number of own queue slots at a given price level, an additional order at that level is ignored. In regular market conditions this constraint has negligible impact on the simulated dynamics, while preventing pathological over-posting during training and preserving queue-priority effects. For a fixed agent i, we define four arrays

$$V_t^{\text{own},b}(i) \in \mathbb{R}^{L \times S_{\text{own}}}, \quad V_t^{\text{own},a}(i) \in \mathbb{R}^{L \times S_{\text{own}}}, \qquad V_t^{\text{oth},b}(i) \in \mathbb{R}^{L \times S_{\text{oth}}}, \quad V_t^{\text{oth},a}(i) \in \mathbb{R}^{L \times S_{\text{oth}}}.$$

in the following fashion. For any bid-side price level $p \in \mathcal{B}_t$, recall from Definition 2 that

$$\mathcal{Q}_t^b(p) = \big((j_{1,p}^b, q_{1,p}^b), \ldots, (j_{n_p,p}^b, q_{n_p,p}^b)\big)_t.$$

For a fixed agent i, define the set of FIFO positions occupied by that agent at price p by

$$\mathcal{I}_t^b(p,i) \coloneqq \big\{m \in \{1,\ldots,n_p\} : j_{m,p}^b = i\big\}.$$

Let

$$K_t^b(p,i) \coloneqq |\mathcal{I}_t^b(p,i)|$$

be the number of stored own bid entries of agent i at price p. Notice that then $K_t^b(p,i) \leq S_{\text{own}}$. If $K_t^b(p,i) \geq 1$, let

$$\pi_{t,1}^b(p,i) < \pi_{t,2}^b(p,i) < \cdots < \pi_{t,K_t^b(p,i)}^b(p,i)$$

denote the elements of $\mathcal{I}_t^b(p,i)$ in increasing FIFO order.

Now fix $\ell \in \{1,\ldots,L\}$ and set $p = p_{t,\ell}^b$, so the ℓ-th price level available on the bid side. The own-liquidity tensor is defined by

$$V_{t,\ell,s}^{\text{own},b}(i) = \begin{cases} q_{\pi_{t,s}^b(p,i),\,p}^b, & 1 \leq s \leq K_t^b(p,i), \\ 0, & K_t^b(p,i) < s \leq S_{\text{own}}. \end{cases}$$

Thus, $V_{t,\ell,s}^{\text{own},b}(i)$ stores the sizes of the S_{own} own bid orders of agent i at level ℓ, preserving FIFO priority.

The foreign-liquidity tensor

$$V_{t,\ell,s}^{\text{oth},b}(i), \qquad s = 1,\ldots,S_{\text{oth}}, \qquad S_{\text{oth}} = S_{\text{own}} + 1,$$

uses the same queue positions and has fixed slot semantics:

$$V_{t,\ell,1}^{\text{oth},b}(i) = \begin{cases} \displaystyle\sum_{\substack{m < \pi_{t,1}^b(p,i) \\ j_{m,p}^b \neq i}} q_{m,p}^b, & K_t^b(p,i) \geq 1, \\ 0, & K_t^b(p,i) = 0, \end{cases}$$

$$V_{t,\ell,s}^{\text{oth},b}(i) = \begin{cases} \displaystyle\sum_{\substack{\pi_{t,s-1}^b(p,i) < m < \pi_{t,s}^b(p,i) \\ j_{m,p}^b \neq i}} q_{m,p}^b, & 2 \leq s \leq K_t^b(p,i), \\ 0, & K_t^b(p,i) < s \leq S_{\text{own}}, \end{cases}$$

$$V_{t,\ell,S_{\text{oth}}}^{\text{oth},b}(i) = \begin{cases} \displaystyle\sum_{\substack{m > \pi_{t,K_t^b(p,i)}^b(p,i) \\ j_{m,p}^b \neq i}} q_{m,p}^b, & K_t^b(p,i) \geq 1, \\ \displaystyle\sum_{\substack{m=1,\ldots,n_p \\ j_{m,p}^b \neq i}} q_{m,p}^b, & K_t^b(p,i) = 0. \end{cases}$$

Hence the competing liquidity is decomposed into volume ahead of the first own order, volume between consecutive own orders, and volume behind the last stored own order. If agent i has no own order at that price level, all competing liquidity is placed in the final slot S_{oth}.

The ask-side tensors are defined analogously. For any ask-side price level $p \in \mathcal{A}_t$, write

$$\mathcal{Q}_t^a(p) = \big((j_{1,p}^a, q_{1,p}^a), \ldots, (j_{n_p,p}^a, q_{n_p,p}^a)\big)_t,$$

define $\mathcal{I}_t^a(p,i)$, $K_t^a(p,i)$ and the ordered positions $\pi_{t,s}^a(p,i)$ in the same way, and then instantiate the definitions above at $p = p_{t,\ell}^a$ (so ℓ-th level on the ask side) to obtain $V_{t,\ell,s}^{\text{own},a}(i)$ and $V_{t,\ell,s}^{\text{oth},a}(i)$.

This representation preserves the economically relevant time-priority information at each level: the agent observes not only how much competing liquidity exists at a given price, but also whether it is ahead of, between, or behind its own queue entries. At the same time, the resulting features are fixed-shape arrays, hence directly compatible with feed-forward policies.

(iii) History embedding. The history tensor is defined as:

$$H_t^{(i)} = \big[\xi_{t-H}^{(i)}, \ldots, \xi_{t-1}^{(i)}\big]^\top \in \mathbb{R}^{H\times 6},$$

where

$$\xi_t^{(i)} = \Big(B_t,\ A_t,\ P_t^{\text{last}},\ \widetilde{S}_t^{(i)},\ q_t^{(i)},\ c_t^{(i)}\Big) \in \mathbb{R}^6.$$

The buffer is padded at episode start to provide a constant shape.

Centralized critic state (CTDE). For centralized training, the critic receives a privileged state vector

$$s_t \in \mathcal{S}.$$

This vector is not part of the decentralized actor observation $o_t^{(i)}$. In addition, the environment returns a one-hot identifier

$$e^{(i)} \in \{0,1\}^N,$$

indicating for which agent the value function is being evaluated.

In the implementation, s_t is a single flat vector built from a global scalar block, all per-agent scalar blocks, and the market-wide queue tensors. Define

$$g_t = \Big(M_t,\ P_t^{\text{last}},\ S_t,\ \tau_t,\ \psi_t^{\text{obs}},\ \sigma_t^{\text{obs}}\Big) \in \mathbb{R}^6,$$

where τ_t is normalized time elapsed, and where $\psi_t^{\text{obs}} = \sigma_t^{\text{obs}} = 0$ whenever the current regime is hidden from the learning pipeline as previously.

For each agent $j \in \{1, \ldots, N\}$, define the per-agent scalar block

$$a_t^{(j)} = \Big(q_t^{(j)},\ c_t^{(j)},\ \alpha_j,\ \phi_j,\ \widetilde{S}_t^{(j)},\ \mathbf{1}\{\text{agent } j \text{ informed}\}\Big) \in \mathbb{R}^6,$$

where α_j, ϕ_j, are the policy specific inventory penalties that play a role in the reward structure in Section 6.5. Using the same queue encoding as in the actor observation, let

$$V_t^{\text{own},b}(j) \in \mathbb{R}^{L\times S_{\text{own}}}, \quad V_t^{\text{oth},b}(j) \in \mathbb{R}^{L\times S_{\text{oth}}}, \quad V_t^{\text{own},a}(j) \in \mathbb{R}^{L\times S_{\text{own}}}, \quad V_t^{\text{oth},a}(j) \in \mathbb{R}^{L\times S_{\text{oth}}},$$

denote the bid/ask own- and foreign-liquidity tensors from agent j's perspective, where $S_{\text{oth}} = S_{\text{own}} + 1$.

The centralized critic state is then the concatenation

$$\begin{aligned} s_t = \Big(& g_t,\ a_t^{(1)}, \dots, a_t^{(N)}, \\ & \text{vec}\big(V_t^{\text{own},b}(1)\big), \dots, \text{vec}\big(V_t^{\text{own},b}(N)\big), \\ & \text{vec}\big(V_t^{\text{oth},b}(1)\big), \dots, \text{vec}\big(V_t^{\text{oth},b}(N)\big), \\ & \text{vec}\big(V_t^{\text{own},a}(1)\big), \dots, \text{vec}\big(V_t^{\text{own},a}(N)\big), \\ & \text{vec}\big(V_t^{\text{oth},a}(1)\big), \dots, \text{vec}\big(V_t^{\text{oth},a}(N)\big)\Big), \end{aligned}$$

where the ordering is agent-major, then level-major, then slot-major within each tensor.

Hence

$$\dim(s_t) = 6 + 6N + N\Big(LS_{\text{own}} + LS_{\text{oth}} + LS_{\text{own}} + LS_{\text{oth}}\Big) = 6 + 6N + 2NL(2S_{\text{own}} + 1).$$

Appendix D Agents' architecture

(i) Scalar encoder.

Let $u_t^{(i)} \in \mathbb{R}^{d_{\text{sc}}}$ denote the concatenation of the scalar keys used by the actor in code (see Appendix C). These scalars are embedded by a one-layer MLP with tanh activation:

$$h_{t,\text{sc}}^{(i)} = \tanh\big(W_{\text{sc}}\bar{u}_t^{(i)} + b_{\text{sc}}\big) \in \mathbb{R}^{d_{\text{sc_out}}}.$$

(ii) Liquidity cross attention.

Recall that each agent observes two depth tensors describing its own outstanding orders and the liquidity posted by others:

$$V_t^{\text{own},b}(i) \in \mathbb{R}^{L\times S_{\text{own}}}, \quad V_t^{\text{own},a}(i) \in \mathbb{R}^{L\times S_{\text{own}}}, \quad V_t^{\text{oth},b}(i) \in \mathbb{R}^{L\times S_{\text{oth}}}, \quad V_t^{\text{oth},a}(i) \in \mathbb{R}^{L\times S_{\text{oth}}},$$

where L is the number of tracked price levels (in ticks away from the reference) and $S_{\text{own}}, S_{\text{oth}}$ are the per-level feature widths in the environment (see Appendix C).

For a fixed agent i and level $\ell \in \{1, \dots, L\}$ we define the *own* and *other* level tokens by concatenation:

$$\mathbf{u}_{t,\ell}^{(i)} \coloneqq \big[\, V_t^{\text{own},b}(i)[\ell,:],\ V_t^{\text{own},a}(i)[\ell,:] \,\big] \in \mathbb{R}^{2S_{\text{own}}},$$

$$\mathbf{v}_{t,\ell}^{(i)} \;\coloneqq\; \big[\, V_t^{\mathrm{oth},b}(i)[\ell,:],\; V_t^{\mathrm{oth},a}(i)[\ell,:] \,\big] \in \mathbb{R}^{2S_{\mathrm{oth}}}.$$

Therefore, these vectors represent FIFO order liquidity at the ℓ-th level from the mid-price on both sides, for both own and foreign liquidity. Although the bid and ask prices differ, the level index is defined relative to the mid-price in both cases. Stacking the level tokens yields two matrices

$$U_t^{(i)} \in \mathbb{R}^{L\times 2S_{\mathrm{own}}}, \quad V_t^{(i)} \in \mathbb{R}^{L\times 2S_{\mathrm{oth}}}, \qquad U_t^{(i)}[\ell,:] = \mathbf{u}_{t,\ell}^{(i)}, \;\; V_t^{(i)}[\ell,:] = \mathbf{v}_{t,\ell}^{(i)}.$$

To combine own exposure with the surrounding liquidity landscape, the actor applies a single cross-attention layer in which each level of the agent's own token sequence queries all levels of the other-liquidity sequence. In particular, with key dimension d_k and output dimension d_{att}, the implementation computes

$$Q = U_t^{(i)} W_q \in \mathbb{R}^{L\times d_k}, \quad K = V_t^{(i)} W_k \in \mathbb{R}^{L\times d_k}, \quad \mathcal{V} = V_t^{(i)} W_v \in \mathbb{R}^{L\times d_{\mathrm{att}}},$$

followed by

$$A = \mathrm{softmax}\!\left(\frac{QK^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^{L\times L}, \qquad C_t^{(i)} = A\,\mathcal{V} \in \mathbb{R}^{L\times d_{\mathrm{att}}}.$$

Thus, for each level ℓ, the row $C_{t,\ell}^{(i)}$ is a convex combination of transformed other-liquidity vectors across all levels, with weights determined by the compatibility between the agent's own queue configuration at ℓ and the competitive liquidity profile at every level.

History encoder.

Using a history tensor $H_t^{(i)} \in \mathbb{R}^{H\times F}$ (as introduced in Appendix C), the actor embeds it with a single-layer GRU (Cho et al. (2014)) of hidden size d_h:

$$(h_{t,1}^{(i)}, \ldots, h_{t,H}^{(i)}) = \mathrm{GRU}\big(H_t^{(i)}\big), \qquad h_{t,\mathrm{hist}}^{(i)} \;\coloneqq\; h_{t,H}^{(i)} \in \mathbb{R}^{d_h}.$$

Policy head.

Let $\mathrm{vec}(\cdot)$ denote row-wise vectorization. The actor feature vector is

$$z_t^{(i)} \;=\; \Big[h_{t,\mathrm{sc}}^{(i)},\; h_{t,\mathrm{hist}}^{(i)},\; \mathrm{vec}\big(C_t^{(i)}\big)\Big] \in \mathbb{R}^{d_{\mathrm{sc_out}}+d_h+d_{\mathrm{att}}L}.$$

A single policy head is a two-layer MLP with tanh nonlinearity and hidden width d_{head}:

$$y_t^{(i)} = W_2\,\tanh(W_1 z_t^{(i)} + b_1) + b_2 \in \mathbb{R}^{d_\pi}.$$

Mixture-of-experts (MoE) actor head.

Policy heads are merged into a mixture-of-experts head. The only difference between agent types is the *gating regime*: agents with access to a fundamental proxy (i.e.

$\mathbf{1}\{\text{agent } i \text{ informed}\} = 1$) use an inventory×mispricing gate, whereas agents without that information use an inventory-only gate.

Let $z_t^{(i)}$ be the actor feature vector defined above and let d_π denote the total number of logits required by RLlib for the discrete action distribution. For a fixed number of experts E, each expert produces logits

$$y_t^{(i,e)} = f_\pi^{(e)}(z_t^{(i)}) \in \mathbb{R}^{d_\pi}, \qquad e \in \{1, \dots, E\},$$

where $f_\pi^{(e)}$ is a two-layer MLP with tanh activations and hidden width 256. The final logits are a convex combination

$$y_t^{(i)} = \sum_{e=1}^{E} w_t^{(i,e)}\, y_t^{(i,e)}, \qquad w_t^{(i,e)} \geq 0, \quad \sum_{e=1}^{E} w_t^{(i,e)} = 1.$$

The gating weights $w_t^{(i,e)}$ are deterministic functions of low-dimensional indicators derived from the observation introduced in the next paragraph.

Inventory gate (all agents). The inventory gate produces weights for negative vs. positive inventory regimes as a hard split,

$$w_{\text{pos}}^{(i)} = \mathbf{1}\{q_t^{(i)} > 0\}, \qquad w_{\text{neg}}^{(i)} = 1 - w_{\text{pos}}^{(i)}.$$

Inventory-only MoE (uninformed agents). For agents with $\mathbf{1}\{\text{agent } i \text{ informed}\} = 0$ the MoE uses $E = 2$ experts corresponding to the two inventory regimes (neg/pos):

$$y_t^{(i)} = w_{\text{neg}}^{(i)}\, y_t^{(i,\text{neg})} + w_{\text{pos}}^{(i)}\, y_t^{(i,\text{pos})}.$$

Inventory × mispricing MoE (agents with fundamental proxy). For agents with $\mathbf{1}\{\text{agent } i \text{ informed}\} = 1$ an additional mispricing gate is applied. Using the actor-side proxy $\widetilde{S}_t^{(i)}$ and mid-price M_t, define the signed mispricing indicator

$$\Delta_t^{(i)} := M_t - \widetilde{S}_t^{(i)}.$$

The under/over weights $(w_{\text{under}}^{(i)}, w_{\text{over}}^{(i)})$ are computed from $\Delta_t^{(i)}$ using the same hard gating form. The MoE then uses $E = 4$ experts corresponding to the Cartesian product of regimes

$$(\text{neg}, \text{pos}) \times (\text{under}, \text{over}),$$

with mixture weights given by products:

$$\begin{aligned} w_t^{(i,\text{neg,under})} &= w_{\text{neg}}^{(i)} w_{\text{under}}^{(i)}, \quad & w_t^{(i,\text{neg,over})} &= w_{\text{neg}}^{(i)} w_{\text{over}}^{(i)}, \\ w_t^{(i,\text{pos,under})} &= w_{\text{pos}}^{(i)} w_{\text{under}}^{(i)}, \quad & w_t^{(i,\text{pos,over})} &= w_{\text{pos}}^{(i)} w_{\text{over}}^{(i)}. \end{aligned}$$

Finally,

$$y_t^{(i)} = \sum_{r\in\{(\text{neg,under}),(\text{neg,over}),(\text{pos,under}),(\text{pos,over})\}} w_t^{(i,r)}\, y_t^{(i,r)}.$$

This design hard-codes economically meaningful regime structure (inventory sign and, when available, mispricing sign) while preserving a fully differentiable mapping from observations to action logits.

Appendix E Centralized critic's architecture

Critic scalar encoder.

The critic first builds a scalar feature vector by concatenating: (i) the global scalars g_t, (ii) all per-agent scalars $a_t^{(j)}$ augmented by an "is-current-agent" indicator extracted from $\mathbf{e}^{(i)}$, and (iii) the history embedding (from the same GRU that is used by the actor) $h_{t,\text{hist}}^{(i)}$ (see Appendix C).

Concretely, define the augmented per-agent tokens

$$\tilde{a}_t^{(j|i)} = \big[a_t^{(j)},\ \mathbf{1}\{j=i\}\big] \in \mathbb{R}^7, \qquad \tilde{a}_t^{(\cdot|i)} = \big[\tilde{a}_t^{(1|i)},\dots,\tilde{a}_t^{(N|i)}\big] \in \mathbb{R}^{7N},$$

and set

$$x_{t,\text{sc}}^{(i)} = \big[g_t,\ \tilde{a}_t^{(\cdot|i)},\ h_{t,\text{hist}}^{(i)}\big].$$

The scalar encoder is a single-layer MLP with tanh activation:

$$h_{t,\text{sc}}^{(i)} = \tanh\big(W_{\text{sc}} x_{t,\text{sc}}^{(i)} + b_{\text{sc}}\big) \in \mathbb{R}^{d_{\text{sc_out}}^{\text{crit}}}.$$

Cross-attention over the market-wide queue state.

Using the same per-level queue vectors as in the actors subsections, define for each agent j and level ℓ:

$$\mathbf{w}_{t,\ell}^{(j)} \coloneqq \big[\,\mathbf{u}_{t,\ell}^{(j)},\ \mathbf{v}_{t,\ell}^{(j)}\,\big] \in \mathbb{R}^{F_q}, \qquad F_q = 2S_{\text{own}} + 2S_{\text{oth}} = 4S_{\text{own}} + 2.$$

Stacking these tokens over all agents and all tracked levels yields

$$W_t \in \mathbb{R}^{N\times L\times F_q}, \qquad W_t[j,\ell,:] = \mathbf{w}_{t,\ell}^{(j)}.$$

Using the one-hot identifier $\mathbf{e}^{(i)}$, define the current-agent query sequence

$$U_t^{(i),\text{critic}} \in \mathbb{R}^{L\times F_q}, \qquad U_t^{(i),\text{critic}}[\ell,:] = \sum_{j=1}^{N} e_j^{(i)}\, W_t[j,\ell,:] = W_t[i,\ell,:],$$

and the market-wide context sequence

$$V_t^{(i),\text{critic}} \in \mathbb{R}^{(NL)\times F_q}, \qquad V_t^{(i),\text{critic}}[(j-1)L+\ell,:] = W_t[j,\ell,:].$$

Then, in the same manner as in the actor part, the critic cross-attention block is written as

$$Q^{\text{critic}} = U_t^{(i),\text{critic}} W_q^{\text{critic}} \in \mathbb{R}^{L \times d_k^{\text{critic}}},$$
$$K^{\text{critic}} = V_t^{(i),\text{critic}} W_k^{\text{critic}} \in \mathbb{R}^{(NL) \times d_k^{\text{critic}}},$$
$$\mathcal{V}^{\text{critic}} = V_t^{(i),\text{critic}} W_v^{\text{critic}} \in \mathbb{R}^{(NL) \times d_{\text{att}}^{\text{critic}}},$$

followed by

$$A_t^{(i),\text{critic}} = \text{softmax}\left(\frac{Q^{\text{critic}} (K^{\text{critic}})^\top}{\sqrt{d_k^{\text{critic}}}} \right) \in \mathbb{R}^{L \times (NL)}, \qquad C_t^{(i),\text{critic}} = A_t^{(i),\text{critic}} \mathcal{V}^{\text{critic}} \in \mathbb{R}^{L \times d_{\text{att}}^{\text{critic}}}.$$

Thus, for each level ℓ, the row $C_{t,\ell}^{(i),\text{critic}}$ is a convex combination of transformed market-wide queue tokens across all agents and all tracked levels, with weights determined by the compatibility between the current agent's queue state at level ℓ and the full market-wide queue configuration.

Value features and MoE value head.

The critic feature vector is

$$x_t^{(i)} = \left[h_{t,\text{sc}}^{(i)},\ C_{t,\ell}^{(i),\text{critic}} \right] \in \mathbb{R}^{d_{\text{sc_out}}^{\text{crit}} + d_{\text{att}}^{\text{crit}} L}.$$

In the implementation the value is produced by a 2×2 mixture-of-experts (four experts), gated by (i) the sign of the current agent's inventory and (ii) the sign of **global mispricing**.

Each expert is a two-layer tanh MLP mapping $x_t^{(i)}$ to a scalar:

$$v_t^{(i,e)} = f_V^{(e)}(x_t^{(i)}) \in \mathbb{R}, \qquad e \in \{1, 2, 3, 4\}.$$

The mispricing proxy is computed from the global scalars as

$$\Delta_t^{\text{critic}} = M_t - S_t.$$

Inventory for gating is extracted from the centralized per-agent block using $\mathbf{e}^{(i)}$:

$$I_t^{(i)} = \sum_{j=1}^{N} e_j^{(i)}\, q_t^{(j)}.$$

Using the same hard/soft gating map $\mathcal{G}$ as in the actor (with different slopes), obtain

$$(w_{\text{neg}}^{(i)}, w_{\text{pos}}^{(i)}) = \mathcal{G}(I_t^{(i)}), \qquad (w_{\text{under}}, w_{\text{over}}) = \mathcal{G}(\Delta_{t,\text{critic}}),$$

and form the 2×2 mixture weights

$$\big(w^{(i)}_{\text{neg,under}}, w^{(i)}_{\text{neg,over}}, w^{(i)}_{\text{pos,under}}, w^{(i)}_{\text{pos,over}}\big) = \big(w^{(i)}_{\text{neg}} w_{\text{under}},\ w^{(i)}_{\text{neg}} w_{\text{over}},\ w^{(i)}_{\text{pos}} w_{\text{under}},\ w^{(i)}_{\text{pos}} w_{\text{over}}\big).$$

The final value is the corresponding convex combination:

$$V^{(i)}_t = w^{(i)}_{\text{neg,under}}\, v^{(i,1)}_t + w^{(i)}_{\text{neg,over}}\, v^{(i,2)}_t + w^{(i)}_{\text{pos,under}}\, v^{(i,3)}_t + w^{(i)}_{\text{pos,over}}\, v^{(i,4)}_t.$$

Appendix F Parameters

This section collects the numerical settings used throughout the simulations and training. In particular, the reported optimization hyperparameters include the Kullback–Leibler (KL) target schedule, the number of stochastic gradient descent (SGD) epochs per update, and the generalized advantage estimation (GAE) parameter used within the proximal policy optimization (PPO) procedure.

Parameter	Symbol	Value	Comment
Number of bins	K	3	
Bins of volumes	–	$[0, 40]$, $[40, 80]$, $[80, 200]$	
Number of exponentials	R	2	
Self-excitation kernel norms (same side)	$A^{AA} = A^{BB}$	(0.3, 0.12, 0.07)	Scaled by $s_\alpha = 0.1$
Cross-excitation kernel norms (across sides)	$A^{AB} = A^{BA}$	(0.07, 0.035, 0.02)	Scaled by $s_\alpha = 0.1$
Exponential decay parameters	β	(3.7, 4.2)	
Exponential mixture weights	w	(0.7, 0.3)	
Baseline intensities	μ	(3.5, 2.5, 1.5)	
Shape parameter	η	2.8	
Scale parameter	ζ	70	

Table F1: Hawkes process parameters.

Regime mode	Description	Switching time	Changes	Regime values
`Regime model 2`	One change at a fixed time to a random regime	$t_c = 50$	1	$(\psi, \sigma) \in \Psi \times \Sigma$
`Regime model 4`	One change at a random time to a random regime	$t_c \sim \mathcal{U}\{10, \ldots, 90\}$	1	$(\psi, \sigma) \in \Psi \times \Sigma$
`Regime model 5`	Fixed number of changes on a fixed grid to random regimes	$t_c \in \{20, 50, 80\}$	3	$(\psi, \sigma) \in \Psi \times \Sigma$
`Regime model 6`	Fixed number of changes on a random grid to random regimes	$t_c^{(m)} \sim \mathcal{U}\{10, \ldots, 90\}$	3	$(\psi, \sigma) \in \Psi \times \Sigma$
`Regime model 7`	Exponential waiting times for random regimes	$\Delta t_c \sim \mathrm{Exp}(\lambda)$	Random	$(\psi, \sigma) \in \Psi_\lambda \times \Sigma$

Table F2: Regime-switching specifications used in the non-stationary market experiments. The variable ψ denotes the informed ratio and σ denotes the market-volatility parameter. Regime models draw the post-change state from the full regime grid $\Psi = \{0.00, 0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95, 0.99, 1.00\}$ and $\Sigma = \{0.001, 0.005, 0.010, 0.020\}$. For each regime model, both known and unknown variants can be evaluated; in the known variant, the current regime variables are included in the agents' observations, whereas in the unknown variant they are hidden from the learning pipeline.

Policy	Agents	ϕ	α	σ_{signal}
Conservative	4	$3 \cdot 10^{-4}$	$3 \cdot 10^{-3}$	0
Quasi-informed conservative	4	$3 \cdot 10^{-4}$	$3 \cdot 10^{-3}$	10^{-5}
Aggressive	4	10^{-4}	10^{-3}	0
Quasi-informed aggressive	4	10^{-4}	10^{-3}	10^{-5}

Table F3: Agent configuration and policy-specific parameters. Additionally, the learning rate for both informed and uninformed conservative policies is scheduled as $\{(0, 9 \cdot 10^{-4}), (15, 3 \cdot 10^{-4}), (40, 10^{-4}), (100, 5 \cdot 10^{-5})\}$, while for both aggressive policies it is scheduled as $\{(0, 8 \cdot 10^{-4}), (15, 3 \cdot 10^{-4}), (40, 10^{-4}), (100, 5 \cdot 10^{-5})\}$. Values between the scheduled points are linearly interpolated for both schedules. In each pair, the first entry denotes the iteration and the second the scheduled value.

Parameter	Value
Total number of agents	16
Agents posting volumes	$[1, 5, 10, 20]$
KL target schedule	$\{(0, 0.08), (10, 0.08), (25, 0.02), (45, 0.01), (100, 0.005)\}$
Entropy schedule	$\{(0, 0.003), (80, 0)\}$
Maximum resting quantity per agent per side	60
Observation history length	8
Protected levels from cancellation	2
Postable levels	4
Observable levels before cancellation	6

Table F4: Common agent constraints and schedules. Values between the scheduled points are linearly interpolated for all schedules. In each pair, the first entry denotes the iteration and the second the scheduled value.

Parameter	Symbol	Value
Training batch size	–	81920
Minibatch size	–	40960
Number of SGD epochs	–	10
Value loss coefficient	c_{vf}	1
KL penalty coefficient	c_{kl}	0.2
Discount factor	γ	0.95
GAE parameter	λ	0.95
Policy clip parameter	$\varepsilon_{\text{clip}}$	0.15
Value clip parameter	ε_v	2

Table F5: PPO and training parameters.

Parameter	Symbol	Value
Interest rate	r	0
Drift	–	0
Rebate	R_{rebate}	0.002
Tick size	Δp	0.01
Initial price	$S_0 = M_0$	\$1

Table F6: Market setup.

Parameter	Symbol	Value
Scalar encoder's output dimension in actor's architecture	$d_{\text{sc_out}}$	64
Key dimension in actor's architecture's cross-attention	d_k	64
Output dimension in actor's architecture cross-attention	d_{att}	32
GRU hidden size	d_h	32
Policy head hidden width	d_{head}	256
Scalar encoder's output dimension in critic's architecture	$d^{\text{crit}}_{\text{sc_out}}$	64
Key dimension in critic's architecture cross-attention	d^{crit}_{k}	64
Output dimension in critic's architecture cross-attention	$d^{\text{crit}}_{\text{att}}$	32

Table F7: Neural network parameters.

References

Amihud, Y., & Mendelson, H. (1980). Dealership market: Market-making with inventory. *Journal of Financial Economics*, *8*(1), 31–53, https://doi.org/10.1016/0304-405X(80)90020-3

Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, *8*(3), 217–224, https://doi.org/10.1080/14697680701381228

Bank, P., Cartea, Á., Körber, L. (2023). Optimal execution and speculation with trade signals. *arXiv preprint arXiv:2306.00621*, *2306.00621*, arXiv:2306.00621, https://doi.org/10.48550/arXiv.2306.00621

Barucci, E., Mathieu, A., Sanchez-Betancourt, L. (2025). Market making with fads, informed, and uninformed traders. *arXiv preprint arXiv:2501.03658*, , https://doi.org/10.48550/arXiv.2501.03658

Black, F., & Scholes, M.S. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81*(3), 637–654, https://doi.org/10.1086/260062

Bodor, H., & Carlier, L. (2025). Deep learning meets queue-reactive: A framework for realistic limit order book simulation. *arXiv preprint arXiv:2501.08822*, *2501.08822*, arXiv:2501.08822, https://doi.org/10.48550/arXiv.2501.08822

Bormetti, G., Calcagnile, L.M., Treccani, M., Corsi, F., Marmi, S., Lillo, F. (2015). Modelling systemic price cojumps with Hawkes factor models. *Quantitative Finance*, *15*(7), 1137–1156, https://doi.org/10.1080/14697688.2014.996586

Bowsher, C.G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, *141*(2), 876–912, https://doi.org/10.1016/j.jeconom.2006.11.007

Campi, L., & Zabaljauregui, D. (2020). Optimal market making under partial information with general intensities. *Applied Mathematical Finance*, *27*(1-2), 1–45, https://doi.org/10.1080/1350486X.2020.1758587

Cartea, Á., & Jaimungal, S. (2016). Incorporating order-flow into optimal execution. *Mathematics and Financial Economics*, *10*(3), 339–364, https://doi.org/10.1007/s11579-016-0162-z

Cartea, Á., Jaimungal, S., Penalva, J. (2015). *Algorithmic and high-frequency trading*. Cambridge, UK: Cambridge University Press.

Cartea, Á., & Wang, Y. (2020). Market making with alpha signals. *International Journal of Theoretical and Applied Finance*, *23*(3), 2050016, https://doi.org/10.1142/S0219024920500168

Chakraborti, A., Muni Toke, I., Patriarca, M., Abergel, F. (2011). Econophysics review: I. empirical facts. *Quantitative Finance*, *11*(7), 991–1012, https://doi.org/10.1080/14697688.2010.539248

Cheridito, P., Dupret, J.-L., Wu, Z. (2025). Abides-marl: A multi-agent reinforcement learning environment for endogenous price formation and execution in a limit order book. *arXiv preprint arXiv:2511.02016*, , https://doi.org/10.48550/arXiv.2511.02016

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1724–1734).

Cohen, K.J., Maier, S.F., Schwartz, R.A., Whitcomb, D.K. (1981). Transaction costs, order placement strategy, and existence of the bid-ask spread. *Journal of Political Economy*, *89*(2), 287–305, https://doi.org/10.1086/260966

Copeland, T.E., & Galai, D. (1983). Information effects on the bid-ask spread. *Journal of Finance*, *38*(5), 1457–1469, https://doi.org/10.1111/j.1540-6261.1983.tb03834.x

Daley, D.J., & Vere-Jones, D. (2003). *An introduction to the theory of point processes* (2nd ed.). Springer.

Demsetz, H. (1968). The cost of transacting. *Quarterly Journal of Economics*, *82*(1), 33–53, https://doi.org/10.2307/1882244

Drissi, F. (2022). Solvability of differential Riccati equations and applications to algorithmic trading with signals. *Applied Mathematical Finance*, *29*(6), 457–493, https://doi.org/10.1080/1350486X.2023.2241130

Filimonov, V., & Sornette, D. (2015). Apparent criticality and calibration issues in the Hawkes self-excited point process model: application to high-frequency financial data. *Quantitative Finance*, *15*(8), 1293–1314, https://doi.org/10.1080/14697688.2015.1032544

Garman, M.B. (1976). Market microstructure. *Journal of Financial Economics*, *3*(3), 257–275, https://doi.org/10.1016/0304-405X(76)90006-4

Gašperov, B., Begušić, S., Posedel Šimović, P., Kostanjčar, Z. (2021). Reinforcement learning approaches to optimal market making. *Mathematics*, *9*(21), 2689, https://doi.org/10.3390/math9212689

Gennaioli, N., & Shleifer, A. (2018). *A crisis of beliefs: Investor psychology and financial fragility.* Princeton, NJ: Princeton University Press.

Glosten, L.R., & Milgrom, P.R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*(1), 71–100, https://doi.org/10.1016/0304-405X(85)90044-3

Gould, M.D., Porter, M.A., Williams, S., McDonald, M., Fenn, D.J., Howison, S.D. (2013). Limit order books. *Quantitative Finance*, *13*(11), 1709–1742, https://doi.org/10.1080/14697688.2013.803148

Hawkes, A.G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, *58*(1), 83–90, https://doi.org/10.1093/biomet/58.1.83

He, X.-Z., & Lin, S. (2022). Reinforcement learning equilibrium in limit order markets. *Journal of Economic Dynamics and Control*, *144*, 104497, https://doi.org/10.1016/j.jedc.2022.104497

Ho, T.S.Y., & Stoll, H.R. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, *9*(1), 47–73, https://doi.org/10.1016/0304-405X(81)90020-9

Ho, T.S.Y., & Stoll, H.R. (1983). The dynamics of dealer markets under competition. *Journal of Finance*, *38*(4), 1053–1074, https://doi.org/10.1111/j.1540-6261.1983.tb02282.x

Jain, K., Firoozye, N., Kochems, J., Treleaven, P. (2024, November). Limit order book dynamics and order size modelling using compound hawkes process. *Finance Research Letters*, *69*(Part A), 106157, https://doi.org/10.1016/j.frl.2024.106157

Lehalle, C.-A., & Neuman, E. (2019). Incorporating signals into optimal trading. *Finance and Stochastics*, *23*(2), 275–311, https://doi.org/10.1007/s00780-019-00382-7

Liang, E., Liaw, R., Nishihara, R., Moritz, P., Fox, R., Goldberg, K., ... Stoica, I. (2018). Rllib: Abstractions for distributed reinforcement learning. *Proceedings of the 35th international conference on machine learning* (pp. 3053–3062). Retrieved from https://proceedings.mlr.press/v80/liang18b.html

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems 30 (neurips 2017)* (pp. 6379–6390).

Mildenstein, E., & Schleef, H.J. (1983). The optimal pricing policy of a monopolistic marketmaker in the equity market. *Journal of Finance*, *38*(1), 218–231, https://doi.org/10.1111/j.1540-6261.1983.tb03637.x

Neuman, E., & Voß, M. (2022). Optimal signal-adaptive trading with temporary and transient price-impact. *SIAM Journal on Financial Mathematics*, *13*(2), 551–575, https://doi.org/10.1137/20M1375486

Ogata, Y. (1981). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, *27*(1), 23–31, https://doi.org/10.1109/TIT.1981.1056305

Paddrik, M., Hayes, R., Todd, A., Yang, S.Y., Beling, P.A., Scherer, W.T. (2012, March). An agent based model of the E-Mini S&P 500 applied to flash crash analysis. *2012 ieee conference on computational intelligence for financial engineering & economics (cifer 2012)* (pp. 257–264). New York City, NY, USA.

Rambaldi, M., Bacry, E., Lillo, F. (2017). The role of volume in order book dynamics: a multivariate hawkes process analysis. *Quantitative Finance*, *17*(7), 999–1020,

https://doi.org/10.1080/14697688.2016.1260759

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, *1707.06347*, arXiv:1707.06347, https://doi.org/10.48550/arXiv.1707.06347

Shleifer, A. (2000). *Inefficient markets: An introduction to behavioral finance*. Oxford: Oxford University Press.

Stoll, H.R. (1978). The supply of dealer services in securities markets. *Journal of Finance*, *33*(4), 1133–1151, https://doi.org/10.1111/j.1540-6261.1978.tb02053.x

Sutton, R.S., & Barto, A.G. (2018). *Reinforcement learning: An introduction* (2nd ed.). The MIT Press.

Szepesvári, C. (2010). *Algorithms for reinforcement learning*. Morgan & Claypool Publishers.

Terry, J.K., Black, B.J., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., ... Ravi, P. (2021). Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in neural information processing systems 34 (datasets and benchmarks track)*. Retrieved from https://openreview.net/forum?id=fLnsj7fpbPI

Tinic, S.M. (1972). The economics of liquidity services. *Quarterly Journal of Economics*, *86*(1), 79–93, https://doi.org/10.2307/1880494

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A.M., Wu, Y. (2022). The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems 35* (pp. 24611–24624). Retrieved from https://openreview.net/forum?id=YVXaxB6L2Pl