

ABIDES-MARL: A Multi-Agent Reinforcement Learning Environment for Endogenous Price Formation and Execution in a Limit Order Book

Patrick Cheridito[†], Jean-Loup Dupret[†] and Zhexin Wu^{*‡}

[†]Department of Mathematics, ETH Zurich, Switzerland

[‡]Department of Management, Technology, and Economics, ETH Zurich, Switzerland

(v1.0 released November 2025)

We present **ABIDES-MARL**, a framework that combines a new multi-agent reinforcement learning (MARL) **methodology** with a new realistic limit-order-book (LOB) **simulation system** to study equilibrium behavior in complex financial market games. The system extends ABIDES-Gym by decoupling state collection from kernel interruption, enabling synchronized learning and decision-making for multiple adaptive agents while maintaining compatibility with standard RL libraries. It preserves key market features such as price-time priority and discrete tick sizes. Methodologically, we use MARL to approximate equilibrium-like behavior in multi-period trading games with a **finite** number of **heterogeneous** agents—an informed trader, a liquidity trader, noise traders, and competing market makers—all with **individual price impacts**. This setting bridges optimal execution and market microstructure by embedding the liquidity trader’s optimization problem within a strategic trading environment. We validate the approach by solving an extended Kyle model within the simulation system, recovering the gradual price discovery phenomenon. We then extend the analysis to a liquidity trader’s problem where market liquidity arises endogenously and show that, at equilibrium, execution strategies shape market-maker behavior and price dynamics. **ABIDES-MARL** provides a reproducible foundation for analyzing equilibrium and strategic adaptation in realistic markets and contributes toward building economically interpretable agentic AI systems for finance.

Keywords: Multi-agent reinforcement learning; Market microstructure; Optimal execution; Endogenous price formation

1. Introduction

Market microstructure theory studies how prices, liquidity, and volatility arise from the interaction of heterogeneous traders. A central objective is to understand what determines market efficiency—how the price of a risky asset is formed and whether it reflects its fundamental value. More precisely, the question is how and at what rate informed traders, such as fundamental investors analyzing company data, release private information into market prices through trading.

Classical models provide analytical insight into this process under asymmetric information. The Kyle model describes how private signals are gradually incorporated into prices by an informed trader interacting with market makers and noise traders (Kyle 1985). Related works explain liquidity provision and bid-ask spreads under inventory and adverse-selection effects (Glosten and Milgrom 1985, Ho and Stoll 1981), competition among informed traders (Holden and Subrahmanyam 1992, Foster and Viswanathan 1993), heterogeneous signals and endogenous timing (Foster and

*Corresponding author: zhexinwu0517@gmail.com

Viswanathan 1996), imperfect competition (Kyle 1989), and recursive linear equilibria under no-manipulation constraints (Huberman and Stanzl 2004). These models form the theoretical basis for price discovery but depend on simplifying assumptions such as perfect competition and Gaussian uncertainty, which limit empirical realism and simulation.

In parallel, optimal execution models developed to address practical aspects of large order trading. Early discrete-time formulations (Bertsimas and Lo 1998) and the Almgren–Chriss mean–variance model (Almgren and Chriss 2001) formalize the trade-off between execution cost and price risk. Later work introduced time-varying impacts (Huberman and Stanzl 2005) and continuous-time control frameworks (Álvaro Cartea *et al.* 2015). More recent studies model transient impact using propagators and limit-order-book (LOB) resilience, ensuring consistency with no-dynamic-arbitrage conditions (Gatheral 2010, Obizhaeva and Wang 2013, Curato *et al.* 2017). Within this stream, Dupret and Hainaut (2025) study optimal liquidation under indirect price impact with propagators and derive a tractable Hamilton–Jacobi–Bellman (HJB) formulation that links high-frequency microstructure with execution models.

Reinforcement learning (RL) provides a data-driven approach to sequential decision-making under uncertainty (Watkins and Dayan 1992, Williams 1992). Deep RL scales this framework to high-dimensional spaces using neural approximations, including off-policy methods such as DQN and DDQN (Hasselt *et al.* 2016) and on-policy actor–critic methods (Mnih *et al.* 2016). Proximal Policy Optimization (PPO) is widely adopted for its training stability (Schulman *et al.* 2017), and deterministic policy gradients extend RL to continuous control settings (Lillicrap *et al.* 2019). In trading applications, differentiable RL methods have shown that combining known model dynamics with policy optimization improves interpretability and stability compared to generic actor–critic training (Jaisson 2022).

Applications of RL to optimal execution and market making typically consider a single adaptive agent trading against an exogenous or partially exogenous price process. Classical work uses tabular Q-learning (Nevmyvaka *et al.* 2006) and its extensions (Hendricks and Wilcox 2014). More recent studies apply deep RL algorithms such as DQN, DDQN, and PPO (Ning *et al.* 2018, Karpe *et al.* 2020, Nagy *et al.* 2023, Hafsi and Vittori 2024, Lin and Beling 2020, 2021, Ganesh *et al.* 2019). Moallemi and Wang (2022) formulate execution timing as an optimal stopping problem and use reinforcement and supervised learning methods to achieve cost reductions on historical data. While these single-agent approaches are effective for policy learning, they do not model strategic interaction among adaptive participants and therefore cannot reproduce endogenous price formation.

Recent machine learning studies have begun to explore the connection between learning and market microstructure. Friedrich and Teichmann (2020) apply adversarial training with deep neural networks to recover the linear equilibrium of Kyle’s single-period model. This work shows how equilibrium can be learned from data but is restricted to static settings. In contrast, our framework extends to a multi-period Kyle model, where multiple agents interact strategically and adapt their policies through learning. This approach allows equilibrium behavior to emerge dynamically from interaction rather than being imposed exogenously.

Multi-agent reinforcement learning (MARL) provides a natural extension of RL to environments with strategic interaction. Independent learners, such as Independent PPO (IPPO), are simple but lead to non-stationary training. Centralized training with decentralized execution (CTDE)—used in algorithms such as MADDPG, MAPPO, and HAPPO—stabilizes learning with centralized critics while maintaining decentralized execution policies (Lowe *et al.* 2020, Yu *et al.* 2022, Kuba *et al.* 2022). In market applications, assuming full observability of opponents’ actions is unrealistic; instead, information is naturally aggregated through the LOB, which serves as a shared market state.

Agent-based simulation environments make MARL practical by providing a realistic trading context compatible with standard RL libraries. ABIDES is an event-driven simulator that models exchange mechanisms such as tick size, price–time priority, order matching, agent registry, and kernel scheduling (Byrd *et al.* 2020). RL libraries like RLlib (Liang *et al.* 2018) and Stable-Baselines3

(Raffin *et al.* 2021) require standardized interfaces, typically implemented through Gymnasium (Towers *et al.* 2024) or PettingZoo (Terry *et al.* 2021). ABIDES-Gym (Amrouni *et al.* 2021) adapts ABIDES to a step-based framework for RL training by introducing an interruptible kernel controlled by a single agent class, the `FinancialGymAgent`. However, because only one such agent can interrupt the kernel, ABIDES-Gym cannot support multiple RL agents. It also removes key agent types—such as zero-intelligence and heuristic belief-learning agents—from the original ABIDES design (Wang *et al.* 2021), limiting its ability to capture endogenous interaction.

In this paper, we introduce **ABIDES-MARL**, which makes contributions at both the system and methodological levels.

First, we present a full-fledged limit-order-book (LOB) simulation system that supports multiple learning agents who trade and adapt their strategies via multi-agent reinforcement learning (MARL). The framework refactors the interruptible-kernel architecture of ABIDES to separate kernel interruption from state collection and communication. A new `StopSignalAgent` coordinates all RL agents and synchronizes their actions, enabling either sequential or simultaneous execution protocols. This design allows multiple adaptive agents to interact through the LOB while remaining fully compatible with major RL training libraries.

Second, we propose a new methodology that uses MARL to approximate equilibrium-like behavior in realistic market settings with **heterogeneous** agents who **individually influence prices**. Our focus is on a **finite** population of agents and their **individual strategic behavior**, in contrast to classical stochastic-control formulations—where the market is represented by an exogenous price process—or mean-field game approaches, which assume agent homogeneity and infinitesimal impact (Lasry and Lions 2007, Huang *et al.* 2006). We embed optimal execution dynamics within a game-theoretic market-microstructure environment by constructing a multi-period trading game inside ABIDES-MARL that includes an informed trader, a liquidity trader, and a group of competing market makers. The equilibrium-like behavior of these agents is approximated by independently trained PPO policies. To our knowledge, this is the first framework to combine MARL and market microstructure in this way.

We validate the approach by first recovering the foundational Kyle model (Kyle 1985) as a special case (by excluding the liquidity trader) and showing empirically that the system reproduces gradual price discovery. Our framework relaxes restrictive analytical assumptions—such as perfect competition among market makers, linear policy parameterization, and continuous prices—thus enabling more realistic models that remain computationally tractable even when analytical equilibria are unavailable.

Beyond validation, we demonstrate that the trained environment serves as a dynamic, endogenous market in which new agents can be introduced and optimized against equilibrium-like counterparts. We illustrate this by solving the liquidity trader’s optimal acquisition problem when trading against learned market participants. Unlike classical execution models where market liquidity is exogenous, here it arises endogenously from strategic interactions between the trading and liquidity-provision sides. We compare three execution approaches: (i) a PPO policy trained in the full game, (ii) an analytical schedule based on a time-varying Kyle-style impact process, and (iii) a post-optimized PPO policy trained against fixed counterparts. This unifies microstructure-based price formation and optimal execution research (Jaisson 2022, Moallemi and Wang 2022, Li *et al.* 2024, Dupret and Hainaut 2025) within a single MARL framework. Recent findings that early price trajectories improve impact-parameter estimation efficiency (Li *et al.* 2024) further underscore the value of trajectory-based MARL environments for testing execution strategies under endogenous price impact.

Contributions.

- (i) **System:** ABIDES-MARL, a game-theoretic LOB simulation environment that enables heterogeneous groups of agents to learn and adapt strategies via reinforcement learning under controlled information and action protocols.

- (ii) **Methodology:** A MARL-based framework to approximate equilibrium-like behavior in multi-period trading games with a **finite, heterogeneous, and strategically interactive** population of agents, bridging market microstructure and optimal execution.

Organization. Sections 2 and 3 describe the ABIDES-MARL framework, the Kyle-style multi-period trading game, and the formulation of the risk-averse optimal acquisition problem. Section 4 presents empirical evidence of price discovery and compares execution strategies. Section 5 concludes with possible extensions, including richer informed-signal structures, alternative market mechanisms, and integration with agentic AI modules. All code and configurations are publicly released for reproducibility.¹

2. System: ABIDES-MARL

ABIDES-MARL converts the Discrete Event Multi-Agent Simulation (DEMAS) system ABIDES into a timestep-based framework compatible with major RL policy training libraries. It builds on the interruptible kernel approach of ABIDES-Gym but introduces a key innovation: the decoupling of kernel interruption from raw state collection and communication.

In ABIDES-Gym, a special agent type, **FinancialGymAgent**, has the exclusive ability to interrupt the priority-queue-based simulation kernel. Upon interruption, this agent collects market information from the ABIDES state, processes it to compute rewards, checks termination and truncation conditions, and stores key information for later use. The result is the standard RL interface tuple (*observation, reward, termination, truncation, info*), which can be sent directly to a policy training algorithm for rollout collection. This design provides a solid foundation for communication between the simulator and external training pipelines.

However, ABIDES-Gym is difficult to extend to a multi-agent setting. Because each **FinancialGymAgent** interrupts the kernel independently, it is impossible to guarantee a consistent action order across RL agents. This creates two issues: (i) sequential games requiring structured order of play cannot be enforced, and (ii) independent interruptions risk leaking information that should be hidden under the intended game design.

To address this, **ABIDES-MARL** introduces a dedicated **StopSignalAgent**. There is exactly one such agent in the simulation, and it alone can interrupt the kernel. All RL agents are synchronized to this interruption point, after which their actions are executed in an explicitly configured order within the timestep-based system. This ensures both consistent synchronization and strict control over the sequential or simultaneous action structure of the game.

We illustrate the system design in Figure 2.

- (i) The kernel is paused upon receiving a **WakeupMsg** from the **StopSignalAgent**.
- (ii) The next wakeup time is scheduled.
- (iii) Raw observations are collected for all RL agents.
- (iv) For each RL agent, the timestep-based environment processes the raw state via user-defined functions to produce standardized observations, rewards, termination & truncation signals and auxiliary information.
- (v) The RL policy receives the current observation and computes an action.
- (vi) The action is transformed into a valid ABIDES-compatible action using a mapping function.
- (vii) Each RL agent submits the order to the **ExchangeAgent**. Then the control is returned to the ABIDES kernel by invoking its **runner** method, allowing the simulation to continue.

To summarize, ABIDES-MARL separates the responsibilities of simulation interaction and decision-making into two distinct layers:

¹Repository: <https://github.com/10258392511/MARLOptExecution>.

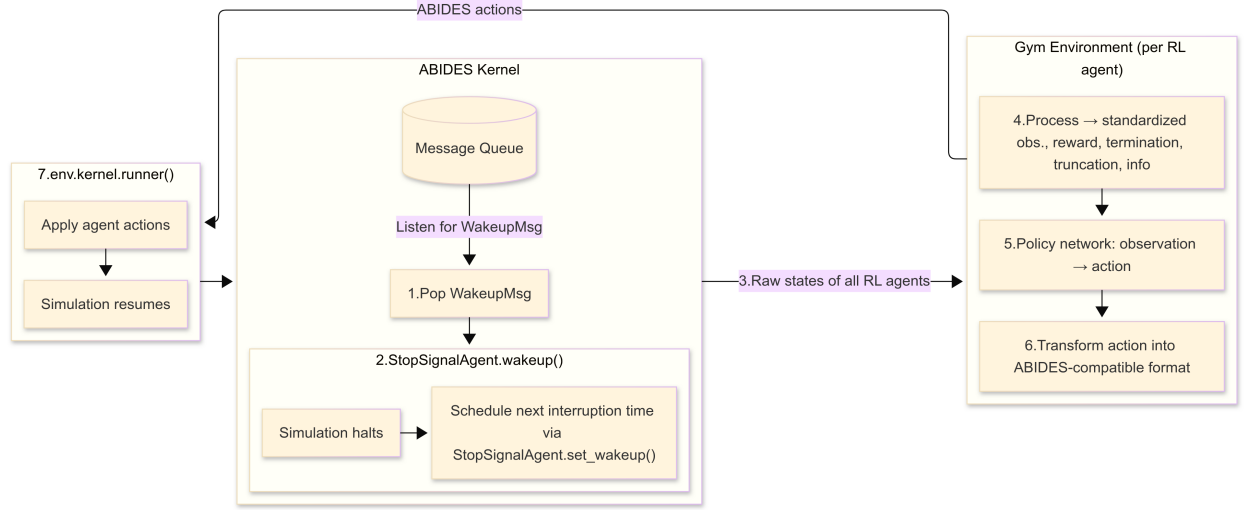


Figure 1. ABIDES-MARL communication cycle.

- **ABIDES agents** act as execution proxies (brokers) embedded in the simulation. They:
 - Observe raw simulation states and convert them into Gym-compatible observations;
 - Receive actions from an external policy and translate them into ABIDES-native order submissions.
- **RL agents**, defined in the timestep-based environment, act as high-level decision-makers (portfolio managers). They maintain policy networks that process observations and generate actions. These agents are referenced by policy ID in the simulation interface and are not explicitly instantiated as ABIDES entities.

Each RL agent is logically paired with an ABIDES agent to form an end-to-end decision-execution loop. This modular separation enables scalable, realistic multi-agent experimentation while preserving compatibility with both Gym-style learning workflows and the event-driven ABIDES kernel.

3. System Validation: Optimal Execution within an Extended Kyle Model

The strategic interaction among heterogeneous agents in a multi-period trading session within **ABIDES-MARL** can be viewed as a sequential game. Each agent repeatedly observes the market state, takes an action, and updates its policy from experience. While one could attempt to derive an analytical equilibrium for this game, doing so typically requires restrictive assumptions that reduce realism. Conversely, obtaining a numerical approximation of the equilibrium is challenging because the learning process of each *individual* agent depends on heterogeneous objectives and limited information.

Our contribution in this section is to show that multi-agent reinforcement learning (MARL) offers a practical approach. MARL allows us to model the adaptive learning behavior of individual agents under minimal structural assumptions, while still recovering key features of the analytical equilibrium in settings where it is known, such as the Kyle model. Beyond these benchmark cases, the same framework can approximate equilibrium behavior in richer market environments that are analytically intractable. To our knowledge, this work is the first to (i) connect foundational microstructure models with optimal execution through a unified simulation framework, (ii) examine how MARL can approximate equilibrium outcomes in heterogeneous-agent trading games, and (iii) relax classical assumptions—such as perfect competition, continuous prices, and linear strategies—while retaining interpretability and equilibrium-like behavior.

Framework overview. We outline the general structure of our MARL framework and demonstrate how it can be used to study multi-period trading games with multiple agent types. In this study, we focus on three learning agent classes: the *informed trader*, the *liquidity trader*, and the *market makers*. Additionally, a representative *noise trader* generates exogenous random orders but does not learn. On the trading side, the informed and liquidity traders submit buy or sell orders to the market. On the liquidity-provision side, a group of market makers compete to quote prices and supply liquidity. Market makers with more competitive quotes transact larger volumes and therefore earn higher profits.

Trading structure and information asymmetry. At each period $n = 1, \dots, N$, the total order flow observed by market makers is given by:

$$q^{(n)} = x^{\text{IT},(n)} + x^{\text{LT},(n)} + u^{(n)}, \quad (1)$$

where:

- $x^{\text{IT},(n)} \in \mathbb{R}$: order submitted by the informed trader,
- $x^{\text{LT},(n)} \in \mathbb{R}$: order submitted by the liquidity trader,
- $u^{(n)} \sim \mathcal{N}(0, \sigma_u^2 \tau)$: random order from the noise trader, iid across periods,
- $q^{(n)} \in \mathbb{R}$: total net order flow available to all market makers.

The trading session proceeds as a hybrid sequential–simultaneous game:

- (i) Traders (informed and liquidity) submit orders to the exchange.
- (ii) Market makers observe the aggregate order flow $q^{(n)}$.
- (iii) Each market maker independently quotes a price.
- (iv) The exchange allocates order flow among market makers according to quote competitiveness (better quotes receive larger allocations).
- (v) The market clears and each agent realizes profit or loss based on its trade.
- (vi) The process repeats at the next period $n + 1$.

This structure yields a hybrid game: within each group (traders or market makers), actions are simultaneous; between the two groups, the interaction is sequential. The resulting dynamics capture key aspects of real-world trading—information asymmetry, competition among liquidity providers, and price formation driven by aggregated order flow—while remaining tractable for MARL-based learning and analysis.

Market makers’ objective. Market makers aim to infer the fundamental value v of the traded asset, which is privately known only to the informed trader. Through trading, the informed trader gradually reveals information about v by submitting orders that influence the observable order flow. Each market maker therefore updates its estimate of v over time based on cumulative market activity.

Formally, let there be M market makers indexed by $i \in \{1, \dots, M\}$. At each trading period n , market maker i quotes a price

$$p_i^{(n)} = \mathbb{E} \left[v \mid q^{(1)}, \dots, q^{(n)} \right], \quad (2)$$

representing its conditional expectation of the asset’s fundamental value given the history of net order flows.

This price-setting behavior can be expressed as an optimization problem. Specifically, each market maker i seeks a quoting function $g_i^{(n)} : \mathcal{B}(\mathbb{R}^n) \rightarrow \mathbb{R}_{\geq 0}$ that minimizes the expected squared deviation

from the true value:

$$\min_{g_i^{(n)} \in \mathcal{B}(\mathbb{R}^n)} \mathbb{E} \left[(g_i(q^{(1)}, \dots, q^{(n)}) - v)^2 \right]. \quad (3)$$

Here, $\mathcal{B}(\mathbb{R}^n)$ denotes the Borel σ -algebra over $\mathbb{R}^n$, representing the measurable space of all possible order-flow histories up to time n . The optimal solution under this quadratic loss corresponds to the conditional expectation above. This formulation highlights that market makers act as Bayesian estimators who continuously refine their beliefs about v through trading interactions with other agents.

To obtain a single transaction price $p^{(n)}$ at each period, individual quotes from all market makers are aggregated using a weighted averaging scheme. This procedure ensures that all traders transact at a common market price while preserving a competitive structure among market makers. The weighting reflects the pro-rata order allocation mechanism—market makers posting more competitive quotes execute a larger share of the total order flow and thus exert greater influence on the average price. The formal definition of this aggregation and its role in the reward computation is provided in Section 3.1.2.

Informed trader's objective. The informed trader possesses private knowledge of the asset's fundamental value v and seeks to exploit this informational advantage to maximize trading profit. Formally, the informed trader solves

$$\max_{(x^{(1)}, \dots, x^{(N)}) \in \mathbb{R}^N} \mathbb{E} \left[\sum_{n=1}^N x^{(n)} (v - p^{(n)}) \right], \quad (4)$$

where $x^{(n)}$ denotes the signed order (positive for buy, negative for sell) submitted at period n , and $p^{(n)}$ is the market-clearing price determined by the aggregation of market-maker quotes. The expected value is taken over the randomness in order flow and other agents' actions. This formulation captures the informed trader's incentive to trade strategically—balancing the immediate gain from exploiting private information against the potential cost of revealing it through trading activity.

Liquidity trader's objective. The liquidity trader aims to acquire a target inventory of size $Q^{(0)} := Q \in \mathbb{R}_{>0}$ within a fixed trading horizon while minimizing the total execution cost. In addition, a risk-averse liquidity trader penalizes excessive intermediate inventory holdings to limit exposure to price fluctuations. Formally, the trader solves

$$\min_{\{x^{(n)} \in \mathbf{M}(H^{(n)})\}_{n=1}^N} \mathbb{E} \left[\sum_{n=1}^N p^{(n)} x^{(n)} + \phi \sum_{n=1}^N (Q^{(n)})^2 \right], \quad (5)$$

subject to the terminal constraint

$$\sum_{n=1}^N x^{(n)} = Q, \quad (6)$$

where $Q^{(n)}$ denotes the remaining quantity to acquire at period n , and $\phi > 0$ controls the trader's degree of risk aversion. The notation $\mathbf{M}(H^{(n)})$ indicates that admissible controls $x^{(n)}$ are measurable with respect to the trader's information set $H^{(n)}$, capturing that actions are conditioned only on information available up to period n .

Approximation to the equilibrium. An analytical solution to the equilibrium of the game described above is difficult to obtain. Unlike the classical Kyle model (Kyle 1985) and its exten-

sions—where agents make decisions based solely on historical information—the liquidity trader’s optimal planning under the dynamic programming principle depends on expectations of future market liquidity. In traditional optimal execution models, market liquidity over the entire trading horizon is exogenously specified, meaning the full liquidity path is known to the trader at the outset. In contrast, in our setting, market liquidity is endogenously determined and evolves dynamically through the actions of market makers. One could assume that the liquidity trader forms estimates of future liquidity from past observations, but there is no canonical or theoretically justified assumption for this relationship. Consequently, our MARL framework offers a practical alternative: it provides an agent-based simulation approach in which equilibrium-like behavior emerges endogenously through repeated learning and strategic interaction among agents.

We now show that both the Kyle model (Kyle 1985) and the optimal execution framework (Almgren and Chriss 2001, Huberman and Stanzl 2005) can be interpreted as special cases of the multi-agent trading game described above. This perspective highlights their shared structure in terms of strategic interaction and information flow. Building on this connection, we then relax several restrictive assumptions—such as perfect competition, exogenous price impact, and linear policies—within a MARL formulation that enables heterogeneous agents to learn equilibrium behavior directly through simulation.

3.1. *Kyle model extended*

By removing the liquidity trader from the game—leaving only the informed trader and the market makers—the setting reduces to the classical Kyle model (Kyle 1985). We first review the key structure of this model and then describe how it can be extended for a MARL formulation.

3.1.1. Model review. The original Kyle model features two strategic participants: an informed trader and a representative market maker. The informed trader seeks to exploit her private knowledge of the asset’s fundamental value v and optimizes the same objective as in Equation 4.

A central assumption of the Kyle model is *perfect competition* among market makers. From a simulation perspective, this corresponds to an infinite number of market makers who instantaneously and repeatedly adjust their quotes each period to eliminate expected profit opportunities. The entire market-making side is therefore summarized by a single representative market maker who sets the price as the conditional expectation:

$$p^{(n)} = \mathbb{E} \left[v \mid q^{(1)}, \dots, q^{(n)} \right],$$

where $q^{(n)}$ denotes the total order flow in period n .

Equilibrium characterization. Under the Kyle equilibrium (Kyle 1985), both the informed trader’s order and the price adjustment rule take linear recursive forms:

$$x^{(n)} = \beta^{(n)} \left(v - p^{(n-1)} \right), \tag{7}$$

$$p^{(n)} - p^{(n-1)} = \lambda^{(n)} q^{(n)}, \tag{8}$$

where $\beta^{(n)}$ represents the insider’s trading intensity and $\lambda^{(n)}$ denotes the price impact coefficient at time step n .

The sequences $\{\beta^{(n)}\}_{n=1}^N$ and $\{\lambda^{(n)}\}_{n=1}^N$ are jointly determined by a system of coupled difference equations, as derived in the original paper. The conditional variance of the pricing error, $\Sigma^{(n)} := \text{Var}(v \mid q^{(1)}, \dots, q^{(n)})$, decreases monotonically over time, reflecting the progressive incorporation of private information into prices.

In equilibrium, the triplet $\{\beta^{(n)}, \lambda^{(n)}, \Sigma^{(n)}\}_{n=1}^N$ characterizes the dynamics of price discovery: smaller $\lambda^{(n)}$ values correspond to deeper and more liquid markets. A full recursive characterization of the equilibrium is provided in Appendix 5.

3.1.2. MARL formulation. The classical Kyle model provides a tractable analytical description of price formation under asymmetric information but remains highly stylized. In particular, its equilibrium relies on strong assumptions—such as an infinite number of perfectly competitive market makers, linear strategies, and Gaussian noise—that limit its realism for simulation and learning-based studies.

To bridge this gap, we formulate the Kyle model within a multi-agent reinforcement learning (MARL) framework. This adaptation serves as a validation for the proposed ABIDES-MARL environment. To our knowledge, this is the first MARL formulation of the Kyle model. It allows agents to learn equilibrium behavior directly from interaction, thereby relaxing the model’s restrictive assumptions on policy linearity, competition, and information flow.

Our formulation introduces several key extensions. First, we replace the assumption of infinitely many market makers with a finite set of competing market makers. Second, we propose a *pro-rata order assignment* mechanism that ensures both (i) a unique, unanimous transaction price across all traders (Claim 3.1), and (ii) a zero-sum game structure among market makers (Lemma 3.2). This structure preserves the essence of competitive quoting without requiring repeated interaction within each trading step. Finally, we redefine the limit order book (LOB) for this discrete, pro-rata setting (Definition 1), while noting that the ABIDES-MARL framework also supports the more general price–time priority structure typical of continuous double auctions. The resulting formulation enables direct comparison with the analytical Kyle equilibrium while allowing richer behavioral dynamics to emerge from learning.

State representations. We distinguish between two scenarios. The first, denoted as **exchange**, assumes that all market participants can observe the full *limit order book (LOB)*. The second, denoted as **OTC**, assumes the LOB is hidden; participants must instead infer its structure through the *volume-weighted average price (VWAP)* derived from their realized PnL.

DEFINITION 1 (LOB) Let $\lambda_i^{(n)}$ denote the market impact coefficient set by market maker $i \in \{1, \dots, M\}$ at time step n . The market depth provided by market maker i is given by $d_i^{(n)} := \frac{1}{|\lambda_i^{(n)}|}$. Let $i_1^{(n)}, \dots, i_M^{(n)}$ be the permutation induced by sorting the prices $\{p_i^{(n)}\}_{i=1}^M$ in ascending order. The limit order book at step n is then defined as

$$\text{LOB}^{(n)} := \begin{bmatrix} d_{i_1^{(n)}}^{(n)} & p_{i_1^{(n)}}^{(n)} \\ \vdots & \vdots \\ d_{i_M^{(n)}}^{(n)} & p_{i_M^{(n)}}^{(n)} \end{bmatrix}. \quad (9)$$

Note that only the sorted depth-price pairs are preserved in the LOB representation; it is not possible to recover the identity of the market maker associated with each quote.

We assume that the market clears at each step, meaning all orders are absorbed by market makers. To compute the VWAP, we must determine the portion of the net order absorbed by each market maker. We assign orders proportionally to market depth: a larger market depth absorbs larger shares. Specifically, let $\text{Order}_i^{(n)}$ denote the volume absorbed by market maker i . Then:

$$\text{Order}_i^{(n)} := q^{(n)} \cdot \frac{d_i^{(n)}}{\sum_{j=1}^M d_j^{(n)}}. \quad (10)$$

Given the allocation rule, we define the VWAP as the average execution price across market makers, weighted by their respective order sizes. Importantly, this value is independent of the total order size:

CLAIM 3.1 (Unanimous VWAP) *Under proportional order assignment, all traders transact at the same VWAP.*

Proof. We verify that the VWAP does not depend on $q^{(n)}$:

$$\begin{aligned}
 \text{VWAP}^{(n)} &=: \bar{p}^{(n)} \\
 &= \frac{\sum_{i=1}^M \text{Order}_i^{(n)} \cdot p_i^{(n)}}{\sum_{i=1}^M \text{Order}_i^{(n)}} \\
 &= \frac{\sum_{i=1}^M q^{(n)} \cdot \frac{d_i^{(n)}}{\sum_{j=1}^M d_j^{(n)}} \cdot p_i^{(n)}}{q^{(n)}} \\
 &= \frac{\sum_{i=1}^M d_i^{(n)} \cdot p_i^{(n)}}{\sum_{i=1}^M d_i^{(n)}}. \tag{11}
 \end{aligned}$$

□

We denote $\bar{p}^{(n)} := \text{VWAP}^{(n)}$ throughout. Furthermore, we can express the VWAP update at round n as follows:

$$\begin{aligned}
 \bar{p}^{(n)} &= \frac{\sum_{i=1}^M p_i^{(n)} d_i^{(n)}}{\sum_{i=1}^M d_i^{(n)}} \\
 &= \frac{\sum_{i=1}^M \left(\bar{p}^{(n-1)} + \lambda_i^{(n)} q^{(n)} \right) d_i^{(n)}}{\sum_{i=1}^M d_i^{(n)}} \\
 &= \bar{p}^{(n-1)} + q^{(n)} \underbrace{\left(\frac{\sum_{i=1}^M \lambda_i^{(n)} d_i^{(n)}}{\sum_{i=1}^M d_i^{(n)}} \right)}_{=: \lambda_{\text{effective}}^{(n)}}, \tag{12}
 \end{aligned}$$

where $d_i^{(n)} = \frac{1}{|\lambda_i^{(n)}|}$ denotes the implied depth associated with market maker i 's quote. The second line follows from expanding each market maker's quote $p_i^{(n)}$ as the previous VWAP $\bar{p}^{(n-1)}$ plus the market impact $\lambda_i^{(n)} q^{(n)}$ resulting from their own pricing rule.

Hence, the effective price impact coefficient at step n becomes:

$$\begin{aligned}
 \lambda_{\text{effective}}^{(n)} &= \frac{\sum_{i=1}^M \lambda_i^{(n)} d_i^{(n)}}{\sum_{i=1}^M d_i^{(n)}} \\
 &= \frac{\sum_{i=1}^M \text{sign}(\lambda_i^{(n)})}{\sum_{i=1}^M \frac{1}{|\lambda_i^{(n)}|}}. \tag{13}
 \end{aligned}$$

This update shows that the VWAP evolves as a linear function of the total order flow $q^{(n)}$, with a slope given by the depth-weighted average of individual market makers' impact coefficients. Conceptually, this update rule behaves similarly to a *harmonic mean* of the market makers' quotes,

consistent with intuition from the Kyle model: more competitive (i.e., tighter) quotes contribute more to price formation.

Agents' observations. Next, we define the agents' observations. Each observation is partitioned into global and individual components. The global component represents market-wide information. Under **exchange**, at timestep n , this is the flattened LOB in addition to VWAP:

$$[\bar{p}^{(n-1)}, d_{i_1^{(n-1)}}^{(n-1)}, p_{i_1^{(n-1)}}^{(n-1)}, \dots, d_{i_M^{(n-1)}}^{(n-1)}, p_{i_M^{(n-1)}}^{(n-1)}],$$

whereas under **OTC** it is simply $[\bar{p}^{(n-1)}]$. The individual observation for the informed trader is $[t^{(n)}, v]$, while for each market maker it is $[t^{(n)}, q^{(n)}]$. We emphasize that at the current timestep n , the global observation reflects only information available at the previous timestep $n - 1$, not the current one. This design ensures that agents cannot infer the actions of their opponents at timestep n , thereby preserving the sequential nature of the game and avoiding information leakage.

Action spaces. We consider both linear and nonlinear policy parameterizations. In the linear setting, policy networks predict the coefficients $\beta^{(n)}$ and $\lambda_i^{(n)}$, following the known analytical structure of the equilibrium. This facilitates direct comparison with the theoretical solution, which is characterized by a difference equation on these coefficients. The predicted coefficients are translated into trading quantities and price quotes as follows:

$$x^{(n)} = \beta^{(n)}(v - \bar{p}^{(n-1)})\tau, \quad (14)$$

$$p_i^{(n)} = \bar{p}^{(n-1)} + \lambda_i^{(n)}(x^{(n)} + u^{(n)}) = \bar{p}^{(n-1)} + \lambda_i^{(n)}q^{(n)}, \quad u^{(n)} \sim \mathcal{N}(0, \sigma_u^2\tau). \quad (15)$$

In contrast, nonlinear policies generalize this structure and directly predict the informed trader's order size $x^{(n)}$ and the individual market maker quote $p_i^{(n)}$. This allows us to assess whether similar convergence to equilibrium arises when the policy class is unconstrained. For evaluation purposes, the implied linear coefficients can be recovered as:

$$\hat{\beta}^{(n)} = \frac{x^{(n)}}{(v - \bar{p}^{(n-1)})\tau}, \quad (16)$$

$$\hat{\lambda}_i^{(n)} = \frac{p_i^{(n)} - \bar{p}^{(n-1)}}{q^{(n)}}. \quad (17)$$

Reward structures. The informed trader receives a reward proportional to her informational advantage at each trading step:

$$R^{\text{IT}} = (v - \bar{p}^{(n)})x^{(n)}. \quad (18)$$

Specifically, if $x^{(n)} > 0$, the informed trader buys from market makers at the average execution price $\bar{p}^{(n)}$ and subsequently sells at the fundamental value v to an external provider. Conversely, if $x^{(n)} < 0$, she shorts at $\bar{p}^{(n)}$ and then covers at v .

For market makers, directly incorporating the fundamental value into their reward results in unrealistically immediate convergence of prices. Moreover, modeling perfect competition requires solving a repeated game within each trading round, which is computationally intractable. To overcome these issues, we propose a zero-sum formulation based on pro-rata order assignment. Each market maker is assumed to close their position at the volume-weighted average price (VWAP)

with an external provider, and their reward at period n is defined as

$$R_i^{\text{MM},(n)} = \text{Order}_i^{(n)} \cdot \left(p_i^{(n)} - \bar{p}^{(n)} \right), \quad (19)$$

This reward design ensures zero aggregate profit across market makers, preserves gradual price discovery which we will show in Section 4, and avoids the need for computing explicit repeated-game equilibria at each timestep.

This structure naturally enforces a zero-profit condition among market makers, as shown below.

LEMMA 3.2 (Zero Profit Among Market Makers) *The reward function in Eqn. 19 implies zero aggregate profit across all market makers.*

Proof. The total profit at step n is:

$$\begin{aligned} \sum_{i=1}^M R_i^{\text{MM},(n)} &= \sum_{i=1}^M \text{Order}_i^{(n)} p_i^{(n)} - \left(\sum_{i=1}^M \text{Order}_i^{(n)} \right) \bar{p}^{(n)} \\ &= q^{(n)} \left(\frac{\sum_{i=1}^M \text{Order}_i^{(n)} p_i^{(n)}}{q^{(n)}} - \bar{p}^{(n)} \right) = 0. \end{aligned}$$

□

Crucially, our model avoids solving a repeated game at each trading step by relaxing the assumption of perfect competition. This introduces a weaker form of competition for computational tractability, while retaining the essential game-theoretic structure of the sequential Kyle framework.

3.2. Optimal execution

We now turn to the optimal execution problem by removing the informed trader from the original game. In this setting, the focus shifts from price discovery to cost-efficient execution.

We begin with the classical formulation in which a single liquidity trader executes against a price oracle. In this benchmark case, market liquidity is treated as exogenous. The problem can be solved analytically using the dynamic programming principle, and we show that it can also be expressed and solved within an RL framework.

Next, we reintroduce the market makers so that market liquidity becomes endogenous. This extension leads naturally to a multi-agent reinforcement learning (MARL) formulation. However, the resulting game is strategically asymmetric: without knowledge of the fundamental value and facing a large execution target, the liquidity trader has no informational advantage over the market makers. As we will demonstrate empirically in Section 4.2, market makers rapidly adjust their quotes toward boundary prices, making it difficult for the liquidity trader to achieve lower transaction costs.

3.2.1. Liquidity trader’s optimization problem with exogenous time-varying market impacts. A liquidity trader is an agent tasked with executing a large inventory within a short time horizon, typically to meet portfolio or client requirements. Without loss of generality, we consider an acquisition task, where the trader must buy the required quantity before the deadline. We formulate a stochastic control problem for the liquidity trader under exogenous time-varying market impact, serving as an intermediate step toward empirically analyzing the equilibrium market impacts that emerge from multi-agent interactions.

In contrast to the Almgren–Chriss model (Almgren and Chriss 2001), which assumes constant market impact parameters, our formulation allows $\lambda^{(n)}$ to vary over time, consistent with the

dynamic microstructure of real markets. Risk aversion is modeled through a running penalty on inventory. Unlike Huberman and Stanzl’s model (Huberman and Stanzl 2005), which uses the variance of total execution cost as a risk measure (thus complicating RL reward design), our approach yields a tractable formulation suitable for reinforcement learning.

Objective with inventory risk aversion. The liquidity trader aims to acquire a target inventory Q over N trading periods by minimizing a cost-plus-risk objective:

$$L(Q, N) := \min_{\{x^{(n)} \in \mathbf{M}(H^{(n)})\}_{n=1}^N} \mathbb{E} \left[\sum_{n=1}^N p^{(n)} x^{(n)} + \phi \sum_{n=1}^N Q^{(n)^2} \right] \quad (20)$$

$$\text{s.t.} \quad \sum_{n=1}^N x^{(n)} = Q,$$

where:

$$\hat{p}^{(n)} = \alpha \hat{p}^{(n-1)} + (1 - \alpha) p^{(n-1)} + \epsilon^{(n)}, \quad (21)$$

$$p^{(n)} = \hat{p}^{(n)} + \lambda^{(n)} (x^{(n)} + u^{(n)}), \quad (22)$$

$$H^{(n)} = \left(\{\hat{p}^{(j)}\}_{j=1}^n, \{p^{(j)}\}_{j=1}^{n-1}, \{x^{(j)}\}_{j=1}^n, \{u^{(j)}\}_{j=1}^{n-1}, \{\epsilon^{(j)}\}_{j=1}^n \right), \quad (23)$$

and $\phi \geq 0$ controls the degree of inventory risk aversion. $M(H^{(n)})$ denotes that the admissible controls $x^{(n)}$ are required to be measurable with respect to the trader’s information set $H^{(n)}$.

Model structure. We focus on the *acquisition* problem (without loss of generality). Key model components are:

- $\hat{p}^{(n)}$: Initial price before execution at time n , evolving with a news component $\epsilon^{(n)} \sim \mathcal{N}(0, \sigma_\epsilon^2)$.
- $p^{(n)}$: Execution price set by a competitive market maker at time n , incorporating temporary impact.
- $x^{(n)}$: Trader’s action (acquisition size) at time n ; the remaining inventory to acquire is $Q^{(n)} = Q^{(n-1)} - x^{(n)}$.
- $u^{(n)}$: $u^{(n)} \sim \mathcal{N}(0, \sigma_u^2)$ which is iid across timesteps. The stepsize τ is absorbed in σ_u^2 . Aggregate order flow from noise traders at time n , independent of $\epsilon^{(n)}$.
- $\lambda^{(n)}$: Time-varying market impact coefficient.
- $\alpha \in [0, 1]$: Weighting between previous initial price and last execution price, modeling permanent impact.

The trader does not observe $p^{(n)}$ or $u^{(n)}$ at time n , but observes $x^{(n)}$, $\hat{p}^{(n)}$, and $\epsilon^{(n)}$. This captures asymmetric information in execution.

Analytical solution via dynamic programming. We next characterize the optimal strategy through a backward difference equation:

THEOREM 3.3 *Let $\mu^{(n)}$ satisfy the backward recursion:*

$$\mu^{(n)} = \alpha \lambda^{(n-1)} + \lambda^{(n)} + \phi - \frac{\lambda^{(n)^2} (1 + \alpha)^2}{4\mu^{(n+1)}}, \quad \forall n \in \{1, \dots, N-1\},$$

$$\mu^{(N)} = \alpha \lambda^{(N-1)} + \lambda^{(N)} + \phi.$$

If $\mu^{(n)} > 0$ for all n , then the unique and time-consistent optimal strategy is:

$$x^{(n)} = \theta^{(n)} Q^{(n)}, \quad \text{where}$$

$$\theta^{(n)} = 1 - \frac{\lambda^{(n)}(1 + \alpha)}{2\mu^{(n+1)}}, \quad \text{for } n < N, \quad \text{and } \theta^{(N)} = 1.$$

The corresponding minimal cost at step n is given by:

$$L^{(n)}(\tilde{p}^{(n-1)}, Q^{(n-1)}, Q^{(n)}) = \left(\tilde{p}^{(n-1)} - \alpha \lambda^{(n-1)} (Q^{(n-1)} + u^{(n-1)}) \right) Q^{(n)} + \mu^{(n)} Q^{(n)^2},$$

where $Q^{(0)} := Q$, with $Q \in \mathbb{R}_{>0}$ the target inventory size; $\lambda^{(0)} = u^{(0)} := 0$, and $\tilde{p}^{(n-1)} := p^{(n-1)} + \epsilon^{(n)}$.

Remarks.

- The optimal strategy depends only on the current inventory $Q^{(n)}$, not on noise trader flow $u^{(n)}$ or news shocks $\epsilon^{(n)}$, due to their zero-mean independence.
- The existence of a unique, time-consistent solution is guaranteed under the positivity condition on $\mu^{(n)}$, which holds under mild assumptions on $\lambda^{(n)}$ and ϕ .

We next integrate market modeling based on the Kyle framework with the liquidity trader's strategy optimization, embedding both into a unified multi-agent environment. Our goal is to demonstrate how the proposed MARL framework can capture the interaction between heterogeneous groups of agents and the resulting equilibrium behavior. We begin by reformulating the optimization problem in Theorem 3.3 as a reinforcement learning problem. We then define the multi-agent setting by specifying the observation, action, and reward structures for all learning agents in the environment.

3.2.2. Single-agent RL environment with exogenous lambdas.

State representation. The observation at step n is $[t^{(n)}, \tilde{p}^{(n-1)}, Q^{(n)}]$, where $\tilde{p}^{(n-1)}$ is defined in Theorem 3.3 and $Q^{(n)}$ denotes the remaining inventory to acquire. Note that the full information set $H^{(n)}$ is not needed, since in Theorem 3.3 we have shown that the minimal cost at step n depends only on $[t^{(n)}, \tilde{p}^{(n-1)}, Q^{(n)}]$ and the action to take i.e., $x^{(n)}$. Since the optimal strategy does not depend on the news component $\epsilon^{(n)}$, we simplify the price process by omitting it. This reduces $\tilde{p}^{(n-1)}$ to $p^{(n-1)}$, which is the execution price incorporating temporary impact from the previous step, without any current-period news effect.

Action space. At each step n , the agent selects a proportion $\theta^{(n)}$, corresponding to an order size $x^{(n)} = \theta^{(n)} Q^{(n)}$. We do not explicitly enforce $\theta^{(N)} = 1$; instead, we incorporate a terminal penalty on unfilled inventory into the reward structure to encourage full acquisition.

Reward Structure. The liquidity trader receives the following step-wise reward:

$$R^{\text{LT},(n)} = - \left(x^{(n)} p^{(n)} + \phi(Q^{(n)})^2 \right), \quad \forall n \in [N - 1]. \quad (24)$$

Here, $p^{(n)}$ is the transaction price at the current step, incorporating temporary impact. To incentivize exact acquisition of the target inventory, we add a terminal penalty on any remaining

unfilled quantity at the last step N :

$$\begin{aligned} Q_{\text{unfilled}} &= Q^{(N)} - x^{(N)}, \\ R^{\text{LT},(N)} &= - \left(x^{(N)} p^{(N)} + \phi(Q^{(N)})^2 \right) - \beta Q_{\text{unfilled}}^2, \end{aligned} \quad (25)$$

where β is chosen to be substantially larger than ϕ .

3.2.3. Multi-agent formulation with endogenous lambdas. We now reintroduce the market makers from the environment of the Kyle model, as presented in Section 3.1.2. The resulting game involves a liquidity trader and a noise trader on the trading side, and multiple market makers providing liquidity. Unlike the single-agent setting in Section 3.2.2, where prices evolve according to an exogenous process, prices here emerge endogenously through interaction among agents.

The state, action, and reward structures for market makers follow those in Section 3.1.2. The liquidity trader retains the structure from the single-agent setup, with an extended observation that includes limit order book (LOB) information. The environment is summarized below.

State representations. Each agent observes a concatenation of a global and an individual component:

- **Global observation at timestep n :**

$$[\bar{p}^{(n-1)}, d_{i_1}^{(n-1)}, p_{i_1}^{(n-1)}, \dots, d_{i_M}^{(n-1)}, p_{i_M}^{(n-1)}],$$

where $\bar{p}^{(n-1)}$ is the VWAP incorporating temporary market impact consistent with Equation 11. If no LOB information is revealed, the market operates in **OTC** mode and the global observation reduces to $[\bar{p}^{(n-1)}]$. If the full LOB is revealed, the market is of type **exchange**.

- **Liquidity trader:** observes $[t^{(n)}, Q^{(n)}]$, where $Q^{(n)}$ is the remaining unfilled inventory.
- **Market maker i :** observes $[t^{(n)}, q^{(n)}]$, where $q^{(n)}$ is the total net order flow from all traders.

Action spaces.

$$x^{\text{LT},(n)} = \theta^{(n)} Q^{(n)}, \quad (26)$$

$$p_i^{(n)} = \bar{p}^{(n-1)} + \lambda_i^{(n)} (x^{\text{LT},(n)} + u^{(n)}), \quad u^{(n)} \sim \mathcal{N}(0, \sigma_u^2 \tau). \quad (27)$$

The liquidity trader controls the trading intensity parameter $\theta^{(n)}$, while each market maker sets either $\lambda_i^{(n)}$ (for linear policies) or a direct quote $p_i^{(n)}$ (for nonlinear policies). For the liquidity trader, linearity is not assumed since the optimal control from Theorem 3.3 is inherently nonlinear.

Reward structures.

- **Liquidity trader:**

$$\begin{aligned} R^{\text{LT},(n)} &= - \left(x^{\text{LT},(n)} p^{(n)} + \phi(Q^{(n)})^2 \right), \quad n < N, \\ R^{\text{LT},(N)} &= - \left(x^{\text{LT},(N)} p^{(N)} + \phi(Q^{(N)})^2 \right) - \beta (Q^{(N)} - x^{\text{LT},(N)})^2. \end{aligned}$$

- **Market maker i :**

$$R_i^{\text{MM},(n)} = \text{Order}_i^{(n)} (p_i^{(n)} - \bar{p}^{(n)}).$$

This game is strategically asymmetric: the liquidity trader lacks private information and must trade a large inventory against adaptive market makers. As shown in Section 4.2, market makers quickly push quotes toward the extremes, making it difficult for the liquidity trader to reduce execution cost. Consequently, we reintroduce the informed trader to restore informational balance and enable non-trivial equilibrium behavior.

3.3. MARL formulation for the full game

We now consider the full MARL formulation with an informed trader, a liquidity trader, and a group of market makers. This is the full game presented at the beginning of Section 3. The informed trader and market makers retain the definitions from Section 3.1.2, while the liquidity trader’s design follows Section 3.2.2, extended with LOB information. The environment is summarized concisely below.

State representations. Each agent observes a concatenated vector of global and individual information:

- **Global observation at timestep n :**

$$[\bar{p}^{(n-1)}, d_{i_1}^{(n-1)}, p_{i_1}^{(n-1)}, \dots, d_{i_M}^{(n-1)}, p_{i_M}^{(n-1)}].$$

- **Liquidity trader:** $[t^{(n)}, Q^{(n)}]$.
- **Informed trader:** $[t^{(n)}, v]$.
- **Market maker i :** $[t^{(n)}, q^{(n)}]$.

Action spaces.

$$x^{\text{LT},(n)} = \theta^{(n)} Q^{(n)}, \tag{28}$$

$$x^{\text{IT},(n)} = \beta^{(n)} (v - \bar{p}^{(n-1)}) \tau, \tag{29}$$

$$p_i^{(n)} = \bar{p}^{(n-1)} + \lambda_i^{(n)} q^{(n)}, \quad u^{(n)} \sim \mathcal{N}(0, \sigma_u^2 \tau). \tag{30}$$

Agents’ action parameters are:

- Liquidity trader: $\theta^{(n)}$;
- Informed trader: $\beta^{(n)}$ under linear policies, or $x^{\text{IT},(n)}$ under nonlinear policies;
- Market maker i : $\lambda_i^{(n)}$ under linear policies, or $p_i^{(n)}$ under nonlinear policies.

Reward structures.

- **Liquidity trader:**

$$R^{\text{LT},(n)} = - \left(x^{\text{LT},(n)} p^{(n)} + \phi(Q^{(n)})^2 \right), \quad n < N,$$

$$R^{\text{LT},(N)} = - \left(x^{\text{LT},(N)} p^{(N)} + \phi(Q^{(N)})^2 \right) - \beta(Q^{(N)} - x^{\text{LT},(N)})^2.$$

- **Informed trader:**

$$R^{\text{IT}} = (v - \bar{p}^{(n)}) x^{\text{IT},(n)}.$$

- **Market maker i :**

$$R_i^{\text{MM},(n)} = \text{Order}_i^{(n)}(p_i^{(n)} - \bar{p}^{(n)}).$$

Including the informed trader restores a non-trivial equilibrium structure. When the liquidity trader and market makers interact alone, the outcome is degenerate: the market makers dominate price formation, and the liquidity trader cannot improve performance. With the informed trader present, the trader side can strategically counterbalance the market makers, allowing the system to recover meaningful equilibrium-like dynamics.

4. Numerical Results

This section presents both qualitative and quantitative evidence of price discovery and demonstrates how **ABIDES-MARL** supports the study of optimal execution under endogenous price formation.

We begin with the extended Kyle model, which includes an informed trader, a group of market makers, and a representative noise trader. We show empirically that prices converge toward the fundamental value known only to the informed trader, and that the speed of convergence depends on market composition and policy parameterization.

Next, we examine the interaction between a liquidity trader and competing market makers. As expected, the liquidity trader finds it difficult to reduce transaction costs: market makers, accustomed to trading against an agent with a known acquisition task, adjust their quotes upward quickly.

Finally, we consider the full game by reintroducing the informed trader alongside the liquidity trader and market makers. This setup enables a direct comparison of alternative execution strategies within an endogenous market. By evaluating these strategies in the same multi-agent environment, we illustrate how adaptive responses from other agents can materially affect execution outcomes and market impact.

4.1. *Price discovery in the extended Kyle model*

We begin by introducing the quantitative measures used to assess price discovery and stylized facts in the simulated market. Let v denote the fundamental value of the traded asset, known only to the informed trader, and define the pricing error as $e^{(n)} := v - p^{(n)}$.

To quantify the rate of convergence toward the fundamental value, we first estimate the *half-life of decay*. Assuming the pricing error follows an autoregressive process of order one,

$$e^{(n)} = \phi e^{(n-1)} + \epsilon^{(n)},$$

where $\{\epsilon^{(n)}\}_n$ is a white-noise process, the corresponding half-life is

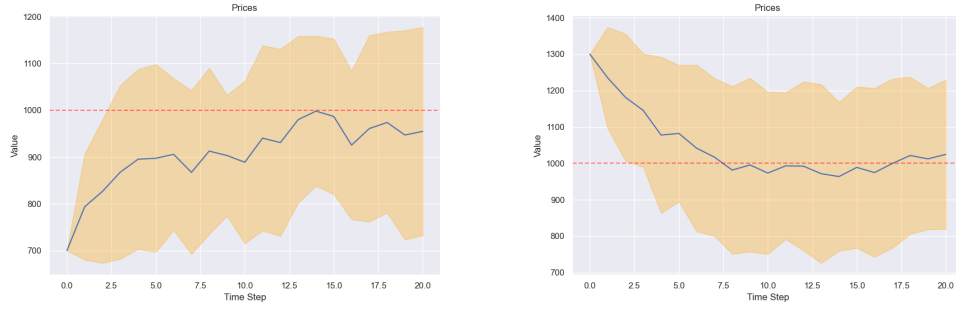
$$n_{1/2} := -\frac{\log(2)}{\log |\phi|}.$$

A shorter half-life indicates faster information incorporation and thus more efficient price discovery.

The second measure follows a Kyle-style regression of price changes on net order flow,

$$\Delta p^{(n)} := p^{(n)} - p^{(n-1)} = \lambda q^{(n)} + \xi^{(n)},$$

where $q^{(n)}$ denotes the net signed order volume and $\xi^{(n)}$ the residual. A significantly positive λ indicates that market makers adjust quotes in the direction implied by order imbalance—raising



(a) Opening price lower than the fundamental value. (b) Opening price higher than the fundamental value.

Figure 2. Linear policy parameterization with 20 market makers and no limit-order-book (LOB) information revealed. Each panel shows the evolution of transaction prices when the opening price was initialized below (a) or above (b) the fundamental value. The fundamental value is marked by the red dashed line. Across evaluation episodes, prices gradually converged toward the fundamental value, indicating that even with limited observability, market participants collectively recovered informational efficiency.

(lowering) prices when net demand is positive (negative)—thereby capturing the informational efficiency of the market.

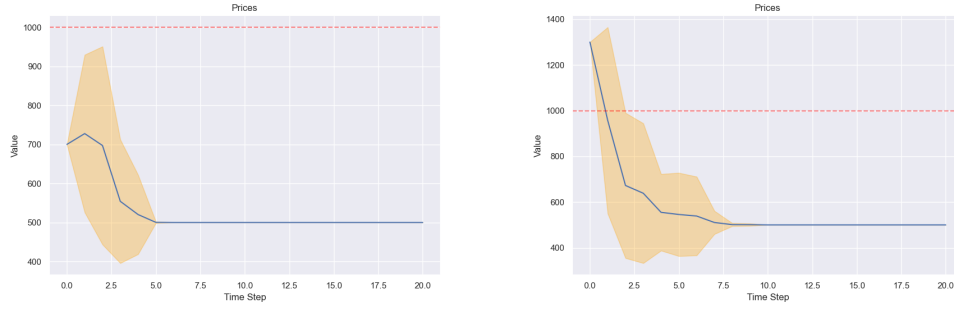
While stylized facts are not the primary focus of this study, we include them to evaluate how closely the endogenous market reproduces key empirical regularities. Specifically, we examine excess kurtosis to detect heavy tails, apply the Anderson–Darling test for normality, and use Engle’s ARCH LM test (Engle 1982) to assess volatility clustering, implemented via `statsmodels` (Seabold and Perktold 2010). Although the present setup—with only three trader archetypes—does not fully reproduce volatility clustering, these diagnostics highlight that realistic statistical properties can progressively emerge as additional agent types and behavioral heterogeneity are introduced. This extension is left for future work.

Experimental configuration. Without loss of generality, all experiments use a symmetric price cap of 50%, with prices bounded by $p_{\min} = 0.5v$ and $p_{\max} = 1.5v$, where v denotes the fundamental value of the traded asset. The fundamental value follows a normal distribution $v \sim \mathcal{N}(\mu_v = 1000, \sigma_v^2 = 100^2)$.

During training, v is independently resampled at the start of each episode, and the opening price is initialized uniformly within the admissible range $[p_{\min}, p_{\max}]$. During evaluation, v is fixed at its mean, and the opening price is set either above or below the mean—specifically at $1.3\mu_v$ (“up”) or $0.7\mu_v$ (“down”)—to test convergence under different initial conditions. Each episode consists of $N = 20$ trading steps. Noise trader order flow is modeled as Gaussian white noise $u_n \sim \mathcal{N}(0, \sigma_u^2)$ with $\sigma_u = 50$.

Reinforcement learning agents are trained using the Proximal Policy Optimization (PPO) algorithm over 1000 episodes and evaluated over 30 holdout episodes. Each policy network is a fully connected architecture with two hidden layers of 64 neurons and `Tanh` activations. Prices are expressed in cents to respect the penny-tick constraint, while order quantities are integer-valued. To maintain numerical stability, rewards are computed in dollars. Reported statistics correspond to averages over the evaluation episodes.

Results. Figures 2 and 3 illustrate the price dynamics under two policy parameterizations—linear and nonlinear—both with 20 market makers and no limit-order-book (LOB) information revealed. Under the linear parameterization, prices converge smoothly toward the fundamental value regardless of whether the opening price is initially above or below it. In contrast, with nonlinear policies, the informed trader dominates the market and drives the price rapidly toward the lower bound (the floor price) irrespective of the initial condition.



(a) Opening price lower than the fundamental value. (b) Opening price higher than the fundamental value.

Figure 3. Nonlinear policy parameterization with 20 market makers and no limit-order-book (LOB) information revealed. Panels (a) and (b) correspond to opening prices initialized below or above the fundamental value, respectively. The fundamental value is marked by the red dashed line. With the more flexible nonlinear policy network, the informed trader dominated the market dynamics, persistently driving transaction prices toward the lower boundary of the admissible price range regardless of the initial condition.

Table 1. Price discovery experiment results. Each configuration reports the estimated AR(1) coefficient (ϕ), corresponding half-life, Kyle regression coefficient (λ), and diagnostics for distributional stylized facts over evaluation episodes.

Act. Type	N	MM	LOB	Mode	ϕ	$p(\phi)$	Half-life	λ	$p(\lambda)$	Kurt.	AD	p-val	ARCH-LM	p-val
linear	20	0	down		0.745	3.18e-88	2.36	2.30	0.000	0.02	4.2e-07		0.275	
linear	20	0	up		0.804	0.000	3.18	2.78	7.05e-80	0.62	1.2e-06		0.028	
linear	20	1	down		0.774	1.23e-249	2.70	2.88	0.000	0.60	7.2e-06		0.240	
linear	20	1	up		0.827	0.000	3.64	2.64	1.66e-41	0.98	4.4e-05		0.505	
linear	2	0	down		0.232	7.09e-09	0.47	5.01	0.000	0.56	2.9e-16		0.679	
linear	2	0	up		0.416	2.36e-31	0.79	5.06	7.72e-124	-0.03	2.2e-08		0.981	
linear	2	1	down		0.503	3.94e-53	1.01	5.13	0.000	1.63	2.0e-12		0.861	
linear	2	1	up		0.625	1.15e-128	1.48	3.66	5.75e-85	3.65	5.6e-30		0.882	
non-linear	20	0	down		0.995	0.000	145.41	-0.14	0.139	12.32	0.000		0.003	
non-linear	20	0	up		0.943	0.000	11.82	0.68	0.317	10.64	0.000		0.068	
non-linear	20	1	down		0.961	0.000	17.59	0.45	0.062	12.85	0.000		0.006	
non-linear	20	1	up		1.009	0.000	-81.16	0.15	0.071	49.34	0.000		0.689	
non-linear	2	0	down		-0.058	0.193	0.24	3.25	0.000	-0.85	0.000		0.241	
non-linear	2	0	up		0.074	0.110	0.27	2.81	0.017	-0.41	0.000		0.048	
non-linear	2	1	down		0.004	0.926	0.12	6.78	0.001	-0.68	0.000		0.373	
non-linear	2	1	up		-0.051	0.289	0.23	14.89	1.98e-49	-0.76	0.000		0.712	

With an unconstrained nonlinear policy, the agent can exploit this structure by first pushing market makers’ beliefs toward the floor price. Once all market makers adopt this belief and no one deviates unilaterally, the informed trader can exploit the situation by executing large buy orders at undervalued prices, generating abnormally high profits.

As a promising direction for future work, introducing multiple informed traders would mitigate this effect by creating strategic competition among them. As in classical extensions of the Kyle model (Foster and Viswanathan 1996), competition among informed traders reduces individual market power, limits aggressive order placement, and leads to more stable price discovery.

Table 1 provides a comprehensive ablation study across three experimental dimensions: (i) policy parameterization (linear versus nonlinear), (ii) the number of market makers (20 or 2), and (iii) the level of LOB information disclosed (0=hidden or 1=full). For each configuration, results are reported under both upward and downward initial price deviations from the fundamental value.

Several key observations emerge.

- **Linear policies.** All estimated autoregressive coefficients (ϕ) are statistically significant at the 5% level, confirming mean reversion to zero of pricing errors. Moreover, the rate of

convergence—as measured by the half-life of decay—is faster when there are fewer market makers, reflecting stronger price adjustment power in less competitive (duopoly) settings. The Kyle regression coefficients (λ) are consistently positive and highly significant, indicating efficient incorporation of order flow information into prices. Excess kurtosis is also more pronounced under linear policies, reflecting occasional large price moves consistent with heavy-tailed return distributions.

- **Nonlinear policies.** In contrast, pricing errors either fail to decay or exhibit statistically insignificant decay rates, implying weak or unstable convergence toward the fundamental value. The Kyle coefficients are generally smaller and less significant, suggesting muted price responsiveness. Excess kurtosis remains high, and in several configurations the price dynamics become dominated by the informed trader rather than market competition.
- **Stylized facts.** All of the return distributions reject the Anderson–Darling test for normality as null hypothesis. However, the Engle’s ARCH–LM test indicates no significant volatility clustering. While this limits the ability of the current minimal setup to reproduce full empirical stylized facts, this outcome is expected given the small number of agent types. Future extensions incorporating additional trader archetypes and heterogeneity in reaction speeds may recover richer volatility dynamics.

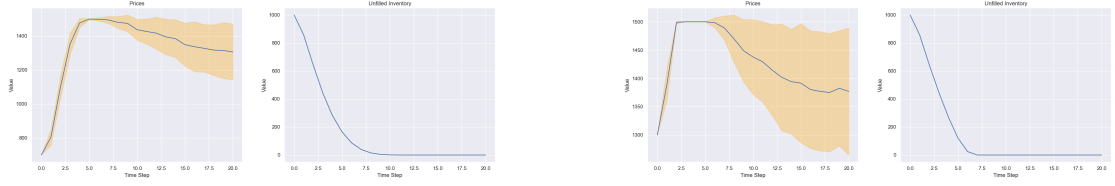
Overall, these results confirm that under realistic structural constraints, **ABIDES-MARL** reproduces key features of price discovery as described in classical microstructure theory. The convergence behavior, sensitivity to market composition, and distinct linear–nonlinear contrasts collectively validate that information is progressively incorporated into prices through agent interaction.

4.2. *Liquidity trader versus market makers: endogenous liquidity formation*

We now consider the setup introduced in Section 3.2.3. Unlike the analytical case in Theorem 3.3, where the market liquidity process is exogenous, here the liquidity trader interacts directly with a group of adaptive market makers whose quoting behavior endogenously determines market liquidity.

Experimental configuration. Episodes consist of $N = 20$ trading steps, and the liquidity trader targets an inventory of $Q = 1000$. The risk-aversion coefficient for the liquidity trader is set to $\phi = 0.01$, and terminal inventory is penalized with $\beta = 10$. The mean-reversion parameter is fixed at $\alpha = 0.0$, maintaining consistency with the Kyle framework. The fundamental value follows $v \sim \mathcal{N}(1000, 100^2)$ during training, while evaluation uses a fixed $v = \mu_v$ with the opening price initialized either at $0.7\mu_v$ (“down”) or $1.3\mu_v$ (“up”). Noise order flow is Gaussian with variance $\sigma_u^2 = 50^2 \times 20/N$, ensuring comparable volatility across horizons. Quotes are bounded within $[0.5v, 1.5v]$ to avoid numerical instability. The liquidity trader’s fractional order size is clipped to $[\theta_{\min}, \theta_{\max}] = [0, 1]$. All agents are trained using PPO with two fully connected hidden layers (64 units, **Tanh** activations). Multi-agent training runs for 1000 episodes, and evaluation statistics are aggregated across 30 holdout episodes.

Results. Figure 4 illustrates the price dynamics and the liquidity trader’s inventory trajectory. The market makers adapt to the trader’s predictable acquisition pattern, leading to adverse price movements regardless of the initial price level. Once the liquidity trader completes the acquisition before the trading horizon ends, the absence of further buy orders introduces divergence among the market makers’ beliefs, increasing uncertainty and widening price fluctuations. Although the average quotes decline during this phase, the effect offers no further cost advantage to the liquidity trader.



(a) Opening price lower than the fundamental value

(b) Opening price higher than the fundamental value

Figure 4. Price and unfilled-inventory dynamics for the liquidity trader trading against 20 market makers under linear policy parameterization, with no limit-order-book (LOB) information revealed. The liquidity trader’s risk aversion is set to $\phi = 0.01$. Panels (a) and (b) correspond to opening prices initialized below and above the fundamental value, respectively. In both cases, market makers rapidly adjust their quotes upward, limiting the trader’s ability to reduce transaction costs. This behavior reflects the market makers’ adaptation to the trader’s predictable acquisition strategy, resulting in consistently adverse price movements.

4.3. Developing execution strategy

Building on the previous experiment, we now reintroduce the informed trader to create a more competitive environment. In this setting, the market comprises one liquidity trader, one informed trader, twenty market makers, and a representative noise trader. The informed trader’s presence restores strategic balance—providing an informational counterweight to the liquidity trader’s predictable behavior and mitigating the market makers’ unilateral control over price dynamics. Within this full multi-agent configuration, we study the liquidity trader’s execution problem under endogenous price formation. Specifically, we consider the risk-averse acquisition objective in Equation 20 and vary the risk-aversion parameter ϕ to examine how execution preferences influence both trading efficiency and the adaptive responses of other participants.

To highlight the flexibility of **ABIDES-MARL** for strategy development, five separate execution policies are evaluated while the informed trader and market makers use the PPO policies obtained from multi-agent training:

- (i) **PPO (multi-agent)**. The liquidity trader deploys the policy learned during the joint MARL training phase (1000 training episodes).
- (ii) **VWAP (trajectory-based)**. Using price–volume trajectories from the PPO evaluation episodes (30 episodes), the trader allocates inventory proportionally to the observed volume share at each step. Formally, letting $V^{(n)}$ denote the average observed trading volume at timestep n and $V_{\text{tot}} = \sum_{n=1}^N V^{(n)}$, the order size is given by

$$x_{\text{VWAP}}^{\text{LT},(n)} = Q \cdot \frac{V^{(n)}}{V_{\text{tot}}} \quad n \in \{1, \dots, N\}.$$

- (iii) **TWAP (benchmark)**. Inventory is split evenly across the trading horizon:

$$x_{\text{TWAP}}^{\text{LT},(n)} = \frac{Q}{N}, \quad n \in \{1, \dots, N\}.$$

- (iv) **Analytical (Kyle- λ)**. A closed-form schedule obtained by solving the risk-averse acquisition problem under a time-varying Kyle-style impact process (Kyle 1985). Concretely, we first compute the sequence of impact coefficients $\{\lambda^{(n)}\}_{n=1}^N$ by solving the system of difference equations that characterize the Kyle equilibrium (Appendix 5). This sequence is then substituted into the difference equation solver derived in Theorem 3.3, yielding the optimal acquisition path $\{x_{\text{Analytical}}^{\text{LT},(n)}\}_{n=1}^N$. The resulting policy is purely analytical and does not incorporate feedback effects from the liquidity trader on other agents, in contrast to the PPO-trained strategies.
- (v) **PPO Single**. Holding the informed trader and market-maker PPO policies fixed (no further

Table 2. Implementation shortfall (IS) comparison under **linear** policy parameterization. Values denote mean $\pm$ standard deviation over 30 evaluation episodes, normalized to the opening price $p^{(0)}$. Lower values indicate better performance. “N/A” in the PPO column denotes cases where the agent did not execute any acquisition during the trading horizon.

Act. Type	LOB	ϕ	Mode	PPO	VWAP	TWAP	Analytical	PPO-Single
linear	0	0	down	N/A	0.698 ± 0.046	0.627 ± 0.046	0.807 ± 0.141	0.597 ± 0.046
linear	0	0	up	N/A	0.008 ± 0.035	-0.065 ± 0.026	0.108 ± 0.020	-0.014 ± 0.026
linear	0	0.01	down	0.876 ± 0.012	0.834 ± 0.021	0.750 ± 0.073	0.775 ± 0.084	0.801 ± 0.079
linear	0	0.01	up	0.149 ± 0.003	0.082 ± 0.012	-0.028 ± 0.038	0.117 ± 0.015	0.055 ± 0.032
linear	0	0.1	down	0.841 ± 0.045	0.811 ± 0.027	0.714 ± 0.066	0.951 ± 0.037	0.703 ± 0.022
linear	0	0.1	up	0.113 ± 0.008	0.072 ± 0.016	-0.007 ± 0.038	0.134 ± 0.008	0.105 ± 0.007
linear	1	0	down	1.014 ± 0.053	0.747 ± 0.059	0.740 ± 0.088	0.784 ± 0.103	0.795 ± 0.130
linear	1	0	up	0.153 ± 0.001	0.112 ± 0.054	0.079 ± 0.059	0.107 ± 0.016	0.111 ± 0.037
linear	1	0.01	down	0.936 ± 0.060	0.775 ± 0.088	0.703 ± 0.100	0.790 ± 0.099	0.799 ± 0.009
linear	1	0.01	up	0.087 ± 0.028	0.046 ± 0.035	0.050 ± 0.030	0.118 ± 0.020	0.144 ± 0.001
linear	1	0.1	down	0.886 ± 0.013	0.767 ± 0.021	0.500 ± 0.019	0.923 ± 0.055	0.548 ± 0.028
linear	1	0.1	up	0.154 ± 0.000	-0.172 ± 0.009	-0.178 ± 0.010	0.135 ± 0.008	-0.052 ± 0.023

Table 3. Implementation shortfall (IS) comparison under **non-linear** policy parameterization. Values denote mean $\pm$ standard deviation over 30 evaluation episodes, normalized to the opening price $p^{(0)}$. Lower values indicate better performance.

Act. Type	LOB	ϕ	Mode	PPO	VWAP	TWAP	Analytical	PPO-Single
non-linear	0	0	down	-0.236 ± 0.061	-0.178 ± 0.020	-0.182 ± 0.024	0.805 ± 0.075	-0.248 ± 0.056
non-linear	0	0	up	-0.152 ± 0.007	-0.142 ± 0.005	-0.141 ± 0.004	0.103 ± 0.029	-0.147 ± 0.006
non-linear	0	0.01	down	-0.286 ± 0.000	-0.258 ± 0.010	-0.255 ± 0.013	0.778 ± 0.089	-0.061 ± 0.003
non-linear	0	0.01	up	0.154 ± 0.001	0.151 ± 0.000	0.152 ± 0.000	0.116 ± 0.017	0.041 ± 0.001
non-linear	0	0.1	down	0.000 ± 0.010	-0.271 ± 0.009	-0.282 ± 0.004	0.916 ± 0.065	-0.200 ± 0.003
non-linear	0	0.1	up	0.051 ± 0.006	-0.413 ± 0.023	-0.432 ± 0.018	0.135 ± 0.008	-0.144 ± 0.010
non-linear	1	0	down	-0.278 ± 0.003	-0.278 ± 0.002	-0.278 ± 0.003	0.783 ± 0.091	-0.286 ± 0.000
non-linear	1	0	up	-0.348 ± 0.121	-0.415 ± 0.072	-0.387 ± 0.082	0.116 ± 0.012	-0.603 ± 0.068
non-linear	1	0.01	down	-0.229 ± 0.016	-0.262 ± 0.008	-0.265 ± 0.003	0.824 ± 0.076	-0.266 ± 0.002
non-linear	1	0.01	up	0.144 ± 0.003	0.143 ± 0.001	0.144 ± 0.001	0.116 ± 0.017	0.150 ± 0.000
non-linear	1	0.1	down	-0.217 ± 0.025	-0.280 ± 0.002	-0.281 ± 0.001	0.925 ± 0.043	-0.168 ± 0.006
non-linear	1	0.1	up	-0.052 ± 0.004	-0.165 ± 0.023	-0.170 ± 0.009	0.134 ± 0.009	0.004 ± 0.007

learning), the environment reduces to a single-agent problem for the liquidity trader, who is retrained with PPO for 500 episodes.

Performance is compared using implementation shortfall,

$$\text{IS} := \left(\frac{\sum_{n=1}^N p^{(n)} q^{(n)}}{Q} - p^{(0)} \right) / p^{(0)},$$

with Q the acquisition target and $p^{(0)}$ the opening price (acquisition case).

Experimental configuration. Episodes have $N = 20$ steps; the liquidity-trader target is $Q = 1000$. Terminal inventory is penalized with $\beta = 10$; the mean-reversion parameter is set to $\alpha = 0$ to align with the original Kyle dynamics. The fundamental value is drawn iid per training episode from $\mathcal{N}(1000, 100^2)$; evaluation fixes $v = \mu_v$ with opening price initialized at $0.7\mu_v$ (“down”) or $1.3\mu_v$ (“up”). Noise order flow is Gaussian with variance scaled as $50^2 \times 20/N$ to keep volatility comparable across horizons. Quotes are clipped to $[0.5v, 1.5v]$ for numerical stability. Actions of the liquidity trader are bounded in $[\theta_{\min}, \theta_{\max}] = [0, 1]$. All agents use PPO with two fully connected hidden layers (64 units, **Tanh**); multi-agent training runs for 1000 episodes, and evaluation statistics aggregate over 30 holdout episodes (PPO-Single retrain the liquidity trader for 500 episodes as noted above).

Results. Tables 2 and 3 compare the five execution strategies using implementation shortfall (IS). Several insights emerge. First, because prices gradually converge toward the fundamental value, the

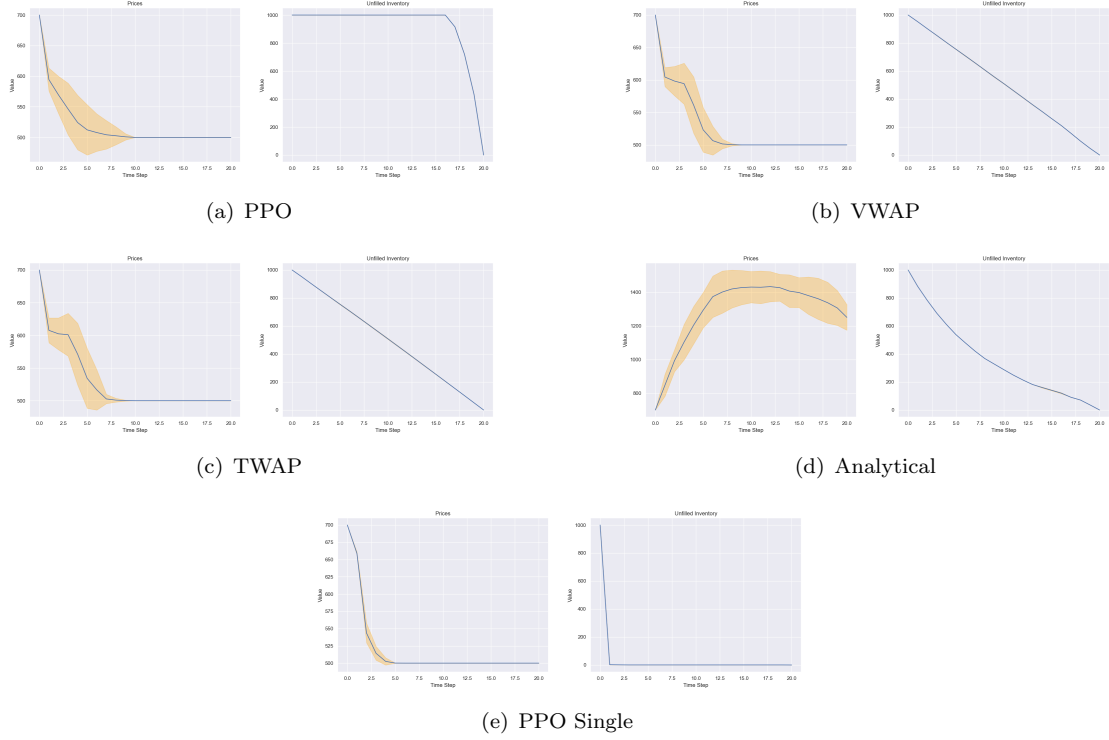


Figure 5. Nonlinear policy parameterization with 20 market makers and no limit-order-book (LOB) information revealed. The liquidity trader’s risk aversion is set to $\phi = 0.01$, and the opening price is initialized below the fundamental value. Each panel reports the price trajectory (left) and unfilled inventory process (right) averaged over 30 evaluation episodes. Among the five strategies, the PPO policy achieves the lowest implementation shortfall by waiting for prices to reach the floor level before initiating acquisition. In contrast, the Analytical strategy performs the worst, as it is the only one that induces an upward price movement during execution.

direction of the initial mispricing determines the liquidity trader’s optimal behavior. Considering the markets with price discovery, when the opening price is below the fundamental value, prices tend to rise, making it advantageous for the trader to execute more quickly—though not so aggressively as to trigger rapid quote revisions by market makers. Conversely, when the opening price is above the fundamental value, it is preferable to delay execution, allowing prices to decline and thereby achieving a lower average acquisition cost (reflected in negative IS).

Second, strategy performance differs substantially across policy parameterizations. Under the **linear** policy setting, the most robust performance is obtained with the simple TWAP schedule. As shown in Section 4.1, the informed trader’s influence is limited under linear parameterization; hence, a non-reactive execution rule such as TWAP—minimizing information leakage—proves most effective. In contrast, under the **non-linear** policy setting, TWAP becomes less robust, while the *PPO-Single* strategy achieves superior results. This suggests that targeted post-optimization of the liquidity trader’s policy, given fixed opponent strategies, yields tangible benefits when agents possess greater expressive power.

Finally, the *PPO (multi-agent)* policy performs poorly relative to alternatives. During training, opposing agents adapt to the liquidity trader’s behavior, effectively learning to exploit its predictable responses. In evaluation, this manifests as adverse interaction dynamics, where market makers and the informed trader optimize their own rewards at the liquidity trader’s expense.

Figures 5 and 6 illustrate representative price and unfilled inventory trajectories under the five execution strategies. A key observation is that, even when employing similar strategies, differences in the opening price lead to distinct market-maker reactions, thereby influencing the effectiveness of the liquidity trader’s acquisition policy. In Figure 5, where the opening price is below the fundamental value, the Analytical strategy performs worst, as it is the only one that fails to push

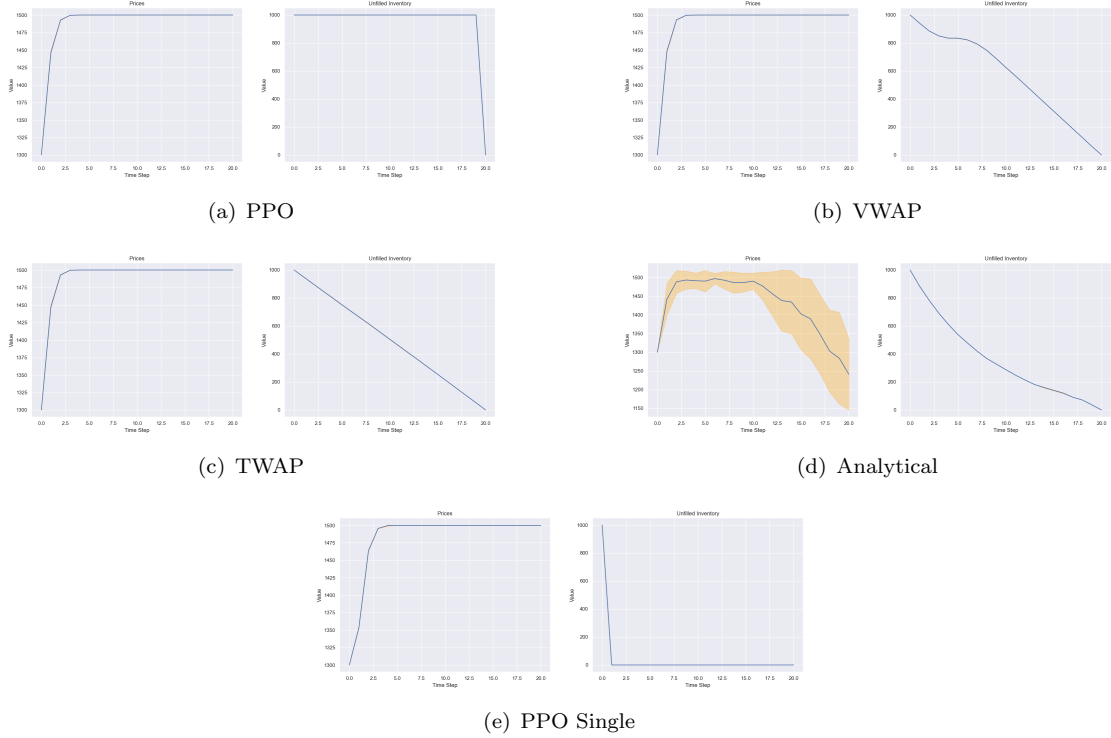


Figure 6. Nonlinear policy parameterization with 20 market makers and no limit-order-book (LOB) information revealed. The liquidity trader’s risk aversion is set to $\phi = 0.01$, and the opening price is initialized above the fundamental value. Each panel reports the price trajectory (left) and unfilled inventory process (right), averaged over 30 evaluation episodes. Among the five strategies, the PPO-Single policy achieves the lowest implementation shortfall, despite not waiting for prices to decline before trading. The Analytical strategy, by inducing downward price adjustments during execution, performs notably better in this setting, ranking as the second-best approach.

the price toward the lower bound. Conversely, in Figure 6, where the opening price is above the fundamental value, the Analytical strategy outperforms PPO, TWAP, and VWAP by successfully inducing downward price adjustment—the only strategy achieving this outcome.

Comparing these results with the nonlinear policy parameterization experiments in Section 4.1 reveals that introducing a liquidity trader fundamentally alters market-maker behavior. Without the liquidity trader, prices consistently drift toward the lower bound regardless of the initial condition. However, once the liquidity trader enters the market, the opening price determines the direction of price adjustment, underscoring how strategic interaction between agents shapes market dynamics.

5. Discussion

This study introduced **ABIDES-MARL**, a multi-agent reinforcement learning (MARL) extension of the ABIDES limit-order-book simulator designed to model endogenous price formation through strategic interaction among heterogeneous agents. By embedding informed traders, market makers, and a liquidity trader within a unified reinforcement learning environment, we demonstrated that classical microstructure phenomena—such as price discovery, information diffusion, and market response to strategic trading—emerge naturally from adaptive behavior rather than being imposed exogenously.

From a quantitative standpoint, our results confirm that under linear policy parameterization, prices converge reliably to the fundamental value with finite half-lives of pricing errors and statistically significant Kyle regression coefficients. The presence of a small number of market makers accelerates convergence, reflecting stronger individual market-making power. In contrast, nonlinear

policy parameterization amplifies the influence of the informed trader and destabilizes convergence, illustrating how greater functional flexibility can shift market equilibrium away from efficiency. These findings reinforce the classical intuition that competition and information asymmetry jointly shape the rate of price discovery.

Introducing a liquidity trader further revealed how execution performance depends on market adaptivity. When the opening price is below the fundamental value, faster execution is preferred, whereas an overvalued market favors patient trading to achieve negative implementation shortfall. Across strategies, the time-weighted average price (TWAP) rule remained the most robust under linear market dynamics, while a retrained single-agent PPO policy outperformed others under non-linear settings. The multi-agent PPO policy underperformed because counterparties had adapted adversarially during training, highlighting that execution costs in adaptive markets are inherently endogenous to the learning dynamics of competitors. These results demonstrate how qualitatively different equilibria depend on learning scope and agent heterogeneity.

Beyond methodological innovation, this work suggests a broader pathway toward integrating MARL-based financial simulations with the emerging class of *agentic AI systems* that combine reasoning and decision modules driven by large language models (LLMs) (Yu *et al.* 2023, Zhang *et al.* 2024, Yang *et al.* 2024, Cao *et al.* 2025). While such systems excel at text-based reasoning, they often lack a quantitatively grounded model of strategic interaction. ABIDES-MARL provides a principled simulation backbone where LLM-based agents could operate as analysts or allocators producing textual signals, while MARL agents act as execution and pricing modules trained under realistic, game-theoretic conditions. This hybrid paradigm opens a research frontier at the intersection of language, learning, and market simulation.

Future extensions will focus on expanding the strategic diversity of the environment. Natural directions include incorporating multiple informed traders with heterogeneous beliefs (Foster and Viswanathan 1996), introducing risk-averse market makers with inventory-based spread adjustments, and extending the model to continuous double-auction mechanisms. Another promising avenue is to integrate historical market replay agents that enable counterfactual training and policy stress-testing. More broadly, embedding MARL as the quantitative core of agentic financial systems may yield a new generation of AI frameworks that combine interpretability, adaptability, and equilibrium-based realism.

In summary, ABIDES-MARL bridges reinforcement learning and market microstructure, offering a foundation for quantitatively rigorous research on strategic adaptation in financial markets and for building the next generation of intelligent, explainable market AI systems.

Supplemental material

Proof of Theorem 3.3

Proof. We proceed by backward induction.

Base Case ($n = N$). At the final period, we evaluate:

$$L^{(N)}(\tilde{p}^{(N-1)}, Q^{(N-1)}, Q^{(N)}) = \mathbb{E}_N \left[p^{(N)} x^{(N)} + \phi Q^{(N)^2} \right].$$

To evaluate this expression in terms of known quantities $(\tilde{p}^{(N-1)}, Q^{(N-1)}, Q^{(N)})$, we substitute for $p^{(N)}$ using the price dynamics equations 21 and 22:

$$\begin{aligned}
p^{(N)} &= \hat{p}^{(N)} + \lambda^{(N)}(x^{(N)} + u^{(N)}) \\
&= \alpha \hat{p}^{(N-1)} + (1 - \alpha)p^{(N-1)} + \epsilon^{(N)} + \lambda^{(N)}(x^{(N)} + u^{(N)}) \\
&= \alpha(\hat{p}^{(N-1)} - p^{(N-1)}) + (p^{(N-1)} + \epsilon^{(N)}) + \lambda^{(N)}(x^{(N)} + u^{(N)}) \\
&= -\alpha\lambda^{(N-1)}(x^{(N-1)} + u^{(N-1)}) + \tilde{p}^{(N-1)} + \lambda^{(N)}(x^{(N)} + u^{(N)}) \\
&= \tilde{p}^{(N-1)} - \alpha\lambda^{(N-1)}(Q^{(N-1)} - Q^{(N)} + u^{(N-1)}) + \lambda^{(N)}(x^{(N)} + u^{(N)}).
\end{aligned}$$

Since $x^{(N)} = Q^{(N)}$, we obtain:

$$p^{(N)}x^{(N)} = \left(\tilde{p}^{(N-1)} - \alpha\lambda^{(N-1)}(Q^{(N-1)} - Q^{(N)} + u^{(N-1)}) + \lambda^{(N)}(Q^{(N)} + u^{(N)}) \right) Q^{(N)}.$$

Taking conditional expectation with respect to $\mathbb{E}_N$, and using the fact that $u^{(N)}$ has zero mean, we obtain:

$$\mathbb{E}_N \left[p^{(N)}x^{(N)} \right] = (\tilde{p}^{(N-1)} - \alpha\lambda^{(N-1)}(Q^{(N-1)} + u^{(N-1)}))Q^{(N)} + (\alpha\lambda^{(N-1)} + \lambda^{(N)})Q^{(N)^2}.$$

Thus, the value function becomes:

$$\begin{aligned}
L^{(N)}(\tilde{p}^{(N-1)}, Q^{(N-1)}, Q^{(N)}) &= (\tilde{p}^{(N-1)} - \alpha\lambda^{(N-1)}(Q^{(N-1)} + u^{(N-1)}))Q^{(N)} \\
&\quad + \underbrace{(\alpha\lambda^{(N-1)} + \lambda^{(N)} + \phi)}_{=:\mu^{(N)}} Q^{(N)^2}.
\end{aligned}$$

Inductive Step. Suppose the result holds for periods $N, N-1, \dots, n+1$, i.e.,

$$L^{(n+1)}(\tilde{p}^{(n)}, Q^{(n)}, Q^{(n+1)}) = (\tilde{p}^{(n)} - \alpha\lambda^{(n)}(Q^{(n)} + u^{(n)}))Q^{(n+1)} + \mu^{(n+1)}Q^{(n+1)^2}. \quad (31)$$

Then, at time n , the Bellman equation becomes:

$$\begin{aligned}
L^{(n)}(\tilde{p}^{(n-1)}, Q^{(n-1)}, Q^{(n)}) &= \inf_{x^{(n)} \in \mathbf{M}(H^{(n)})} \mathbb{E}_n \left[p^{(n)}x^{(n)} + \phi Q^{(n)^2} + L^{(n+1)}(\tilde{p}^{(n)}, Q^{(n)}, Q^{(n+1)}) \right] \\
&= \inf_{x^{(n)} \in \mathbf{M}(H^{(n)})} \mathbb{E}_n \left[p^{(n)}x^{(n)} + \phi Q^{(n)^2} + (\tilde{p}^{(n)} - \alpha\lambda^{(n)}(Q^{(n)} + u^{(n)}))Q^{(n+1)} \right. \\
&\quad \left. + \mu^{(n+1)}Q^{(n+1)^2} \right].
\end{aligned}$$

After substituting $p^{(n)}, \tilde{p}^{(n)}, Q^{(n+1)}$ in terms of known quantities $(\tilde{p}^{(n-1)}, Q^{(n)}, x^{(n)})$, we obtain:

$$\begin{aligned}
L^{(n)}(\tilde{p}^{(n-1)}, Q^{(n-1)}, Q^{(n)}) &= \inf_{x^{(n)} \in \mathbf{M}(H^{(n)})} \mu^{(n+1)}x^{(n)^2} + (\alpha\lambda^{(n)}Q^{(n)} + \lambda^{(n)}Q^{(n)} - 2Q^{(n)}\mu^{(n+1)})x^{(n)} \\
&\quad + (-\alpha\lambda^{(n)} + \alpha\lambda^{(n-1)} + \mu^{(n+1)} + \phi)Q^{(n)^2} \\
&\quad + (\tilde{p}^{(n-1)} - \alpha\lambda^{(n-1)}(Q^{(n-1)} + u^{(n-1)}))Q^{(n)}.
\end{aligned}$$

This is a strictly convex quadratic in $x^{(n)}$ when $\mu^{(n+1)} > 0$, yielding the unique minimizer:

$$x^{(n)*} = \underbrace{\frac{2\mu^{(n+1)} - \lambda^{(n)}(1 + \alpha)}{2\mu^{(n+1)}}}_{=: \theta^{(n)}} Q^{(n)}. \quad (32)$$

Substituting this optimal $x^{(n)*}$ back, we recover:

$$L^{(n)}(\tilde{p}^{(n-1)}, Q^{(n-1)}, Q^{(n)}) = \left(\tilde{p}^{(n-1)} - \alpha \lambda^{(n-1)} (Q^{(n-1)} + u^{(n-1)}) \right) Q^{(n)} + \mu^{(n)} Q^{(n)^2}, \quad (33)$$

with

$$\mu^{(n)} = \alpha \lambda^{(n-1)} + \lambda^{(n)} + \phi - \frac{\lambda^{(n)^2} (1 + \alpha)^2}{4\mu^{(n+1)}}. \quad (34)$$

This completes the induction. $\square$

Characterization of kyle model's linear recursive equilibrium

We restate the main result of the discrete-time Kyle model (Kyle 1985) using the notation adopted in this paper.

THEOREM 5.1 (Recursive Linear Equilibrium of the Kyle Model) *There exists a unique linear equilibrium, which is a recursive linear equilibrium. In this equilibrium, there are constants $\beta^{(n)}, \lambda^{(n)}, \alpha^{(n)}, \delta^{(n)}$, and $\Sigma^{(n)}$ such that, for each trading period $n \in \{1, \dots, N\}$,*

$$x^{IT,(n)} = \beta^{(n)} (v - p^{(n-1)}) \Delta t^{(n)}, \quad (35)$$

$$p^{(n)} = p^{(n-1)} + \lambda^{(n)} (x^{IT,(n)} + u^{(n)}), \quad (36)$$

$$\Sigma^{(n)} := \text{Var}(v \mid q^{(1)}, \dots, q^{(n)}), \quad (37)$$

$$\mathbb{E} \left[\sum_{k=n}^N x^{IT,(k)} (v - p^{(k)}) \mid p^{(1)}, \dots, p^{(n-1)}, v \right] = \alpha^{(n-1)} (v - p^{(n-1)})^2 + \delta^{(n-1)}, \quad (38)$$

where $\Delta t^{(n)} := t^{(n)} - t^{(n-1)} = \tau$ denotes the time step size.

Given $\Sigma^{(0)}$, the constants $\beta^{(n)}, \lambda^{(n)}, \alpha^{(n)}, \delta^{(n)}, \Sigma^{(n)}$ are the unique solution to the recursive system:

$$\alpha^{(n-1)} = \frac{1}{4\lambda^{(n)} (1 - \alpha^{(n)}\lambda^{(n)})}, \quad (39)$$

$$\delta^{(n-1)} = \delta^{(n)} + \alpha^{(n)} \left(\lambda^{(n)} \right)^2 \sigma_u^2 \Delta t^{(n)}, \quad (40)$$

$$\beta^{(n)} \Delta t^{(n)} = \frac{1 - 2\alpha^{(n)}\lambda^{(n)}}{2\lambda^{(n)} (1 - \alpha^{(n)}\lambda^{(n)})}, \quad (41)$$

$$\lambda^{(n)} = \frac{\beta^{(n)}\Sigma^{(n)}}{\sigma_u^2}, \quad (42)$$

$$\Sigma^{(n)} = \left(1 - \beta^{(n)}\lambda^{(n)} \Delta t^{(n)} \right) \Sigma^{(n-1)}, \quad (43)$$

subject to $\alpha^{(N)} = \delta^{(N)} = 0$ and the second-order condition

$$\lambda^{(n)}(1 - \alpha^{(n)}\lambda^{(n)}) > 0. \quad (44)$$

References

- Almgren, R. and Chriss, N., Optimal Execution of Portfolio Transactions. *The Journal of Risk*, 2001, **3**, 5–39.
- Amrouni, S., Moulin, A., Vann, J., Vyetrenko, S., Balch, T. and Veloso, M., ABIDES-gym: Gym Environments for Multi-Agent Discrete Event Simulation and Application to Financial Markets. In *Proceedings of the Proceedings of the Second ACM International Conference on AI in Finance*, ICAIF'21, November, pp. 1–9, 2021, ACM.
- Bertsimas, D. and Lo, A.W., Optimal control of execution costs. *Journal of Financial Markets*, 1998, **1**, 1–50.
- Byrd, D., Hybinette, M. and Balch, T.H., ABIDES: Towards High-Fidelity Multi-Agent Market Simulation. In *Proceedings of the Proceedings of the 2020 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, SIGSIM-PADS '20, Miami, FL, Spain, p. 11–22, 2020 (Association for Computing Machinery: New York, NY, USA).
- Cao, B., Wang, S., Lin, X., Wu, X., Zhang, H., Ni, L.M. and Guo, J., From Deep Learning to LLMs: A survey of AI in Quantitative Investment. ArXiv preprint arXiv:2503.21422, 2025.
- Curato, G., Gatheral, J. and Lillo, F., Optimal execution with non-linear transient market impact. *Quantitative Finance*, 2017, **17**, 41–54.
- Dupret, J.L. and Hainaut, D., Optimal liquidation under indirect price impact with propagator. *Quantitative Finance*, 2025, **25**, 359–381.
- Engle, R.F., Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 1982, **50**, 987–1007.
- Foster, F.D. and Viswanathan, S., The Effect of Public Information and Competition on Trading Volume and Price Volatility. *The Review of Financial Studies*, 1993, **6**, 23–56.
- Foster, F.D. and Viswanathan, S., Strategic Trading When Agents Forecast the Forecasts of Others. *The Journal of Finance*, 1996, **51**, 1437–1478.
- Friedrich, P. and Teichmann, J., Deep Investing in Kyle's Single Period Model. *arXiv preprint*, 2020.
- Ganesh, A., D'Adamo, G., Firoozye, N. and Skoulakis, S., Reinforcement Learning for Market Making in a Multi-agent Dealer Market. *arXiv preprint*, 2019 ArXiv preprint arXiv:1911.05892.
- Gatheral, J., No-dynamic-arbitrage and market impact. *Quantitative Finance*, 2010, **10**, 749–759.
- Glosten, L.R. and Milgrom, P.R., Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders. *Journal of Financial Economics*, 1985, **14**, 71–100.
- Hafsi, Y. and Vittori, E., Optimal Execution with Reinforcement Learning. ArXiv preprint arXiv:2411.06389, 2024.

- Hasselt, H.v., Guez, A. and Silver, D., Deep reinforcement learning with double Q-Learning. In *Proceedings of the Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, Phoenix, Arizona, p. 2094–2100, 2016, AAAI Press.
- Hendricks, D. and Wilcox, D., A reinforcement learning extension to the Almgren-Chriss framework for optimal trade execution. In *Proceedings of the IEEE Conference on Computational Intelligence for Financial Engineering & Economics, CIFE 2014, London, UK, March 27-28, 2014*, pp. 457–464, 2014, IEEE.
- Ho, T. and Stoll, H.R., Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, 1981, **9**, 47–73.
- Holden, C.W. and Subrahmanyam, A., Long-Lived Private Information and Imperfect Competition. *The Journal of Finance*, 1992, **47**, 247–270.
- Huang, M., Caines, P.E. and Malhamé, R.P., Large population stochastic dynamic games: closed-loop McKean–Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems*, 2006, **6**, 221–252.
- Huberman, G. and Stanzl, W., Price Manipulation and Quasi-Arbitrage. *Econometrica*, 2004, **72**, 1247–1275.
- Huberman, G. and Stanzl, W., Optimal Liquidity Trading. *Review of Finance*, 2005, **9**, 165–200.
- Jaisson, T., Deep differentiable reinforcement learning and optimal trading. *Quantitative Finance*, 2022, **22**, 1429–1443.
- Karpe, M., Fang, J., Ma, Z. and Wang, C., Multi-Agent Reinforcement Learning in a Realistic Limit Order Book Market Simulation. In *Proceedings of the Proceedings of the First ACM International Conference on AI in Finance*, ICAIF '20, New York, New York, Oct., 2020, ACM.
- Kuba, J.G., Chen, R., Wen, M., Wen, Y., Sun, F., Wang, J. and Yang, Y., Trust Region Policy Optimisation in Multi-Agent Reinforcement Learning. ArXiv preprint arXiv:2109.11521, 2022.
- Kyle, A.S., Continuous Auctions and Insider Trading. *Econometrica*, 1985, **53**, 1315–1335.
- Kyle, A.S., Informed Speculation with Imperfect Competition. *The Review of Economic Studies*, 1989, **56**, 317–355.
- Lasry, J.M. and Lions, P.L., Mean field games. *Japanese Journal of Mathematics*, 2007, **2**, 229–260.
- Li, F., Ihnatiuk, V., Chen, Y., Lin, J., Kinnear, R.J., Schneider, A., Nevmyvaka, Y. and Lam, H., Do price trajectory data increase the efficiency of market impact estimation?. *Quantitative Finance*, 2024, **24**, 545–568.
- Liang, E., Liaw, R., Moritz, P., Nishihara, R., Fox, R., Goldberg, K., Gonzalez, J.E., Jordan, M.I. and Stoica, I., RLlib: Abstractions for Distributed Reinforcement Learning. ArXiv preprint arXiv:1712.09381, 2018.
- Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D. and Wierstra, D., Continuous Control with Deep Reinforcement Learning. arXiv preprint arXiv:1509.02971, 2019.
- Lin, S. and Beling, P.A., An End-to-End Optimal Trade Execution Framework based on Proximal Policy Optimization. In *Proceedings of the Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, edited by C. Bessiere, 7 Special Track on AI in FinTech, pp. 4548–4554, 2020, International Joint Conferences on Artificial Intelligence Organization.
- Lin, S. and Beling, P.A., An Agent-Based Market Simulator for Back-Testing Deep Reinforcement Learning Based Trade Execution Strategies. In *Proceedings of the Neural Information Processing*, edited by T. Mantoro, M. Lee, M.A. Ayu, K.W. Wong and A.N. Hidayanto, pp. 644–653, 2021 (Springer International Publishing: Cham).
- Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P. and Mordatch, I., Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. ArXiv preprint arXiv:1706.02275, 2020.
- Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D. and Kavukcuoglu, K., Asynchronous Methods for Deep Reinforcement Learning. In *Proceedings of the Proceedings of The 33rd International Conference on Machine Learning*, edited by M.F. Balcan and K.Q. Weinberger, Vol. 48 of *Proceedings of Machine Learning Research*, 20–22 Jun, pp. 1928–1937, 2016 (PMLR: New York, New York, USA).
- Moallemi, C.C. and Wang, M., A reinforcement learning approach to optimal execution. *Quantitative Finance*, 2022, **22**, 1051–1069.
- Nagy, P., Calliess, J.P. and Zohren, S., Asynchronous Deep Double Dueling Q-learning for trading-signal execution in limit order book markets. *Frontiers in Artificial Intelligence*, 2023, **Volume 6 - 2023**.
- Nevmyvaka, Y., Feng, Y. and Kearns, M., Reinforcement Learning for Optimized Trade Execution. In *Proceedings of the Proceedings of the 23rd International Conference on Machine Learning*, pp. 673–680, 2006.

- Ning, B., Lin, F.H.T. and Jaimungal, S., Double Deep Q-Learning for Optimal Execution. Papers 1812.06600, arXiv.org, 2018 ArXiv preprint arXiv:1812.06600.
- Obizhaeva, A.A. and Wang, J., Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, 2013, **16**, 1–32.
- Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M. and Dormann, N., Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research*, 2021, **22**, 1–8.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O., Proximal Policy Optimization Algorithms. ArXiv preprint arXiv:1707.06347, 2017.
- Seabold, S. and Perktold, J., statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the Proceedings of the 9th Python in Science Conference*, edited by S. van der Walt and J. Millman, pp. 57–61, 2010.
- Terry, J.K., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L., Perez, R., Horsch, C., Dieffendahl, C., Williams, N.L., Lokesh, Y. and Ravi, P., PettingZoo: Gym for Multi-Agent Reinforcement Learning. ArXiv preprint arXiv:2009.14471, 2021.
- Towers, M., Kwiatkowski, A., Terry, J., Balis, J.U., Cola, G.D., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J.J., Tan, H. and Younis, O.G., Gymnasium: A Standard Interface for Reinforcement Learning Environments. ArXiv preprint arXiv:2407.17032, 2024.
- Wang, X., Hoang, C., Vorobeychik, Y. and Wellman, M.P., Spoofing the Limit Order Book: A Strategic Agent-Based Analysis. *Games*, 2021, **12**.
- Watkins, C.J.C.H. and Dayan, P., Q-learning. *Machine Learning*, 1992, **8**, 279–292.
- Williams, R.J., Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, 1992, **8**, 229–256.
- Yang, H., Zhang, B., Wang, N., Guo, C., Zhang, X., Lin, L., Wang, J., Zhou, T., Guan, M., Zhang, R. and Wang, C.D., FinRobot: An Open-Source AI Agent Platform for Financial Applications using Large Language Models. ArXiv preprint arXiv:2405.14767, 2024.
- Yu, C., Velu, A., Vinitisky, E., Gao, J., Wang, Y., Bayen, A. and Wu, Y., The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games. ArXiv preprint arXiv:2103.01955, 2022.
- Yu, Y., Li, H., Chen, Z., Jiang, Y., Li, Y., Zhang, D., Liu, R., Suchow, J.W. and Khashanah, K., FinMem: A Performance-Enhanced LLM Trading Agent with Layered Memory and Character Design. ArXiv preprint arXiv:2311.13743, 2023.
- Zhang, W., Zhao, L., Xia, H., Sun, S., Sun, J., Qin, M., Li, X., Zhao, Y., Zhao, Y., Cai, X., Zheng, L., Wang, X. and An, B., A Multimodal Foundation Agent for Financial Trading: Tool-Augmented, Diversified, and Generalist. In *Proceedings of the Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, Barcelona, Spain, p. 4314–4325, 2024 (Association for Computing Machinery: New York, NY, USA).
- Álvaro Cartea, Jaimungal, S. and Penalva, J., *Algorithmic and High-Frequency Trading*, 2015, Cambridge University Press.