Peng Liu, Xiangyu Sun, Yiqiao Sun, Yize Wang

Massachusetts Institute of Technology

15.S06 Project

Reinforcement Learning in Simulated Financial Markets

Project Instructor: Leonid Kogan

December 12, 2025

Contents

Chapter 1

Introduction

1.1 Background and Motivation

Reinforcement learning (RL) is increasingly adopted in financial markets for portfolio allocation, execution, and market making. Major exchanges and trading firms already deploy AI-driven order types and trading bots, while regulators worry that algorithms may learn to coordinate in ways that undermine market quality even without explicit communication or intent. Recent work on algorithmic collusion in product markets (e.g. Calvano et al., 2020) and securities markets (Dou et al., 2025) shows that self-interested RL agents can converge to outcomes that resemble tacit collusion, raising new challenges for market oversight and antitrust enforcement.

At the same time, modern electronic markets are organized around limit-order books (LOBs) operated by market makers. These intermediaries aggregate heterogeneous order flow, infer information from prices and quantities, and manage inventory risk while setting quotes. Classical microstructure models such as Kyle (1985) have clarified how informed traders, noise traders, and competitive market makers interact in partial equilibrium, but these models typically impose linear strategies and rational expectations. They leave open how algorithmic agents might *learn* to trade and quote in a realistic LOB environment and how the resulting equilibrium affects price efficiency and surplus sharing between informed traders and liquidity providers.

This paper builds a simulated limit-order-book market in which reinforcement-learning agents trade a short-lived asset. We start from a simple automated market maker that mechanically clears limit orders at a single price, and then move to an order-book setting where a rule-based market maker posts a three-level LOB and faces an RL informed trader. Finally, we allow two competing market makers to learn their own quoting policies via DQN in the same environment. This stepwise design allows us to isolate how the sophistication of market makers and the learning capability of informed traders jointly shape price efficiency, inventory dynamics, and emergent trading strategies.

Our central questions are: (i) how does an RL-informed trader exploit a naive order-book market maker, and how much of its private information is ultimately incorporated into prices; (ii) when market makers themselves learn via RL, do they compete or (implicitly) collude, and how does this affect the distribution of profits; and (iii) to what extent can a regulator or outside observer infer the presence of informed trading or AI collusion from public LOB data alone. To answer these questions, we simulate thousands of episodes, document the training dynamics of the RL agents, and analyze the implied price paths, inventories, and policies.

Our contributions are threefold. First, we develop a microstructure-level RL environment that nests classical informed-trading models in a realistic LOB with inventory-constrained market makers. Second, in the single-maker environment we show that the RL-informed trader endogenously learns to hide its orders when prices are far from its private signal, to condition on recent price trends, and to exploit a mechanically adaptive but non-learning market maker. Third, when two RL market makers compete, we study how their learned quoting rules affect price efficiency and the detectability of informed trading from public data, connecting the AI-collusion literature to the concrete structure of LOB markets.

1.2 Related Literature

Our framework builds on the canonical microstructure benchmark of Kyle (1985), which formalizes price discovery with asymmetric information in a dynamic trading environment. In Kyle's model, a risk-neutral insider trades on private information about the terminal liquidation value X alongside noise traders, while competitive, risk-neutral market makers set prices efficiently (in the semi-strong sense) given their information set. At each auction, market makers observe only aggregate order flow, so innovations in order flow are the primitive source of price changes. The insider optimally trades as an intertemporal monopolist, exploiting private information while using noise trading as camouflage; in equilibrium, prices become increasingly informative over time and incorporate all private information by the end of trading. A key implication is that competitive market makers earn zero expected profit, even though the insider earns positive profits, precisely because market makers price at the conditional expectation of fundamentals given observed order flow.

Our paper extends this logic to a limit-order-book (LOB) setting in which trading and liquidity provision are generated by reinforcement-learning (RL) agents rather than imposed linear strategies. The different setting matters because Kyle-style models deliver sharp equilibrium predictions under strong structural assumptions (normality and linearity), whereas modern electronic markets feature discrete quoting, depth dynamics, and inventory constraints. In our environment, the asset payoff is drawn at the start of each episode and remains fixed over T steps, and the informed trader and market makers receive noisy private signals with the informed trader being more precise. These design choices allow us to study how prices converge toward fundamentals when both liquidity demand and liquidity supply are strategic and learned.

We are also related to the literature on algorithmic pricing and AI-enabled tacit collusion. Calvano et al. (2020) show, in a repeated oligopoly pricing environment, that standard Q-learning agents can learn to sustain supracompetitive prices without communication, typically through history-dependent strategies that punish deviations for a finite period and then gradually return to cooperation. This line of work highlights that collusion-like outcomes can emerge endogenously from learning dynamics rather than explicit coordination, and it motivates concerns that algorithmic decision-makers may alter the traditional feasibility of tacit collusion.

Most closely, Dou et al. (2025) adapt the experimental RL-collusion agenda to a microstructure setting with asymmetric information, noise trading, and market makers, and they study how collusive outcomes affect market quality and efficiency. Their simulations emphasize that stronger AI collusion can be associated with reduced liquidity and lower price informativeness, and they identify distinct algorithmic channels (e.g., price-trigger-like mechanisms versus learning biases such as over-pruning) through which collusion can arise. Compared to Dou et al. (2025), our focus is less on cataloging collusive mechanisms and more on the interplay between (i) an RL informed trader's exploitation and concealment incentives and (ii) competitive versus potentially collusive behavior among multiple RL market makers in an explicit LOB environment, while using additional tests to evaluate informational efficiency and detectability from public quotes.

Relative to these literatures, our contribution is methodological and substantive. Methodologically, we construct a microstructure-level RL environment that nests Kyle's key informational structure (private signals, noise trading, and pricing from aggregated information) but operationalizes trading through discrete LOB states and inventory-constrained liquidity provision. Substantively, we use this environment to evaluate (a) how quickly private information is incorporated into prices under learned quoting and trading policies, (b) whether two learning market makers converge to competitive outcomes or collusion-like pricing, and (c) whether an external observer can infer informed trading using only public LOB features. These questions connect Kyle-style informational efficiency to the modern concern that learning agents may change equilibrium market quality, and to the practical regulatory question of whether public market data contains reliable signatures of private information or algorithmic coordination.

Chapter 2

Market Design and Learning Setup

2.1 Asset, payoff, and private information

We consider a short-lived asset that pays a random terminal value X at the end of each episode. At the beginning of an episode, nature draws

$$X \sim \log \mathcal{N}(\mu, \sigma),$$

and this payoff remains fixed for T trading steps. An informed trader (IT) and each market maker i receive private signals

$$s^{\mathrm{IT}} = X + \varepsilon^{\mathrm{IT}}, \qquad s_i^{\mathrm{MM}} = X + \varepsilon_i^{\mathrm{MM}},$$

with $\varepsilon^{\mathrm{IT}} \sim \mathcal{N}(0, \sigma_{\mathrm{IT}}^2)$ and $\varepsilon_i^{\mathrm{MM}} \sim \mathcal{N}(0, \sigma_{\mathrm{MM}}^2)$, where $\sigma_{\mathrm{IT}} \ll \sigma_{\mathrm{MM}}$. Thus the informed trader has more precise information about X than the market makers.

2.2 Agents and order flow

The market hosts three types of agents.

Informed trader. The informed trader observes its private signal s^{IT} and the top-of-book bid and ask quotes (b_t, a_t) from the consolidated LOB. It chooses a signed market-order quantity

$$a_t^{\mathrm{IT}} \in \{-200, -100, 0, 100, 200\}$$

at each step t, and aims to maximize terminal wealth

$$R_T^{\mathrm{IT}} \ = \ \frac{\big(\mathrm{cash}_T + \mathrm{inv}_T X\big) - \mathrm{cash}_0}{1000}.$$

The state vector at step t is six-dimensional:

$$s_t = \left[\frac{s^{\mathrm{IT}} - \mathrm{mid}_t}{\mathrm{mid}_t},\ \frac{\mathrm{spread}_t}{\mathrm{mid}_t},\ \frac{t}{T},\ \frac{\mathrm{inv}_t}{5000},\ r_{1,t},\ r_{2,t} \right],$$

where $\mathrm{mid}_t = (a_t + b_t)/2$, $\mathrm{spread}_t = a_t - b_t$, and $r_{1,t}, r_{2,t}$ denote the last one-step and two-step mid-price returns. The trader is modelled as a DQN agent taking s_t as input and outputting Q-values for each action.

Noise trader. The noise trader provides non-strategic liquidity and order-flow randomness. At each step it draws a signed quantity

$$q_t^{\mathrm{NZ}} \sim \mathcal{N}(\mu_{\mathrm{NZ}}, \sigma_{\mathrm{NZ}}^2)$$

and rounds it to the nearest integer. The order is submitted as a market order each period. The noise trader has no private information and does not learn.

Market makers. We consider two LOB-based market structures built on the same environment:

1. **Single naive market maker** A single market maker posts a three-level LOB at each step following a deterministic quoting rule. It updates a reference price p_t^{ref}, adjusts the half-spread h_t as a function of its inventory, and then computes an inventory-skewed reservation price $p_t^\star$ from which it derives the three bid and ask levels. This market maker does not use RL; instead, it mechanically reacts to net order flow and inventory.
2. **Two RL market makers** Two symmetric market makers, MM_A and MM_B, each observe their own noisy signals s_i^{MM} and maintain internal reference prices $p_{\mathrm{ref},t}^{(i)}$. They post three-level LOBs that are subsequently merged into a consolidated book. Each maker is an RL agent with a twelve-dimensional state (including signal gap, inventory, spread, and order-flow statistics) and a discrete action set controlling spread, skew, and quote sizes. Both market makers learn via DQN and receive terminal rewards equal to their mark-to-payoff P&L minus an inventory penalty.

In both structures, the informed and noise traders always act as liquidity takers via market orders.

2.3 Single-market-maker quoting rule

In the single-maker environment, the market maker first updates its reference price (its current estimate of fair value) according to net order flow:

$$p_{t+1}^{\mathrm{ref}} = p_t^{\mathrm{ref}} + \mathrm{mm\_impact} \cdot \tanh\left(\frac{\mathrm{NetFlow}_t}{3 \cdot \mathrm{base\_size}}\right) \cdot h_t,$$

where the half-spread h_t widens with inventory deviation,

$$h_t = \mathrm{base\_half\_spread} \cdot \big(1 + \mathrm{inv\_spread\_slope} \cdot \Delta\mathrm{inv}_t\big),$$

and $\Delta\mathrm{inv}_t$ denotes the deviation of inventory from its initial level. The market maker then calculates an inventory-skewed reservation price

$$p_{t+1}^\star = p_{t+1}^{\mathrm{ref}} - \mathrm{inv\_fair\_skew} \cdot \Delta\mathrm{inv}_{t+1},$$

and finally posts three bid and ask levels:

$$\mathrm{bid}_{t+1}^{(k)} = p_{t+1}^\star - (k+0.5)h_{t+1}, \qquad \mathrm{ask}_{t+1}^{(k)} = p_{t+1}^\star + (k+0.5)h_{t+1}, \quad k = 0, 1, 2.$$

This "tedious" update rule is necessary because the market maker's signal is much noisier than the informed trader's signal. Without such a mechanism, the mid-price could remain far from true value for the entire episode, eliminating trading opportunities; the quoting rule, combined with order flow, allows the maker to partially learn about fundamentals.

2.4 Single Market Maker: Simulation Process and Model Training

2.4.1 Initialization

At each episode, the market start with an empty order book. To initialize, the market maker will receives an estimate of the fundamental value (reference price) of the asset, then post limit orders based on this reference price and its default spread.

2.4.2 Per-step market dynamics

1. **Market Maker:** update fair value $\hat{v}_t$ and volatility $\hat{\sigma}_t$, compute reservation price r_t and spread, generate 3 levels of (b_t, a_t).
2. **Noise Trader:** with probability p (= 1 for now), send small buy/sell market order.
3. **Informed Agent:** receive private signal s (fixed),
   form state $x_t = f(s_t, \text{market}, \text{inventory}, t/T)$,
   choose action $a_t = \pi_\theta(x_t)$,
   execute market order.
4. **Matching Engine:** Orders interact under price–randomized time priority.
5. **Settlement:** Update cash, shares, and apply maker/taker fees (= 0 for now).
6. **RL Transition:** record $(x_t, a_t, \text{mid}_t, x_{t+1})$, perform Q-learning update.

2.4.3 Training Reward

At the beginning of each episode, a new payoff X is drawn and all agents are reset. The market then runs for T steps as described above. At the end of the episode, we compute terminal rewards for informed trader:

$$R_T^{\text{IT}} = \frac{\big(\text{cash}_T + \text{inv}_T X\big) - \text{cash}_0}{1000},$$

2.5 Dual Market Makers: Quoting, Internal Crossing, and Order Cancellation

In the third market configuration, two symmetric RL market makers (MM_A and MM_B) participate in the market simultaneously. At the beginning of each episode, MMs will have an initial estimate of the fundamental value:

$$s_i^{MM} \sim N(FV_i, \sigma_{MM}^2)$$

Where FV_i is the fundamental value of the asset at episode i. σ_{MM} is the variance of estimation of MM. In the experiment we have $\sigma_{MM} = 10$. Each MM maintains its own private signal s_i^{MM}, reference price $p_t^{\mathrm{ref},(i)}$, inventory $\mathrm{inv}_t^{(i)}$, and a DQN-based quoting policy. At every step, both market makers independently update their internal states and post three-level limit-order books before their quotes are merged into a consolidated book. This environment introduces competition between liquidity providers and allows for strategic interactions such as internal crossing, quote improvement, and dynamic cancellation.

State space. MMs will observe the current environment(orderbook and flow information) and act based on their obervations. The state space include:

- Mid Deviation: $\frac{Mid_{i,t} - P_{ref,t}}{P_{ref,t}}$
- Relative Spread: $\frac{Spread_{i,t}}{Mid_{i,t}}$
- Time Fraction: $\frac{t}{T}$
- Scaled Inventory: $\frac{Inv_{i,t}}{10000}$
- Flow: $(1-\alpha) \cdot flow + \alpha \cdot flow_t$
- Imbalance: $\frac{Bid\_Qty_{i,t} - Ask\_Qty_{i,t}}{Bid\_Qty_{i,t} + Ask\_Qty_{i,t}}$
- Last Action: Last Spread, Bid, Ask (Normalized)

Reference-price update and inventory adjustment. Each market maker updates its internal reference price using both its private signal and the order flow it receives:

$$p_{t+1}^{\mathrm{ref},(i)} = p_t^{\mathrm{ref},(i)} + \kappa h_t^{(i)} \tanh\left(\frac{q_t^{(i)}}{3 \cdot \mathrm{base\_size}} \right) + \alpha\big(s_i^{\mathrm{MM}} - p_t^{\mathrm{ref},(i)}\big),$$

where $q_t^{(i)}$ is the net market-order flow hitting MM_i at step t, and $h_t^{(i)}$ is the maker-specific half-spread. The second term incorporates learning from order flow, while the final term pulls the reference price toward the maker's private signal.

Each maker computes an inventory-adjusted fair price,

$$p_t^{\mathrm{fair},(i)} = s_i^{\mathrm{MM}} - \phi\, \Delta \mathrm{inv}_t^{(i)},$$

and then combines the two sources of information to form a reservation center:

$$p_t^{\star,(i)} = (1-\beta) p_t^{\mathrm{ref},(i)} + \beta\, p_t^{\mathrm{fair},(i)}.$$

RL-controlled quoting. The DQN action for MM_i consists of three discrete multipliers:

$$a_t^{(i)} = (\text{spread multiplier, skew multiplier, size multiplier}),$$

which jointly determine

$$h_t^{(i)}, \qquad p_t^{\star,(i)}, \qquad q_t^{(i,k)}, \quad k = 0, 1, 2.$$

Given the action, MM_i posts:

$$\begin{aligned} \text{bid}_t^{(i,k)} &= p_t^{\star,(i)} - (k + 0.5)h_t^{(i)} + \text{skew adjustment}, \\ \text{ask}_t^{(i,k)} &= p_t^{\star,(i)} + (k + 0.5)h_t^{(i)} + \text{skew adjustment}, \end{aligned}$$

where $q_t^{(i,k)}$ follows geometric size decay.

Internal crossing between market makers. Before any external market orders interact with the LOB, we resolve potential "internal crosses" between MM_A and MM_B. Whenever

$$\max_k \text{bid}_t^{(A,k)} \;\geq\; \min_k \text{ask}_t^{(B,k)},$$

a trade occurs directly between the market makers at the mid-price of the overlapping quotes. Symmetrically, if B's bid crosses A's ask, the makers trade in the opposite direction.

This mechanism enforces competition:

- a maker posting too aggressive quotes gets "picked off" by the rival,
- makers learn to balance informativeness, inventory risk, and adverse-selection risk imposed by each other,
- makers are incentivized to shade their quotes strategically rather than blindly narrowing spreads.

Quote merging and cancellation. After internal crossing, the two books are merged into a consolidated LOB:

$$\text{OrderBook}_t = \text{merge}(\text{Book}_A, \text{Book}_B),$$

where all remaining live quotes are sorted by price and time.

All quotes from previous periods are *automatically cancelled* when a new period begins. This reflects standard exchange behavior: each maker posts a fresh three-level book at every step, preventing "stale" quotes from accumulating or conflicting with RL-driven adjustments.

Execution of informed and noise traders. Once the consolidated book is constructed:

1. the informed trader and noise trader submit signed market orders,
2. their execution priority is randomized to avoid systematic bias,
3. fills propagate to the inventories and cash accounts of MM_A, MM_B, and the traders.

Terminal rewards and learning. After T steps, each market maker receives:

$$R_T^{\mathrm{MM}_i} = \frac{\left(\mathrm{cash}_T^{(i)} + \mathrm{inv}_T^{(i)} X\right) - \left(\mathrm{cash}_0^{(i)} + \mathrm{inv}_0^{(i)} X\right) - \lambda_{\mathrm{inv}} (\mathrm{inv}_T^{(i)})^2}{1000},$$

which captures trading profits and an inventory-penalty term discouraging large risk exposure. The informed trader receives its payoff-based P&L reward, and both RL systems update based on replay-buffer samples.

Overall, the dual-maker environment enables rich strategic interactions: competition for order flow, defensive widening of spreads, mutual learning via intern

2.6 Dual RL Market Makers: Simulation Process and Model Training

2.6.1 Initialization

At the start with each episode, the market starts with an empty order book. To initialize the market, the market makers will:

1. Receives an initial estimation (s_{MM}) of the fundamental value.
2. Use neutral action (Action at the mid position of the action space) to post limit order books.

2.6.2 Per-step market dynamics

Within each episode, the market evolves over $t = 1, \ldots, T$ according to the following loop:

1. **Market Makers (MMs) Observe:** MMs A and B observe the current order book and state variables.
2. **MM Action:** `mmA_agent` and `mmB_agent` select actions (`a_mmA`, `a_mmB`) based on their RL policy (π_{MM}) and exploration (ϵ).
3. **Order Posting:** mmA post limit orders on bookA, mmB on bookB.
4. **Internal Crossing:**
   - Resolve crosses between `bookA` and `bookB` (A's bid matches B's ask).
   - Update inventory and cash of MMs
5. **Merge Books:** Create the **Consolidated Book (**`OrderBook`**)** containing all live quotes.
6. **Informed Trader (IT) Action**
   (a) IT observes the best bid/ask of `cbook` (`proxy_book`).

(b) IT agent selects market order size (`a_it`).

7. **Noise Trader (NZ) Action:** Samples a random market order (`q_nz`).
8. **Execution:** IT and NZ market orders are executed sequentially against the `orderbook` (In random order).
9. **Market Maker:** Update reference price and flow information.
10. **Reward Calculation:**
    - If $t < T$: $reward = 0$ for all agents(mmA,mmB and IT).
    - if $t = T$:
      - $reward_{MM} = FinalPV - InitPV - invPenalty$
      - $reward_{IT} = FinalPV - InitPV$
11. **Agent Training(MM & IT):**
    (a) Store state, action, reward, and next state in the replay buffers.
    (b) Sample from replay buffer and back propagation.

2.6.3 Training reward

At the beginning of each episode, a new payoff X is drawn and all agents are reset. The market then runs for T steps as described above. At the end of the episode, we compute terminal rewards:

$$R_T^{\mathrm{IT}} = \frac{\big(\mathrm{cash}_T + \mathrm{inv}_T X\big) - \mathrm{cash}_0}{1000},$$

and, for each market maker i in the dual-maker environment,

$$R_T^{\mathrm{MM}_i} = \frac{\big(\mathrm{cash}_T^{(i)} + \mathrm{inv}_T^{(i)} X\big) - \big(\mathrm{cash}_0^{(i)} + \mathrm{inv}_0^{(i)} X\big) - \lambda_{\mathrm{inv}}(\mathrm{inv}_T^{(i)})^2}{1000}.$$

There are no intermediate rewards; all RL agents learn purely from terminal P&L. To stabilize training, we increase the sampling weight of terminal transitions in the replay buffer and run multiple gradient steps at the end of each episode.

Chapter 3

Results: Single Naive Market Maker and RL Informed Trader

We now focus on the second market structure, where a single rule-based market maker posts the LOB and only the informed trader learns via RL. This environment is particularly transparent for understanding how the informed agent exploits a naive liquidity provider and how much price efficiency can be achieved when the maker does not learn.

3.1 Training dynamics and convergence

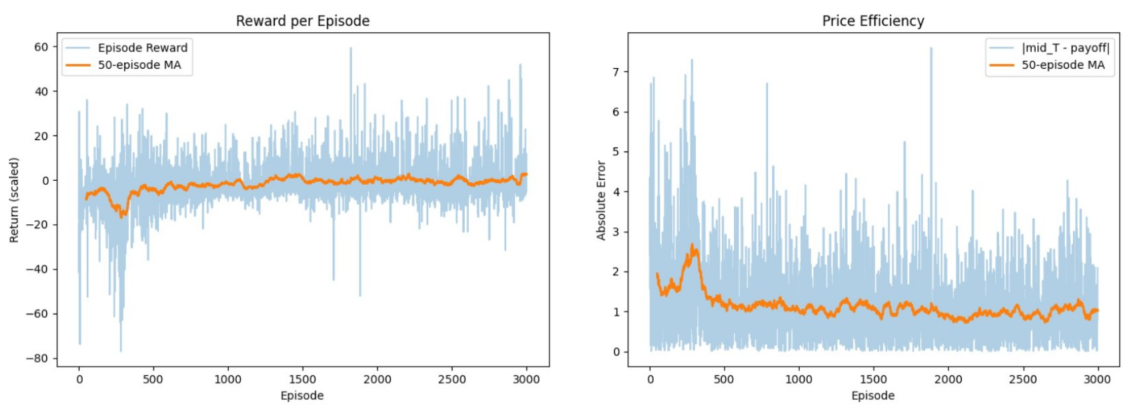

Figure 3.1: Training Dynamics

Figure 3.1 reports the training dynamics over 3,000 episodes. The left panel shows the terminal reward of the informed trader in each episode (light blue) together with a 50-episode moving average (orange). The average reward increases steadily and stabilizes at a positive level, indicating that the DQN converges to a profitable trading policy. Interestingly, the volatility of episode-level rewards first decreases and then increases again as training progresses. Intuitively, the agent initially learns to avoid obviously bad trades, which compresses the distribution of outcomes, and later begins to exploit more extreme states (for example, taking larger positions when its signal is far from the mid-price), which increases the dispersion of P&L even as the mean continues to rise.

The right panel of Figure 3.1 plots the absolute pricing error $|\mathrm{mid}_T - X|$ at the end of each episode along with its moving average. Price efficiency improves over time: the average absolute error declines significantly as the informed trader learns to trade in the "right direction" and pushes prices closer to fundamentals. However, the error does not converge to zero, reflecting the fact that no agent knows X exactly and the market maker has only limited ability to infer X from order flow given its noisy signal and mechanical quoting rule.

3.2 Price efficiency and inventory paths

To illustrate the resulting market outcomes, Figure 3.2 shows a representative "test episode" using the trained policy.

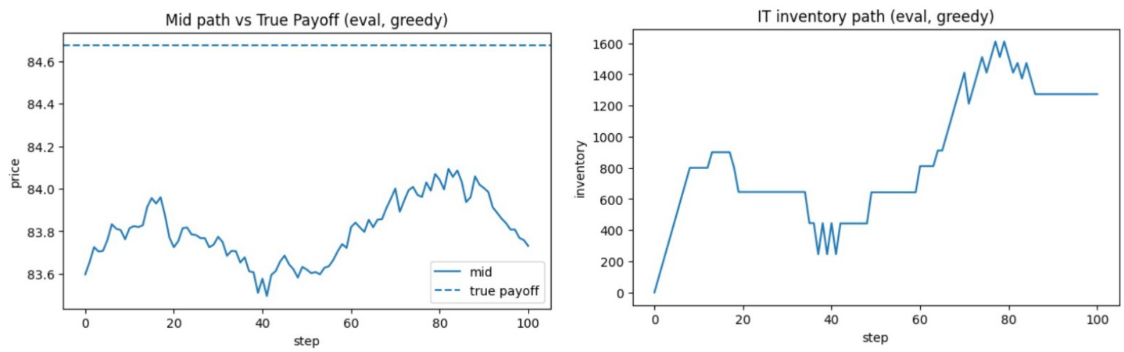

Figure 3.2: Pricing Inventory Path

The left panel plots the mid-price path mid_t alongside the realized payoff X. The mid-price tracks the payoff reasonably well and gradually drifts toward X, but it does not fully converge: even near the end of the episode, the mid remains modestly below the payoff. This residual mispricing reflects two frictions: first, nobody in the market observes X directly, not even the informed trader; second, the market maker is only partially able to extract fundamental information from order flow through its deterministic update rule.

The right panel of Figure 3.2 displays the informed trader's inventory over the same episode. The inventory path shows that the trader builds up a sizable long position when its signal is above the prevailing mid-price and gradually reduces or stabilizes the position as prices approach its valuation. In the particular test episode plotted here, the informed trader realizes roughly a 30% profit relative to initial capital, illustrating that substantial informational rents remain even though prices become somewhat informative.

3.3 Learned trading strategy of the informed trader

Beyond aggregate P&L, we analyze the structure of the learned policy. Figure 3.3 summarizes the informed trader's strategy after training.

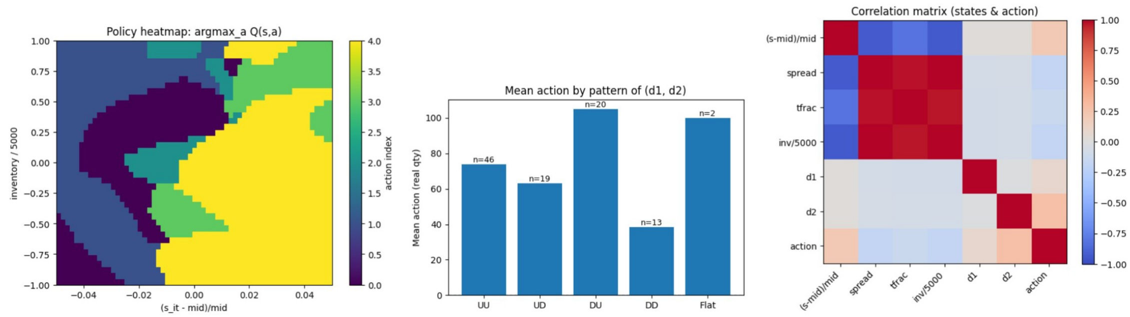

Figure 3.3: Policy Heatmaps

The left panel reports a policy heatmap of the greedy action $\arg\max_a Q(s, a)$ on a grid over two key state variables: the signal–price gap $(s^{\mathrm{IT}} - \mathrm{mid}_t)/\mathrm{mid}_t$ and scaled inventory $\mathrm{inv}_t/5000$. Several patterns emerge. When the signal is only slightly above the mid-price and inventory is low, the trader optimally buys (positive quantities), whereas when the signal is slightly below and inventory is low it sells. As inventory becomes large and positive, the trader refrains from further buying even if its signal remains above price, and symmetrically for large negative inventory. Most importantly, when the signal–price gap is very large in absolute value, the agent *does not* submit the maximum order size. Instead, it "hides" by trading more cautiously to avoid moving the price too aggressively and revealing its information. This endogenous order-hiding behavior closely mirrors the intuition from Kyle-type models that informed traders trade less aggressively when their information advantage is easily detectable.

The middle panel aggregates actions by short-run price patterns, grouping states according to whether the last two mid-price changes are up or down. The bars show that the informed trader tends to sell when prices have been moving down for two consecutive steps (bearish momentum) and to buy when prices move from down to up (bullish reversal). In other words, the policy combines momentum-following in persistent downtrends with contrarian behavior around potential turning points, depending on how the signal compares to the current mid-price.

Finally, the right panel plots the correlation matrix between the six state variables and the realized actions. The pairwise correlations between the action and individual state components are relatively small in magnitude, suggesting that the DQN is not simply implementing a linear threshold rule on any single variable. Instead, the policy is genuinely nonlinear and interacts multiple dimensions—signal gap, spread, time, and inventory, together with recent returns—to determine order size. This provides evidence that the neural network architecture is effectively exploited by the RL algorithm rather than merely rediscovering a simple linear strategy.

Overall, the single-maker environment demonstrates that even when facing a mechanically adaptive but non-learning market maker, an RL-informed trader can develop sophisticated execution patterns: hiding orders when its information advantage is extreme, conditioning on recent price dynamics, and exploiting the maker's limited ability to infer fundamentals from order flow. These findings provide a useful benchmark for interpreting the more complex interactions that arise once market makers themselves begin to learn via RL in the dual-maker environment.

Chapter 4

Results: Dual RL Market Makers and an RL Informed Trader

This section examines the equilibrium properties of a simulated limit-order-book market populated by two symmetric reinforcement-learning market makers (MM_A and MM_B), an informed trader (IT) trained via reinforcement learning, and a noisy trader. Our analysis focuses on three main questions:

(i) how competition affects market-maker and informed-trader behaviors

(ii) whether prices efficiently incorporate private information

(iii) whether informed trading leaves detectable traces in public order-book data

4.1 Dual RL Market Makers and Competitive Quoting Behavior

We analyze the behavior and profitability of the two reinforcement-learning market makers. Each market maker dynamically chooses a reservation-price adjustment ("lean") and an effective half-spread based on inventory, order-flow conditions, and public signals. The result is shown below:

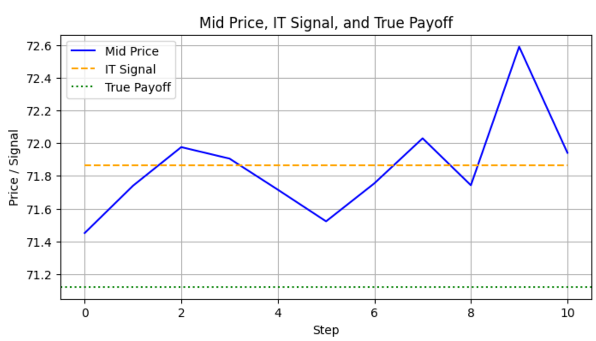

Figure 4.1: Price Evolution

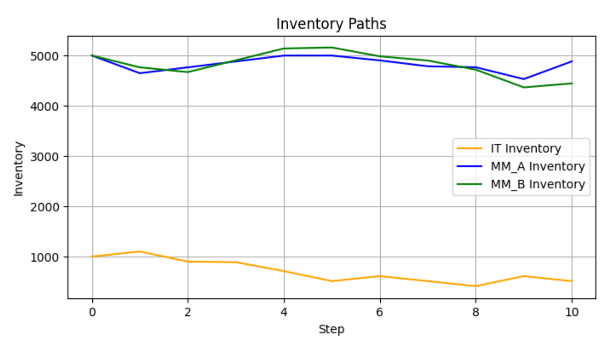

Figure 4.2: Inventory Path

The learned policies display economically intuitive structure. From Figure 4.1 and Figure 4.2, we observe that the informed trader is able to move prices through the submission of buy and sell market orders. The informed trader's inventory trajectory closely mirrors the evolution of the mid price: periods of aggressive buying coincide with upward price movements, while selling pressure pushes prices downward. This co-movement reflects the mechanism through which private information is incorporated into prices.

Moreover, the mid price tends to fluctuate around the informed trader's private signal rather than around the true terminal payoff. This is consistent with the informational structure of the market. Since no agent directly observes the fundamental payoff, the informed trader's signal represents the most informative estimate available. As the informed trader trades on this signal, it becomes partially revealed through order flow, leading the market makers to update their reservation prices accordingly. Consequently, the price process tracks the informed signal—its best publicly inferable component, while the true payoff remains unknown until the end of the episode.

4.2 RL Agents Terminal Portfolio Value

One core research goal of this project is to demonstrate the ability of using RL in trading to make money. We plotted the terminal portfolio value of one test round of all RL trained agents:

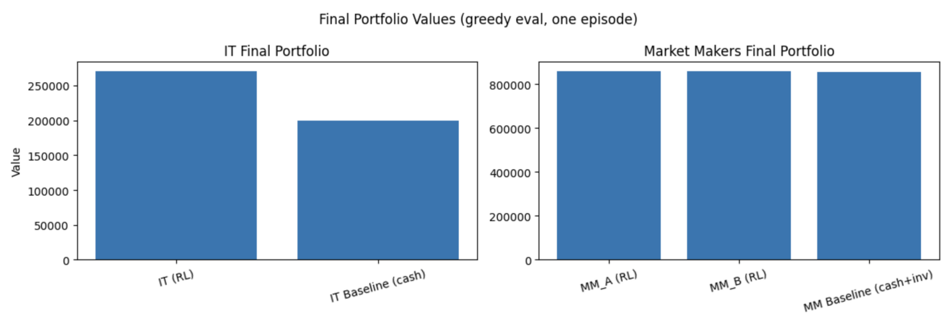

Figure 4.3: Portfolio Value

From the terminal portfolio value plot, we introduce a baseline "cash-and-hold" benchmark for both the informed trader and the market makers. Under this benchmark, each agent simply retains its initial cash and inventory positions and takes no actions until the end of the episode; the corresponding marked-to-fundamental portfolio value serves as a non-strategic reference point. We then compare this baseline to the realized terminal portfolio values of the informed trader, MM_A, and MM_B under their learned policies.

The results indicate that the informed trader is able to generate positive profits relative to the baseline, exploiting the advantage provided by its private signal. In contrast, the market makers (MM_A and MM_B) are largely unable to earn significant profits. The presence of the informed trader introduces adverse selection: market makers face a constant risk of trading against a better-informed agent, which reduces their expected gains from spread capture and inventory management. Despite their ability to post limit orders and hold larger inventories, the informational advantage of the IT constrains the profitability of the market makers.

Moreover, there is no evidence of collusive behavior between MM_A and MM_B (see detail in Section 4.3). On the contrary, competition between the two market makers erodes potential profits: each market maker's attempt to attract order flow and manage inventory leads to tighter spreads and offsetting trades that cancel out gains. This competitive interaction ensures that any surplus created by other participants, including the informed trader, is partially absorbed, leaving market makers with minimal net profit. The observed dynamics highlight how RL market makers, when competing under informational asymmetry, may struggle to generate positive returns.

4.3 Market Maker Trading Behavior

Figure 4.4 illustrates the learned price adjustments, i.e. lean, and half-spreads of the two market makers as functions of inventory and bias from the signal. Several features of these policy maps indicate competitive behavior rather than collusion.

First, the two market makers exhibit different lean policies across identical states. For MM_A, its strategy is almost inventory dependent: small inventories are associated with downward reservation-price shifts (leaning toward buying), while large inventories induce upward shifts (leaning toward selling). In contrary, MM_B's policy reflects a more conservative strategy as there exist regions in which MM_B raises its reservation price despite low inventory, indicating that adverse-selection risk or competitive undercutting outweighs the incentive to rebuild inventory. Such asymmetric, state-contingent behavior is characteristic of competitive liquidity provision under inventory risk and is inconsistent with tacit collusion, which would require synchronized pricing strategies. The heatmap of lean differences demonstrates that MM_A and MM_B tend to adjust prices in opposite directions in many states more clearly, indicating independent inventory management and active undercutting, rather than coordinated price setting.

Secondly, looking at the middle action heatmaps of half-spreads, we can see that spreads tighten near inventory-neutral regions and widen only when inventory positions become extreme. The pattern reflects competitive pressure to provide liquidity when risk is low, with spreads expanding only to compensate for inventory and adverse-selection risk. In contrast, collusive market makers would maintain uniformly wide spreads even in low-risk states.

Taken together, these policy maps provide clear evidence of a competitive equilibrium. Although market makers anticipate informed trading, competition prevents coordinated collusion and drives expected profits toward zero, consistent with standard principles of competitive market making.

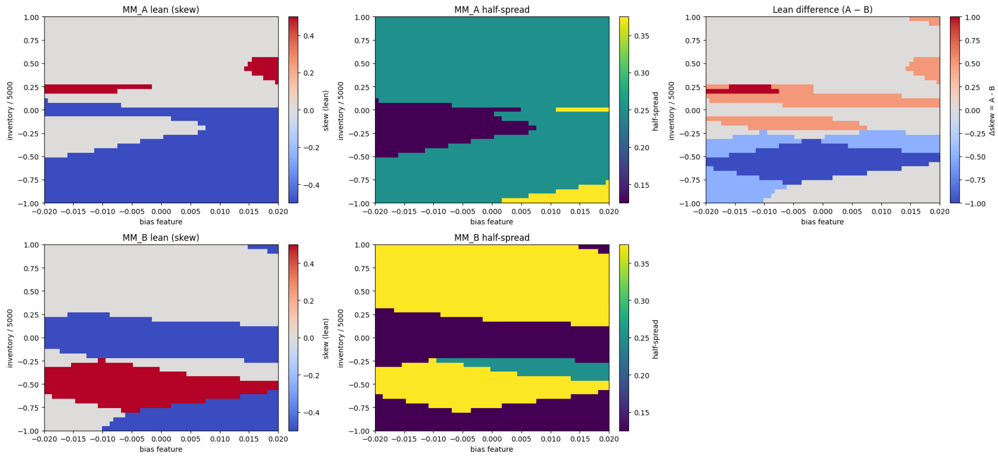

Figure 4.4: MM Action Heatmap

4.4 Informed Trader Trading Behavior

In this section, we try to summarized the informed trader behavior.

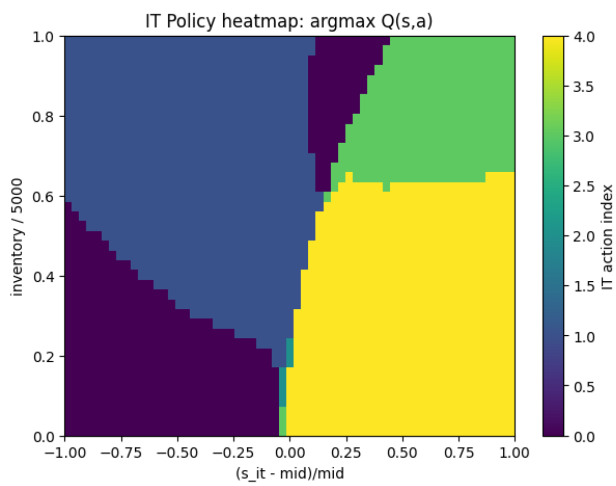

Figure 4.5: IT Action Heatmap

In the above heatmap, the x-axis represents the standardized market mispricing, defined as the difference between the informed trader's signal and the current mid price. Values on the left indicate that the asset is overpriced relative to the signal, whereas values on the right indicate underpricing. The y-axis represents the standardized inventory of the informed trader. The heatmap coloring reflects trading intensity: lighter (yellow) colors correspond to stronger buy actions, while darker (purple) colors indicate stronger sell actions.

The heatmap reveals a nearly vertical decision boundary around zero signal difference, suggesting that the trader's directional bets are primarily driven by the signal. Specifically, when the asset is overpriced relative to the signal, the trader tends to sell; when underpriced, the trader tends to buy.

Trading quantity, however, is also influenced by the informed trader's inventory. Since the total number of shares in the market is fixed, a high inventory for the informed trader implies that the market makers hold less inventory, which in turn can result in less favorable pricing. This introduces a trade-off between executing a desired order and obtaining a favorable price. The heatmap shows that when the informed trader holds a large inventory, trading actions become more conservative, reflecting this trade-off.

Finally, the decision boundary is not perfectly linear. This is expected given that the policy is learned through a deep Q-network, whose nonlinear function approximation cannot be fully summarized in a simple two-dimensional plot.

Overall, informed trading is state-dependent and contributes to price discovery while leaving a limited footprint in standard public market statistics.

4.5 Market Efficiency Test Using an Arbitrageur-Style Predictor

We then evaluate whether publicly observable limit-order-book information contains systematic predictive power for the terminal fundamental value. To do this, we introduce an arbitrageur-style neural predictor: a feedforward neural network that observes the market but does not trade and therefore does not affect prices. Its sole purpose is to test for informational inefficiencies.

At each time t, the predictor receives the public state

$$s_t = \left[(\text{ask}_t - \text{bid}_t)/\text{mid}_t,\ \text{mid}_t,\ t/(T-1),\ \text{spread}_t\right],$$

and is trained to minimize mean squared error when forecasting the terminal payoff:

$$\min_\theta\ \mathbb{E}\left[(\hat{X}_\theta(s_t) - X)^2\right],$$

where $\hat{X}_\theta(s_t)$ denotes the model's prediction.

Formally, let $\mathcal{F}_t^{\text{public}}$ denote the public information set at time t, generated by the limit order book (quotes, depth, etc.). In a Kyle-style informationally efficient equilibrium,

$$p_t = \mathbb{E}[X \mid \mathcal{F}_t^{\text{public}}],$$

which implies that no measurable function of public information should systematically forecast deviations between the market price and the fundamental value. The scatter plot of prediction error in the test round and the histogram are plotted below:

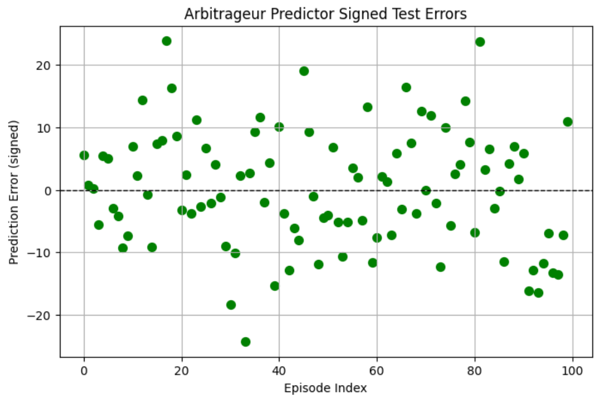

Figure 4.6: Prediction error

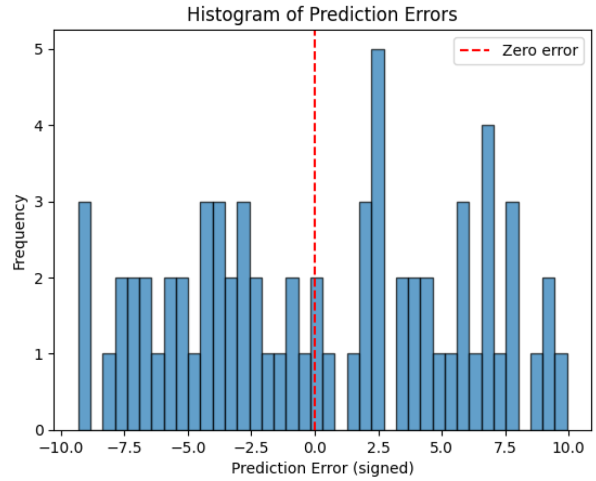

Figure 4.7: Histogram of prediction errors

Empirically, prediction errors are centered tightly around zero and show no statistically significant dependence on public state variables. It is somewhat uniformly distributed. Regressions of errors on observable features yield insignificant coefficients. These findings suggest that although prices reflect information on average, publicly available data do not allow precise recovery of the terminal payoff beyond what is already embedded in market prices.

4.6 Detecting Informed Trading from Public Quotes

We next examine whether the presence of informed trading can be detected from publicly observable limit-order-book features. For this purpose, we train a binary classifier that observes only public information and predicts whether there is informed trading present in the episodes.

The classifier observes a feature vector

$$z_t = \phi(\mathcal{F}_t^{\text{public}}),$$

where $\phi(\cdot)$ maps public quotes into a finite-dimensional representation including spreads, mid-price levels, short-horizon returns, depth imbalances, and quote slopes.

Each episode receives a label

$$y = \begin{cases} 1, & \text{if the informed trader's signal is correlated with } X, \\ 0, & \text{if the informed trader's signal is independent of } X, \end{cases}$$

where X denotes the terminal fundamental payoff.

At the start of each episode, we draw a fundamental payoff X from the same payoff distribution used throughout the environment. In the informative regime ($y = 1$), the informed trader receives a noisy private signal centered on X,

$$s^{IT} \mid X \sim \mathcal{N}(X, \sigma_{IT}^2),$$

so that $\mathrm{Cov}(s^{IT}, X) > 0$.

In the uninformative regime ($y = 0$), we generate a surrogate payoff $\tilde{X}$ independently from the same payoff distribution,

$$\tilde{X} \sim \mathcal{D}_X, \qquad \tilde{X} \perp\!\!\!\perp X,$$

and we draw the trader's signal around this surrogate payoff,

$$s^{IT} \mid \tilde{X} \sim \mathcal{N}(\tilde{X}, \sigma_{IT}^2).$$

By construction, s^{IT} is independent of X under $y = 0$: $\mathrm{Cov}(s^{IT}, X) = 0$, and $\mathbb{E}[s^{IT} \mid X] = \mathbb{E}[s^{IT}]$. Intuitively, the uninformed signal has the same marginal distribution and noise level as an informative signal, but it carries no information about the true payoff X.

A key identification feature of our design is that the informed trader follows the same trained policy in both regimes. Only the information content of the private signal differs across episodes. Thus, any detectability must arise from differences in market outcomes generated by private information, rather than from mechanical differences in trading behavior.

The classifier estimates a conditional probability

$$\hat{p}(z_t) = \mathbb{P}(y = 1 \mid z_t),$$

using a neural network trained with the binary cross-entropy objective

$$\min_{\theta} \ \mathbb{E}\big[-y \log \sigma(f_\theta(z_t)) - (1 - y) \log\big(1 - \sigma(f_\theta(z_t))\big)\big],$$

where f_θ denotes the classifier logit and $\sigma(\cdot)$ is the logistic function.

Out-of-sample performance is essentially indistinguishable from random guessing. Classification accuracy is near 50%, and the area under the ROC curve is approximately 0.5. These results imply that

$$y \perp\!\!\!\perp \mathcal{F}_t^{\text{public}},$$

that is, the presence of private information is statistically independent of public order-book states.

A natural alternative design would be to remove the informed trader entirely in the uninformative regime. We also experiment with this environment by setting informed trader's market order quantity to 0 in all episodes, completely deleting informed trader's presense in the market. It's worth noticing that the average accuracy and AUC are still around 50%. The fact that these two environments—one with active but uninformative trading and one with no informed trader participation at all—are equally indistinguishable from the informative regime highlights a deeper equilibrium property of the market. In both cases, competitive market makers optimally adjust quotes based on observed order flow and public state variables, internalizing any informational content into prices while neutralizing its detectability in public order-book statistics.

When the informed trader is active but uninformative, its trades are statistically indistinguishable from noise and are absorbed by market makers without generating systematic patterns in LOB dynamics. When the informed trader is absent, aggregate order flow is reduced, but market makers optimally respond by tightening spreads and adjusting liquidity provision so that the resulting public order-book states remain similar. In equilibrium, both environments therefore generate order books that are observationally equivalent from the perspective of public information. This implies that the failure to detect informed trading is not driven by the specific implementation of the uninformed regime, but by the equilibrium behavior of competitive market makers: regardless of whether informed trading is absent or present but uninformative, liquidity provision internalizes order-flow information into prices while eliminating residual statistical signatures that could reveal the presence of private information. As a result, public quotes reflect expectations of fundamentals conditional on order flow, but do not encode whether that flow was generated by informed or uninformed motives.

Taken together, the consistently random classification performance across both settings provides strong evidence of informational efficiency: private information, when present, is incorporated into prices, yet its presence cannot be inferred from public limit-order-book data alone.

4.7 Summary

The dual market-maker environment converges to a competitive equilibrium characterized by efficient price discovery, zero expected profits for liquidity providers, and positive but bounded profits for informed traders. Public order-book information neither allows accurate prediction of fundamentals nor reveals the presence of informed trading, illustrating how strategic interaction between informed traders and competitive market makers endogenously generates informational efficiency.

Chapter 5

Conclusion

The project studies how reinforcement-learning agents interact in a limit-order-book market with asymmetric information. By progressively enriching the market structure—from a single rule-based market maker facing an RL-informed trader to a fully competitive environment with two RL market makers—we isolate the roles of learning, competition, and private information in shaping price dynamics and market efficiency.

In the single–market-maker setting, the informed trader learns to exploit the maker's mechanical quoting rule by conditioning on its private signal, recent price movements, and inventory constraints. Prices become more informative over time but remain imperfect, allowing the informed trader to earn informational profits. The learned strategy exhibits order-hiding behavior and nonlinear state dependence, closely mirroring predictions from classical Kyle-style models.

When two reinforcement-learning market makers compete, the market converges to a markedly different equilibrium. Market makers adopt state-contingent quoting policies that aggressively manage inventory risk while undercutting each other to attract order flow. This competition erodes expected market-maker profits and eliminates evidence of tacit collusion. At the same time, prices track the informed trader's signal closely and incorporate private information rapidly, indicating a high degree of informational efficiency.

We further test market efficiency using two complementary approaches. First, an arbitrageur-style neural predictor that observes only public limit-order-book information is unable to forecast the terminal payoff beyond what is already embedded in prices. Second, a classifier trained to detect the presence of informed trading from public quotes performs no better than random guessing. These failures persist even when the informed trader is entirely removed in the uninformed regime. Together, these results imply that public order-book states are observationally equivalent across regimes once market makers optimally internalize order flow.

Overall, the findings support a strong notion of Kyle-style informational efficiency: private information is incorporated into prices through strategic trading and competitive liquidity provision, yet it leaves no detectable statistical footprint in public order-book data. Reinforcement learning provides a flexible framework in which such equilibria emerge endogenously, without imposing linear strategies or rational-expectations assumptions.

Chapter 6

Limitations and Future Work

While the reinforcement-learning framework provides a flexible environment for studying market microstructure with asymmetric information, several limitations remain.

First, the market structure is intentionally stylized. The environment features a single asset, a finite trading horizon, and a small number of strategic agents. Although this design allows for clean identification of informational and competitive effects, it abstracts away from many institutional features of real markets, such as heterogeneous trader arrival times, multiple assets, and persistent order-flow autocorrelation across episodes. Extending the framework to richer market settings would help assess the robustness of the equilibrium outcomes documented here.

Second, the informed trader's information structure is deliberately simple. Private information enters through a single noisy signal about the terminal payoff, and the informed trader follows a fixed, trained policy across regimes. While this choice is essential for isolating informational effects in the detection experiments, it limits the scope of strategic information acquisition and learning. Future work could allow informed traders to adapt their trading intensity dynamically based on signal strength, volatility, or perceived market-maker behavior.

Third, although market makers learn state-contingent quoting strategies, their objective function is limited to short-horizon profit and inventory management. Incorporating alternative objectives, such as risk preferences and capital constraints, could generate richer liquidity dynamics and potentially alter the degree of informational efficiency.

Several natural extensions follow from these limitations. One direction is to introduce additional market makers with heterogeneous speeds, inventories, or learning algorithms, allowing entry and exit over time. Such extensions would permit a systematic study of how competition intensity affects spreads, depth, and price efficiency. Another direction is to incorporate RL-based arbitrageurs that attempt to exploit short-horizon mispricings or cross-agent statistical patterns, providing a dynamic stress test of semi-strong efficiency in real time.

Finally, the framework can be extended to study regime uncertainty. Rather than fixing the informed trader's information structure exogenously, future environments could randomize the presence and quality of private information across episodes, leaving market makers uncertain about whether informed trading is occurring. This would allow a direct analysis of whether market makers can infer informational regimes from order flow alone and how such uncertainty affects equilibrium pricing and liquidity.

Together, these extensions would deepen our understanding of how learning, competition, and private information jointly shape market efficiency, while preserving the core insight of this paper: competitive liquidity provision internalizes information into prices without leaving detectable signatures in public order-book data.

Bibliography

Dou, W. W., I. Goldstein, and Y. Ji (2025). AI-Powered Trading, Algorithmic Collusion, and Price Efficiency. *NBER Working Paper* No. 34054, National Bureau of Economic Research, Cambridge, MA. doi:10.3386/w34054.

Calvano, E., G. Calzolari, V. Denicolò, and S. Pastorello (2020). Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review* 110(10), 3267–3297. doi:10.1257/aer.20190623.

Kyle, A. S. (1985). Continuous Auctions and Insider Trading. *Econometrica* 53(6), 1315–1335. doi:10.2307/1913210.