Market Making with Learned Beta Policies

Yongzhao Wang
University of Liverpool
United Kingdom
The Alan Turing Institute
United Kingdom
wangyzh@liverpool.ac.uk

Rahul Savani
University of Liverpool
United Kindom
The Alan Turing Institute
United Kingdom
rahul.savani@liverpool.ac.uk

Anri Gu
University of Michigan
US
anrigu@umich.edu

Chris Mascioli
University of Michigan
US
cmasciol@umich.edu

Theodore Turocy
University of East Anglia
United Kingdom
The Alan Turing Institute
United Kingdom
t.turocy@uea.ac.uk

Michael Wellman
University of Michigan
US
wellman@umich.edu

Abstract

In market making, a market maker (MM) can concurrently place many buy and sell limit orders at various prices and volumes, resulting in a vast action space. To handle this large action space, *beta policies* were introduced, utilizing a scaled beta distribution to concisely represent the volume distribution of an MM's orders across different price levels. However, in these policies, the parameters of the scaled beta distributions are either fixed or adjusted only according to predefined rules based on the MM's inventory. As we show, this approach potentially limits the effectiveness of market-making policies and overlooks the significance of other market characteristics in a dynamic market. To address this limitation, we introduce a general adaptive MM based on beta policies by employing deep reinforcement learning (RL) to dynamically control the scaled beta distribution parameters and generate orders based on current market conditions. A sophisticated market simulator is employed to evaluate a wide range of existing market-making policies and to train the RL policy in markets with varying levels of inventory risk, ensuring a comprehensive assessment of their performance and effectiveness. By carefully designing the reward function and observation features, we demonstrate that our RL beta policy outperforms baseline policies across multiple metrics in different market settings. We emphasize the strong adaptability of the learned RL beta policy, underscoring its pivotal role in achieving superior performance compared to other market-making policies.

Keywords

Market Making, Deep Reinforcement Learning, Beta Policy

ACM Reference Format:
Yongzhao Wang, Rahul Savani, Anri Gu, Chris Mascioli, Theodore Turocy, and Michael Wellman. 2024. Market Making with Learned Beta Policies. In *5th ACM International Conference on AI in Finance (ICAIF '24), November 14–17, 2024, Brooklyn, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3677052.3698623

ICAIF '24, November 14–17, 2024, Brooklyn, NY, USA

ACM ISBN 979-8-4007-1081-0/24/11
https://doi.org/10.1145/3677052.3698623

1 Introduction

In a limit-order driven market, a market maker (MM) operates in both sides of the market, consistently offering to buy and sell to facilitate trading. By constantly presenting in the market, MMs provide continuous liquidity, enabling immediate trading at prices that fairly represent current market conditions. The primary objective of an MM is to profitably capture a *spread* by engaging in transactions on both sides. A significant challenge for MMs lies in managing *inventory risk* associated with trading against better-informed traders. This situation exposes MMs to *adverse selection*, where counterparties exploit informational or technological advantages during transactions. As a result, MMs may accumulate a positive or negative inventory position, which could lead to losses when adverse price movements occur, such as when the MM has net sold to the market just before a significant price increase. Despite the risks, profitable MMs exist across many types of markets and they are generally recognized to be beneficial for stabilizing prices and aiding in the discovery of accurate market prices.

At any given time, an MMs' orders typically cover a range of price levels, with varying volumes at each price level, so a comprehensive representation is usually high dimensional. To represent the MMs' orders in a succinct manner, Jerome et al. [20] introduced a general policy representation known as the *scaled beta policy*. A scaled beta policy utilizes scaled beta distributions to determine the volume profiles of bids (buys) and asks (sells) placed by an MM. The shape of a scaled beta distribution can be controlled precisely by its parameters, making the scaled beta policy a generalization of many existing market-making policies, including single price-level policy, the ladder policy, and "market making at the touch". Figure 1 demonstrates the application of scaled beta distributions to describe order volume profiles across various price levels. Given a fixed total volume 100 and a fixed number of price levels 5, bid volumes in this example were derived from a scaled beta distribution with $\alpha = 2$ and $\beta = 2$, and ask volumes from another scaled beta distribution with $\alpha = 1$ and $\beta = 2$.

Figure 1: A sample limit order book, demonstrating the distribution of orders modeled by scaled beta distributions.

However, in [20] the parameters of the scaled beta distribution are either fixed or adaptively determined based solely on predefined rules and the MM's inventory. We show that this potentially limits the flexibility of market-making policies and overlooks the significance of other market characteristics in an evolving market. To address this limitation, we propose a deep reinforcement learning (RL) market-making policy for an MM using the beta policy representation. This removes the constraints on the market-making policy's functional form and enables the inclusion of various market features, which are essential for the adaptability required in varying market conditions. Specifically, we treat the parameters of the scaled beta distributions as actions and use deep RL to learn a policy for determining these actions. At each time step, the RL market-making policy takes in observations of the current market conditions and outputs the parameters for the scaled beta distributions as actions. These parameters then dictate the volumes of bids and asks that the MM submits at different price levels.

We evaluate the performance of the RL market-making policy in two market settings: one with informed background traders and one without, comparing with a comprehensive collection of existing market-making policies from the literature. We first show that there is no single static MM policy that performs well across all the market settings, highlighting the need for adaptive strategies. Then we find that the presence of informed background traders can lead to significant profit losses for MMs and incorporating inventory risk control methods into the policies can mitigate these losses. With the RL market-making policy, we show that the MM can effectively adjust volume distributions in response to current market conditions, consistently outperforming the baseline policies. Interestingly, we find that the optimal policy against the adverse selection may not be minimizing inventory but rather adapting swiftly by broadening its trading spread when risks arise and narrowing the spread when market conditions become more certain.

The contributions of this work include:

(1) We develop an RL market-making policy built on the scaled beta distribution, enabling it to adjust volume distributions according to current market conditions.
(2) We evaluate the performance of the RL market-making policy by comparing it against a variety of existing market-making policies across two different market settings: one with informed background traders and one without. Our results demonstrate that the RL market-making policy consistently outperforms the baseline policies.
(3) We reveal the impact of adverse selection on different market-making policies and identify the key factors that contribute to the success of these policies in various market settings.

2 Related Work

2.1 MM Policies

Single Price-Level Policy. A natural two-dimensional action space for the MM involves selecting two half-spreads, namely a bid and ask offset from the midprice. In the literature, this setup typically assumes that all orders have a constant volume. The MM adjusts these half-spreads at each time step based on the market conditions and their inventory. This method of choosing half-spreads is the primary approach in financial stochastic control literature on market making [3, 6, 8, 10, 13, 17].

While the financial stochastic control literature often utilizes continuous models with corresponding continuous half-spreads, many simulators for RL operate in a discrete setting. In these simulators, the problem involves choosing the number of ticks away from the touch (the best bid and ask prices) at which to quote the bid and ask. Note that the action space, being the product space of bid and ask actions, can become very large unless the MM is restricted to placing actions very close to the best prices.

The first application of RL to market making, by Chan and Shelton [12], used such a policy. To address the issue of the large action space, they chose to adjust their quotes by increasing or decreasing them from a small set of actions. Subsequently, Kim et al. [22] fitted an input-output hidden Markov model to order data from Nasdaq and used RL to determine actions within the model. They allowed their MMs to increase, decrease, or maintain their bid, ask, and both associated volumes by at most one tick or unit of the asset.

More recently, Spooner et al. [32] utilized a realistic market simulator incorporating five levels of order book data and transactions. The action space in their study consists of a collection of pre-specified half-spread pairs, along with an action that clears the entirety of the MM's inventory using a market order. Note that some of these actions are skewed to favor filling on one side. This approach allows for a basic form of inventory control, while maintaining a manageable action space size. This paper was the first to use such a finite pre-specified selection of actions, a method that has since been adopted by many other works [25, 30, 40].

Marin et al. [26] studied the application of RL to the model by Avellaneda and Stoikov [3]. Instead of directly determining the limit order to place, they employed the RL algorithm to adjust the risk aversion parameter and skew the quotes given by the Avellaneda-Stoikov algorithm according to recent market activity trends. Spooner and Savani [33] explored a robust version of the model by Avellaneda and Stoikov [3], where an adversarial market agent controls the drift of the financial market. In this model, the action space consists of four continuous parameters that control the mean and variance of the agent's bids and asks. The parameters are learned by approximating the value function using cubic polynomials and then performing least squares policy iteration [23]. The same model has also been studied by Nyström et al. [28].

A final form of action space within the single price-level category involves selecting a continuous half-spread on each side of the book and then quantizing it to submit orders on the price grid. This approach was employed by Gasperov and Kostanjcar [15], who used neuro-evolution to train a policy represented by a deep neural network.

Ladder Policy. Another relevant strand of the existing literature was initiated by Chakraborty and Kearns [11], who introduced and studied ladder policies. These policies place a unit of volume at all prices within two price intervals, one on each side of the book. Chakraborty and Kearns [11] theoretically proved the utility of these policies in mean-reverting markets with Ornstein-Uhlenbeck price dynamics.

Inspired by this work, Abernethy and Kale [1] considered related order placement policies where limit orders for one unit of volume were placed at all price levels outside a window around the midprice, defining the MM's spread. They presented an online learning scheme that adjusts between parameterizations of their ladder policies, guaranteeing competitive performance with the best parameter choice in hindsight.

Wah et al. [35] explored how market making influences market performance, focusing on allocative efficiency and the payoffs from trades received by background traders. They analyzed a parameterized ladder policy for MMs, with its parameters equilibrated through *empirical game-theoretic analysis* [38, 39].

Market-Making at the Touch. "Market-making at-the-touch" refers to submitting limit buy or sell orders at the best bid (i.e., the highest buy price) or the best ask (i.e., the lowest sell price). In the continuous setting, it has been studied by Cartea et al. [7, 9], using a continuous time and space mathematical model of the market. In the RL literature, this approach was adopted by Zhong et al. [41]. In their model, the MM's actions involved choosing whether or not to place an order at both the bid and ask sides of the book at each time step. By discretizing the observation space similarly to the work by Spooner et al. [32], RL algorithms such as Q-learning can be applied.

Beta Policy. Jerome et al. [20] introduced a parametric representation of order volume profiles using scaled beta distributions, developing the scaled beta policy with fixed distribution parameters. This representation significantly reduces the dimension of representing an MM's orders, creating the opportunity of applying RL to determine the MM's orders. Additionally, they extended this representation to create an inventory-driven approach that dynamically adjusts the parameters of the distributions to minimize the inventory risk based on the MM's current inventory. This policy specifically manages to increases sales as the MM's inventory grows and boosts purchases as the inventory diminishes.

2.2 Market Simulation

Since the early 1990s, researchers have used simulation techniques to explore financial markets. Early studies utilizing the Santa Fe Artificial Stock Market [29] initiated a branch of agent-based finance research [24], which generated numerous insights for financial modeling.

PyMarketSim [27] is a Python reimplementation and enhancement of MarketSim, originally developed in Java by Elaine Wah for an agent-based examination of latency arbitrage [34]. MarketSim featured a discrete-event scheduler and a modular structure designed to flexibly accommodate diverse trading strategies and market mechanisms, built around an efficient order book architecture. The initial version was applied to studies on topics such as spoofing [36], welfare impacts of market making [35], and benchmark manipulation [31]. The updated version also supports training of novel trading strategies using deep reinforcement learning and the integration of trained agents into the simulation.

ABIDES (Agent-Based Interactive Discrete Event Simulation) [5] is similarly grounded in discrete-event processing and structured to facilitate the flexible integration of agent trading strategies. ABIDES additionally uses a uniform message-passing system aligned with standard market protocols and has gained broad adoption within the research community, particularly with support from JP Morgan AI Research. An enhanced version, ABIDES-Gym [2], includes an interface compatible with the OpenAI Gym environment for dRL.

Alternatively, simulations can focus on interaction with a limit order book (LOB) governed by an external order process. This approach enables backtesting of trading strategies with historical data or the use of a mathematical LOB model [19]. Recently, Frey et al. [14] introduced JAX-LOB, a simulator that leverages GPU computation through JAX [4] libraries. Jerome et al. [21] also developed a LOB simulator with interfaces designed to facilitate dRL agent training.

3 Preliminaries

3.1 Scaled Beta Distribution

As illustrated in Figure 1, we employ the scaled beta distribution for describing the volume profiles for bids and asks. The scaled beta distribution is a rescaling of the beta distribution, where the scaling is done to hit a desired total volume of orders, distributed across tick-based price levels. Specifically, the probability density function of a beta distribution with parameters $\alpha > 0$ and $\beta > 0$ can be written as

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\,, \tag{1}$$

where $x \in [0,1]$ and

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt\,.$$

Since the beta distribution is defined over the interval $[0,1]$, a variable transformation is required to extend its domain to encompass multiple price levels (i.e., the prices at which the market maker will place orders). Denote the number of price levels on one side of the limit order book as *n_levels*. Define a new variable $y = n\_levels \cdot x$ and therefore $y \in [0, n\_levels]$. By changing the variable x in Equation 1, we have

$$g(y) = \left|\frac{dx}{dy}\right| \cdot f\left(\frac{y}{n\_levels}\right) = \frac{1}{n\_levels} \cdot f\left(\frac{y}{n\_levels}\right) \tag{2}$$

which is the probability density function of the scaled beta distribution with support $[0, n\_levels]$. We describe how these two distributions determine the volume profiles later in Section 4.

3.2 Market Simulator

Our agent-based market simulator [27] employs a Continuous Double Auction mechanism over a single security with discrete time steps and trades occurring over a finite horizon T, following the prior works [35, 37]. Agents in the simulator submit limit orders, which specify the maximum (or minimum) price at which they are willing to buy (or sell) along with the number of units to trade. The *fundamental value*, denoted as r, evolves dynamically over the course of the simulation, representing the intrinsic economic value of the security. It follows a mean-reverting stochastic process:

$$r_t = \max(0, \kappa\bar{r} + (1-\kappa)r_{t-1} + u_t), \quad r_0 = \bar{r}\,, \tag{3}$$

where r_t is the fundamental value at time $t \in [0, T]$. The parameter $\kappa \in [0, 1]$ denotes the mean-reversion strength parameter, which defines the degree to which the fundamental value reverts back to the mean $\bar{r}$. The incremental shock in the fundamental value at time t is sampled from a zero-mean Gaussian distribution: $u_t \sim N(0, \sigma_s^2)$ with some variance σ_s^2.

3.3 ZI Agents as Background Traders

3.3.1 Valuation Model. Background agents represent traditional investors with preferences for holding either long or short positions in the underlying security. In our simulator, we consider Zero Intelligence (ZI) agents, as described by Gode and Sunder [16], which enter the market with an equal probability of being assigned to buy or sell.

The value of a ZI agent's portfolio at the end of a simulation T is a sum of its cash and its holdings' value. The holdings' value is based on the agent's private values, which are agent-specific factors that affect a security's value such as liquidity requirements, and the liquidation at the final fundamental value r_T. Specifically, the private values for ZI agent i can be represented by a vector Θ_i denoting differences in private benefits of trading given the trader's net position. The vector is of size $2q_{max}$, where $q_{max} > 0$ is the maximum number of units the agent can be long or short at any time, with

$$\Theta_i = (\theta^{-q_{max+1}}, \ldots, \theta^0, \theta^{+1}, \ldots, \theta^{q_{max}})\,.$$

Element θ_i^q is the incremental private benefit obtained from selling one unit of the security given current position q, where positive (negative) q indicates a long (short) position. Similarly, θ_i^{q+1} is the i marginal private gain from buying an additional unit given current net position q.

We generate θ_i^q from a set of $2q_{max}$ values drawn independently from a Gaussian distribution. Let $\hat{\theta} \sim N(0, \sigma_{PV}^2)$ denote one of these drawn values. To ensure that the valuation reflects diminishing marginal utility, that is, $\theta^{q'} \geq \theta^q$ for all $q' \leq q$, we sort the $\hat{\theta}$ and set the θ_i^q to respective values in the sorted list. With the private values, we can define the valuation or payoff of a ZI agent as

$$payoff_{ZI} = positional\_value + cash\,,$$

where the positional value includes the holdings' value liquidated at r_T and the sum of its private values on the current position:

$$positional\_value = r_T \cdot q_i + \begin{cases} \sum_{k=1}^{k=q_i} \theta_i^k & \text{if } q_i > 0 \\ -\sum_{k=q+1}^{k=0} \theta_i^k & \text{if } q_i < 0\,. \end{cases}$$

3.3.2 Trading Policy. ZI agents arrive at the market according to a Poisson process with rate λ_a. On arrival, they are assigned to buy or sell (with equal probability), and accordingly submit an order to buy or sell a single unit. Agents may trade any number of times, as long as their net positions do not exceed q_{max} (either long or short).

At the time of market entry t, a ZI agent can assess its payoff at the end of simulation, using an estimate $\hat{r}_t$ of the terminal fundamental r_T. The estimate is based on the current fundamental r_t with additional Gaussian noise z_t, adjusted to account for mean reversion:

$$\hat{r}_t = (1 - (1-\kappa)^{T-t})\bar{r} + (1-\kappa)^{T-t}(r_t + z_t)\,. \tag{4}$$

The ZI agent then submits a bid shaded from this estimate and the incremental private value by a random offset—the degree of extra payoff it demands from the trade. The amount of shading is drawn uniformly from range $[R_{min}, R_{max}]$, where R_{min} and R_{max} are predefined parameters such that $0 < R_{min} < R_{max}$. Specifically, a ZI trader i arriving or re-entering at time t with current position q submits a limit order for a single unit of the security at price

$$p_i \sim \begin{cases} U\left[\hat{r}_t + \theta_i^{q+1} - R_{max}, \hat{r}_t + \theta_i^{q+1} - R_{min}\right] & \text{if buying} \\ U\left[\hat{r}_t + \theta_i^{q} + R_{min}, \hat{r}_t + \theta_i^{q} + R_{max}\right] & \text{if selling}\,. \end{cases}$$

To expose MMs to adverse selection, we introduce informed ZI agents. Such an agent is assumed to know the true final fundamental value, which can be implemented by sampling all the fundamentals throughout a simulation a priori. This knowledge allows the agent, indexed by j, to be informed about the true value of the security, enabling them to submit bids accordingly at price

$$p_j \sim \begin{cases} U\left[r_T + \theta_j^{q+1} - R_{max}, r_T + \theta_j^{q+1} - R_{min}\right] & \text{if buying} \\ U\left[r_T + \theta_j^{q} + R_{min}, r_T + \theta_j^{q} + R_{max}\right] & \text{if selling}\,. \end{cases}$$

4 RL Beta Policy

To construct MM's orders, an MM should first determine the relevant price levels. For example, in Figure 1, the starting price level for bids is 5.05 and for asks is 5.07, and $n\_levels = 5$. The MM then distributes the total volume according to the scaled beta distribution. In our model, we assume that the number of price levels, *n_levels*, the rung size (i.e., the price gap between two price levels in MM's orders), and the total volume distributed across these levels are predefined parameters.

4.1 Selecting the Starting Price Level

We select the starting price levels of MM's orders based on the estimated final fundamental $\hat{r}_t$ and a pre-specified spread ω. Specifically, we first compute the initial bid and ask price levels, B_t' and S_t', by adding or subtracting half of the spread from the estimated

final fundamental. That is,

$$B'_t = \hat{r}_t - \frac{1}{2}\omega \quad S'_t = \hat{r}_t + \frac{1}{2}\omega \ .$$

The rationale behind this choice is to allow the MM to trade around its estimate of the security's true value, $\hat{r}_t$, profiting from a specified spread, ω. To prevent the initial price levels from crossing the current bid-ask spread, we further adjust B'_t and S'_t by truncating them to the current best bid BID_t (i.e., the highest buy price at time t) and the best ask ASK_t (i.e., the lowest sell price at time t) as follows:

$$B_t = \min(B'_t, ASK_t) \quad S_t = \max(S'_t, BID_t) \ .$$

Starting from B_t and S_t, an MM can construct a sequence of price levels for placing orders, comprised of K rungs, each spaced ξ ticks apart:

$$\begin{cases} [B_t - (K-x)\xi, B_t - (K-x+1)\xi, \ldots, B_t - K\xi] & \text{for bids} \\ [S_t + (K-x)\xi, S_t + (K-x+1)\xi, \ldots, S_t + K\xi] & \text{for asks} \ , \end{cases} \tag{5}$$

with $S_t > B_t$ and predefined $K, \xi > 0$, where $x > 0$ specifies the rung immediately above BID_t (for sell orders) or below ASK_t (for buy orders).

4.2 Distributing the Total Volume

To allocate the total volume to the selected price levels with a scaled beta distribution, an RL beta policy should first determine the parameters for the corresponding beta distribution. As discussed in Section 3.1, a beta distribution can be characterized by the parameters α and β, which will be the action outputs of our learned RL policy. Specifically, at each time step, an action a from an RL policy is defined by a four-tuple, which specifies two scaled beta distributions for bids and asks, respectively.

$$a = (\alpha^{ask}, \beta^{ask}, \alpha^{bid}, \beta^{bid}) \ .$$

We further simplify the action space dimension by assuming that the beta distributions on both sides of the limit order book are symmetric, meaning that $\alpha^{ask} = \alpha^{bid}$ and $\beta^{ask} = \beta^{bid}$. As a result, the action space can be reduced to two dimensions as follows:

$$a = (\alpha, \beta) \ .$$

With the output of the RL policy $a = (\alpha, \beta)$, we can define the beta distributions and subsequently the scaled beta distributions for bids and asks.

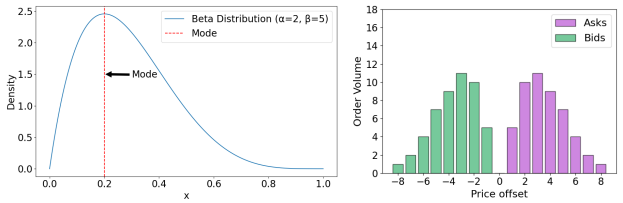
Using the scaled beta distributions, we can allocate the total volume to the selected price levels as described in Equation 5. Conceptually, these price levels can be viewed as bins, with the scaled beta distribution indicating the proportion of total volume assigned to each bin. Consequently, the volume for each price level is calculated by multiplying the total volume by its corresponding proportion. Specifically, if the total number of price levels (viewed as bins) is denoted as *n_levels*, we can index the boundaries of these bins using integers from 0 to *n_levels* (e.g., 0 and 1 are the boundary indices for the first bin). Recall that the scaled beta distribution spans these *n_levels* bins. Thus, we can evaluate the cumulative distribution function $G(y)$ of the scaled beta distribution (can be derived from $g(y)$ in Equation 2) at the boundary indices. The difference between two successive indices represents the proportion corresponding to the respective bin. For example, assuming $x = K$ (i.e., no truncation) in Equation 5, the volume assigned to the first price level with price B_t will be $[G(1) - G(0)] \times total\_volume$. Similarly, the volume $[G(2) - G(1)] \times total\_volume$ will be allocated to the second price level. All volumes will be rounded to the nearest integers.

4.3 Interpreting an Action

As components of an action, α and β are not particularly interpretable, however, they can be mapped to the mode and concentration of the beta distribution, providing a clearer interpretation. Specifically, the concentration κ of a beta distribution is defined as $\kappa = \alpha + \beta$. When $\alpha > 1$ and $\beta > 1$, the mode $\mu = \frac{\alpha-1}{\alpha+\beta-2}$. With κ and μ specified,

$$\alpha = \mu(\kappa - 2) + 1, \beta = (1 - \mu)(\kappa - 2) + 1 \ . \tag{6}$$

By converting α and β to the mode and concentration, the distribution of volumes becomes visually discernible, as shown in Figure 2, making an action more interpretable.



(a) Beta distribution. (b) Volume distributions.

Figure 2: An illustration of the mode in a beta distribution with $\alpha = 2$ and $\beta = 5$. The mode directly corresponds to the highest volume in the volume distributions, mirrored for bids and asks.

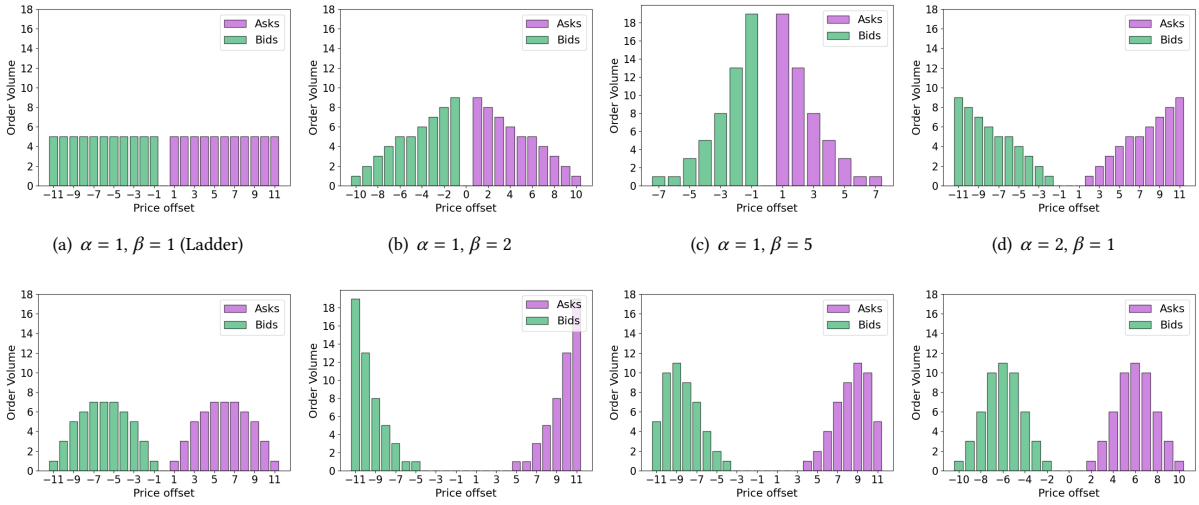
4.4 Observation Space

The specific action that the RL policy outputs depends on the current market conditions, which we refer to as the *adaptability* of the policy. We consider the following conditions as features of the observation space for RL, which are common market summary statistics. These features are normalized to $[0, 1]$ for RL training.

- The number of time steps left $T - t$.
- The current fundamental value r_t.
- The current best bid price BID_t (if any).
- The current best ask price ASK_t (if any).
- The MM's inventory I.
- Mid-price move from $t - 1$ to t.
- Other market statistics including volume imbalance, queue imbalance, volatility, Relative Strength Index, suggested in the work [32].

4.5 Valuation and Rewards

Similar to ZI agents, the MM liquidates its inventory at the end of the trading horizon, with the liquidation price being the final fundamental value, r_T. Unlike ZI agents, the MM does not incorporate private values in its valuation. Therefore, the MM's payoff can be



(a) $\alpha = 1, \beta = 1$ (Ladder) (b) $\alpha = 1, \beta = 2$ (c) $\alpha = 1, \beta = 5$ (d) $\alpha = 2, \beta = 1$

(e) $\alpha = 2, \beta = 2$ (f) $\alpha = 5, \beta = 1$ (g) $\alpha = 5, \beta = 2$ (h) $\alpha = 5, \beta = 5$

Figure 3: Volume Distributions for different α and β combinations.

computed as the sum of its cash and its holdings' value at the end of the simulation T:

$$payoff_{MM} = r_T \cdot q_i + cash. \tag{7}$$

Since the payoff is only realized at the end of a simulation without any intermediate rewards, the training of RL may face the problem of sparse rewards. To tackle this, we design a "period by period" reward function

$$R_{t+k} = (r_T \cdot q_{t+k} + cash_{t+k}) - (r_T \cdot q_t + cash_t),$$

which is the valuation difference between two successive orders of the MM at time step t and $t + k$. Intuitively, the reward at time step $t + k$ is the change in valuation after taking an action at t. Note that this reward function applies reward shaping in hindsight, which is a common reward shaping approach in RL, since r_T is only available at the end of the simulation T. It can be easily shown that these intermediate rewards will sum to the final payoff.

5 Experiment Settings

5.1 Simulation Setup

We evaluate the performance of MMs in a market setting with $N = 25$ ZI agents as background traders who enter the market at a rate of $\lambda_a = 0.075$ and submit a single-unit order. In the market setting with informed background traders, 13 out of 25 regular ZI agents are replaced by the informed ZI agents. The variance for the private value vector σ^2_{PV} and the variance for the noisy estimated final fundamental σ^2_z for all ZI agents (both the regular and the informed ones) are set to be 5×10^6 and 1×10^6, respectively. The maximum number of units that ZI agents can hold, either long or short, at any time is $q_{max} = 10$. The amount of shading is drawn uniformly from the range $[250, 500]$.

The MM submits orders at a rate of $\lambda_{MM} = 0.005$. At each entry, the MM submits orders with a fixed total volume 100 distributed across 21 price levels with rung size 50 on each side of the limit order book. The volume distribution is determined by the MM's policy.

Each simulation runs for $T = 1 \times 10^5$ time steps. The fundamental value follows a mean-reversion process with a mean of $\bar{r} = 1 \times 10^5$ and a parameter $\kappa = 0.05$. The minimum tick size is fixed at 1. We interpret the tick size as one-thousandth of a dollar ($0.001), so the mean fundamental value corresponds to $100. To account for the stochastic nature of the simulations, including fluctuations in market fundamentals, variations in agent arrival rates, and diverse private valuations, we average all the results over 2000 simulations during testing.

5.2 Baseline MMs

We explore four types of market-making policies as baselines, each with distinct configurations: MMs at the touch, ladder MMs, fixed-parameter beta MMs, and inventory-driven beta MMs. MMs at the touch place orders with the entire volume of 100 at the best bid and ask prices. Ladder MMs distribute the total volume evenly across predefined price levels. Beta MMs allocate the volume according to scaled beta distributions with fixed parameters. For ladder MMs and beta MMs, we vary the spread $\omega \in \{10, 60\}$, which affects the first price level to place orders as well as their payoffs. For beta MMs, we conduct parameter optimization over a set of Cartesian product of $\alpha \in \{1, 2, 3, 4, 5\}$ and $\beta \in \{1, 2, 3, 4, 5\}$ and selectively report some combinations that are representative in performance. Note that when $\alpha = 1$ and $\beta = 1$ for both bids and asks, beta MMs reduce to ladder MMs (we distinguish between ladder MMs and beta MMs for clarity). We illustrate how the volume distributions look like in Figure 2(b) and Figure 3 with a total volume of 50 and 11 price levels.

The inventory-driven beta-based policy by Jerome et al. [20] adjusts the volume distributions based on the MM's current inventory to minimize the inventory risk. Instead of directly computing α and β, it computes the modes, μ^{bid} and μ^{ask}, of the volume distributions

Index	Policy Type	ω	α	β	Averaged Payoff	Averaged Spread	Market Share
1	**MM at the touch**	-	-	-	1.41×10^6	386	22.8%
2	**Ladder MM**	10	-	-	4.34×10^6	330	24.9%
3	**Beta MM**	10	1	2	4.04×10^6	271	29.0%
4	**Beta MM**	10	1	5	3.25×10^6	196	33.8%
5	**Beta MM**	10	2	1	4.49×10^6	452	15.7%
6	**Beta MM**	10	2	2	4.76×10^6	378	21.3%
7	**Beta MM**	10	2	5	4.41×10^6	284	28.2%
8	**Beta MM**	10	5	1	3.00×10^6	574	6.4%
9	**Beta MM**	10	5	2	3.77×10^6	534	9.3%
10	**Beta MM**	10	5	5	4.82×10^6	455	16.3%
11	**Invt MM**	10	-	-	4.08×10^6	278	28.6%
12	**RL MM**	10	-	-	$\mathbf{5.08 \times 10^6}$	424	17.9%

Table 1: The performance of MMs without informed ZI agents.

Index	Policy Type	ω	α	β	Averaged Payoff	Averaged Spread	Market Share
1	**MM at the touch**	-	-	-	7.95×10^6	422	18.1%
2	**Ladder MM**	10	-	-	-9.00×10^7	569	27.2%
3	**Beta MM**	10	1	2	-9.60×10^7	560	28.9%
4	**Beta MM**	10	1	5	-1.00×10^8	505	30.9%
5	**Beta MM**	10	2	1	-6.43×10^7	628	20.9%
6	**Beta MM**	10	2	2	-8.44×10^7	608	25.3%
7	**Beta MM**	10	2	5	-9.01×10^7	536	28.8%
8	**Beta MM**	10	5	1	-5.09×10^7	702	13.3%
9	**Beta MM**	10	5	2	-5.54×10^7	674	15.8%
10	**Beta MM**	10	5	5	-7.28×10^7	648	21.8%
11	**Invt MM**	10	-	-	-4.18×10^7	699	12.7%
12	**RL MM**	10	-	-	$\mathbf{-1.47 \times 10^7}$	542	11.3%

Table 2: The performance of MMs with informed ZI agents.

on both sides of the book and the concentration κ. Then it converts them to $\alpha^{bid}, \beta^{bid}, \alpha^{ask}, \beta^{ask}$ using Equation 6. The policy takes the following form

$$f_1(\text{inv}_t) := \mu_0 \left[1 + \left(\frac{1}{\mu_0} - 1\right) \text{clamp}\left(\left|\frac{\text{inv}_t}{\text{max\_inv}}\right|^p\right)\right],$$
$$f_2(\text{inv}_t) := \mu_0 \left[1 - \text{clamp}\left(\left|\frac{\text{inv}_t}{\text{max\_inv}}\right|^p\right)\right],$$

$$\mu^{bid}(\text{inv}_t) := \mathbf{1}_{\text{inv}_t \geq 0} f_1(\text{inv}_t) + \mathbf{1}_{\text{inv}_t < 0} f_2(\text{inv}_t),$$
$$\mu^{ask}(\text{inv}_t) := \mathbf{1}_{\text{inv}_t < 0} f_1(\text{inv}_t) + \mathbf{1}_{\text{inv}_t \geq 0} f_2(\text{inv}_t),$$

$$\kappa := (\kappa_{max} - \kappa_{min}) \text{clamp}\left(\left|\frac{\text{inv}_t}{\text{max\_inv}}\right|^p\right) + \kappa_{min}.$$

In our experiments, we set the concentration $\kappa_{min} = 5$ and $\kappa_{max} = 20$, the maximum absolute inventory $max\_inv = 20$, the exponent $p = 2$, and the mode $\mu_0 = 0.2$ for μ^{ask} and μ^{bid}. The function $\text{clamp}(x) = \min(1, \max(-1, x))$. Intuitively, when the MM's inventory inv_t accumulates, the policy skews the volume distributions on both sides of the limit order book (e.g., buy less and sell more when inv_t becomes large) to reduce the inventory and control the risk.

6 Experimental Results

6.1 Performance

In Tables 1 and 2, the performance of twelve market-making policies is presented in two market settings: one without informed ZI agents and one with informed ZI agents. For each market setting, we train an RL MM policy using Soft Actor-Critic [18] with a total number of RL steps 5×10^4. We evaluate four metrics to measure the trading effect of each market-making policy (including baselines and RL): the MM's payoff, the bid-ask spread, and the MM's market share.

In the market without informed ZI agents (Table 1), the RL MM policy achieved the highest average payoff among all market maker policies, while the "MM at the touch" policy performed the worst. The average payoff of the beta MM policies varied significantly with different parameters. For instance, the beta policy with $\alpha = 5$ and $\beta = 5$ achieved the second-highest average payoff, but its performance dropped sharply with $\alpha = 5$ and $\beta = 1$. The ladder policy, a special beta policy, showed decent performance, supporting its widespread use in prior research. Additionally, the inventory-driven policy earned less payoff in this setting, as deliberately controlling inventory could limit ones ability to benefits from other agents in the market. Lastly, the market share of the best-performing policy (i.e., the RL MM) was moderate, indicating that both excessive and insufficient trading can lower the overall payoff for market making.

In the second market setting (Table 2), 13 out of 25 regular ZI agents were replaced with informed ones. After introducing informed traders, compared to Table 1, we observed a substantial decline in the average payoffs of all MM policies, except for the MM at the touch, underscoring the detrimental effect of adverse selection. Notice that with the beta policies, adverse selection caused the MM's inventory to grow to 100 times the size it would have been without informed traders. This substantial increase in inventory led to a significant loss in value, highlighting the danger of inventory risk.

An interesting observation in the results with informed traders is that the MM at the touch is the only policy to earn a positive payoff.

This can be attributed to the large number of informed traders who bid around the true fundamental, which causes the midprice to become a better estimate of the final fundamental compared to the one used by the MMs. With this improved estimate, the MM at the touch policy can exploit the regular ZI agents and generate profits. Alongside the MM at the touch policy, the RL MM was found to be the most resilient policy to adverse selection, with the inventory-driven policy coming in as the second most resilient. However, the RL MM underperformed compared to the MM at the touch due to its restricted action space, which was confined to several price levels around the estimated fundamental. This limitation possibly excluded the current best bid and best ask prices. Theoretically, the number of price levels can be set extremely high to encompass all prices, including any best bid and best ask. Nevertheless, an extensive number of price levels will significantly flatten the volume distributions (making them nearly uniform), which would cause the RL policy to resemble the ladder policy, thereby diminishing the effectiveness of market making. This flattening effect can be easily verified by examining Equation 2 and considering a large *n_levels*. Despite this restriction, the RL MM demonstrated much stronger resilience to adverse selection than other market-making policies. Moreover, the inventory-driven policy, designed to prevent inventory accumulation, also showed good resilience compared to beta policies. Interestingly, in many simulations, we observed that the RL MM consistently held a higher inventory compared to the inventory-driven policy. This indicates that the optimal strategy to counter adverse selection may not solely rely on controlling inventory. Instead, it involves quickly adapting by broadening the trading spread when risks increase and narrowing it when market conditions become more stable.

Policy Type	ω	α	β	Uninformed	Informed
Ladder MM	60	-	-	4.48×10^6	-8.67×10^7
Beta MM	60	1	2	4.28×10^6	-8.81×10^7
Beta MM	60	1	5	3.68×10^6	-1.05×10^8
Beta MM	60	2	1	4.46×10^6	-6.86×10^7
Beta MM	60	2	2	4.81×10^6	-7.89×10^7
Beta MM	60	2	5	4.61×10^6	-9.87×10^7
Beta MM	60	5	1	2.89×10^6	-4.98×10^7
Beta MM	60	5	2	3.65×10^6	-5.53×10^7
Beta MM	60	5	5	4.75×10^6	-6.89×10^7
Invt MM	60	-	-	4.32×10^6	-4.20×10^7
RL MM	60	-	-	$\mathbf{4.98 \times 10^6}$	$\mathbf{-2.10 \times 10^7}$

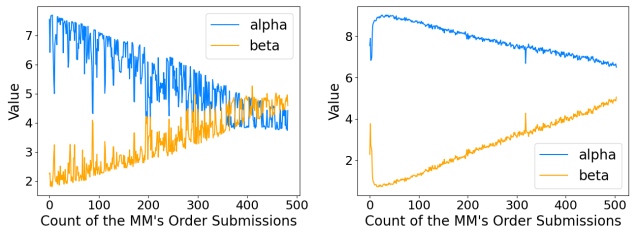
Table 3: The payoffs of MMs with a wider spread.

Furthermore, we found that the beta policy with $\alpha = 5$ and $\beta = 1$ outperformed all other beta policies. This can be visually interpreted by examining the volume distribution shown in Figure 3(f). Due to the inaccuracy in the estimated final fundamental, placing orders near this estimate can result in payoff losses. Consequently, the parameters that led to superior performance tend to distribute volume away from the estimate, meaning buy orders are submitted at lower prices and sell orders at higher prices.

Finally, we observed that while the ladder MM achieved good payoffs in the absence of informed traders, its performance significantly declined with their introduction. This suggests that the effectiveness of a ladder MM is highly dependent on the market setting, particularly the types of market participants; otherwise, it could fail to market make effectively. In Table 3, we provide more results on the performance of market-making policies with a wider spread $\omega = 60$. Our findings for $\omega = 10$ and for $\omega = 60$ are consistent, with the exception that the optimal parameters for the beta policy adjust in response to the change in spread.

6.2 Adaptability of RL Policy

The success of the RL market-making policy can be attributed to three main factors: its adaptability (within its trained market), the use of a general functional form (such as a neural network), and features other than inventory. We specifically examine the adaptability in the two market settings separately. In Figure 4(a), we plot the changes in the RL policy outputs, α and β, during a single simulation in the market with uninformed background traders. Initially, α becomes very large while β becomes very small, resulting in volume being distributed far from the spread (similar to Figure 3(f)). This occurs because, at the early stage, the MM cannot accurately estimate the final fundamental value. Consequently, the RL policy learns to reduce the risk of trading with the inaccurate estimate, thus distributing volume away from the estimate (i.e., placing buy orders at lower prices and sell orders at higher prices). As the simulation advances, the time step left in the observation declines and then the estimate becomes increasingly accurate. Since all agents trade around the same estimate (i.e., there are no informed traders), the market spread begins to narrow and the RL policy also narrows its trading spread gradually to get its orders transacted. As a result, the values of α and β quickly converge, leading to normal trading activity and producing volume distributions similar to those shown in Figure 3(h).



(a) Without informed ZI (b) With informed ZI

Figure 4: The evolution of α and β in a sample simulation.

However, when informed background traders are introduced, the estimated fundamental value becomes much less accurate, further increasing the risk of trading around it. This heightened risk can lead to substantial losses for both RL and uninformed traders. Consequently, the RL policy adjusts by placing buy orders at even lower prices and sell orders at even higher prices, significantly distributing volume away from the estimate. This can be achieved by maintaining a high value for α and a low value for β until the very end of the simulation, when the accuracy of the estimate improves, as shown in Figure 4(b). This example demonstrates the RL's ability to adapt to different market conditions, which is crucial for the success of RL in market making.

7 Conclusion and Discussion

We combine deep RL with beta distributions to develop a market-making policy capable of adapting to varying market conditions. Leveraging deep RL, our approach avoids assuming a specific functional form for the market-making policy and does not depend on strong assumptions about prior knowledge of the underlying market. By meticulously designing the reward function and observation features, we demonstrate that our RL beta policy outperform the baseline policies in two market settings with differing levels of inventory risk. Finally, we highlight the strong adaptability of the learned RL beta policy, underscoring its crucial role in achieving superior performance compared to other market-making policies.

Based on this work, several future research directions are worth exploring. Firstly, in the experiments, each RL beta policy is trained specifically for an individual market, limiting each policy to the specific market conditions it was designed for. A promising research avenue would be to investigate whether a single RL beta policy can be effectively trained across multiple markets with varying characteristics (e.g., different proportions of informed versus uninformed traders). Secondly, as these experiments are conducted within a market simulator rather than using real-world trading data, another potential research direction would be to evaluate the RL beta policy's performance on actual trading data, such as that provided by LOBSTER. Thirdly, further investigation into the performance of the RL beta policy could involve adjusting some of the initial experimental assumptions. For instance, our experiments assume that MMs are uninformed and passively provide liquidity, while, in reality, MMs are often sophisticated, informed agents. Thus, future research could examine the RL beta policy under conditions where the MM is modeled as a more informed, competitive agent.

References

[1] Jacob D. Abernethy and Satyen Kale. 2013. Adaptive Market Making via Online Learning. In *Proc. of NIPS*.

[2] Selim Amrouni, Aymeric Moulin, Jared Vann, Svitlana Vyetrenko, Tucker Balch, and Manuela Veloso. 2021. ABIDES-Gym: Gym environments for multi-agent discrete event simulation and application to financial markets. In *2nd ACM International Conference on AI in Finance*. 30:1–30:9.

[3] Marco Avellaneda and Sasha Stoikov. 2008. High-frequency trading in a limit order book. *Quantitative Finance* 8, 3 (2008), 217–224.

[4] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. JAX: composable transformations of Python+NumPy programs. http://github.com/google/jax

[5] David Byrd, Maria Hybinette, and Tucker Hybinette Balch. 2020. ABIDES: Towards high-fidelity multi-agent market simulation. In *34th ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*. 11–22.

[6] Álvaro Cartea, Ryan Donnelly, and Sebastian Jaimungal. 2017. Algorithmic Trading with Model Uncertainty. *SIAM Journal on Financial Mathematics* (2017).

[7] Alvaro Cartea, Ryan Donnelly, and Sebastian Jaimungal. 2018. Enhancing trading strategies with order book signals. *Applied Mathematical Finance* (2018).

[8] Álvaro Cartea and Sebastian Jaimungal. 2015. Risk Metrics and Fine Tuning of High-Frequency Trading Strategies. *Mathematical Finance* (2015).

[9] Álvaro Cartea, Sebastian Jaimungal, and José Penalva. 2015. *Algorithmic and High-Frequency Trading*. Cambridge University Press.

[10] Álvaro Cartea, Sebastian Jaimungal, and Jason Ricci. 2014. Buy low, sell high: A high frequency trading perspective. *SIAM Journal on Financial Mathematics* (2014).

[11] Tanmoy Chakraborty and Michael Kearns. 2011. Market Making and Mean Reversion. In *Proc. of ACM EC*. 307–314.

[12] Nicholas Tung Chan and Christian Shelton. 2001. An Electronic Market-Maker. (2001).

[13] Fayçal Drissi. 2022. Solvability of differential Riccati equations and applications to algorithmic trading with signals. *Applied Mathematical Finance* 29, 6 (2022), 457–493.

[14] Sascha Yves Frey, Kang Li, Peer Nagy, Silvia Sapora, Christopher Lu, Stefan Zohren, Jakob Foerster, and Anisoara Calinescu. 2023. JAX-LOB: A GPU-Accelerated limit order book simulator to unlock large scale reinforcement learning for trading. In *4th ACM International Conference on AI in Finance*. 583–591.

[15] Bruno Gasperov and Zvonko Kostanjcar. 2021. Market Making With Signals Through Deep Reinforcement Learning. *IEEE Access* 9 (2021), 61611–61622.

[16] Dhananjay K. Gode and Shyam Sunder. 1993. Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *Journal of Political Economy* 101, 1 (1993), 119–137.

[17] Olivier Guéant, Charles-Albert Lehalle, and Joaquin Fernandez-Tapia. 2011. Dealing with the Inventory Risk: A solution to the market making problem. *Mathematics and Financial Economics* (2011).

[18] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*. PMLR, 1861–1870.

[19] Konark Jain, Nick Firoozye, Jonathan Kochems, and Philip Treleaven. 2024. *Limit order book simulations: A review*. Technical Report. University College London. Available at SSRN, 4745587.

[20] Joseph Jerome, Gregory Palmer, and Rahul Savani. 2022. Market making with scaled beta policies. In *ICAIF*. 214–222.

[21] Joseph Jerome, Leandro Sánchez-Betancourt, Rahul Savani, and Martin Herdegen. 2022. Model-based gym environments for limit order book trading. (2022). arXiv:2209.07823

[22] Adlar J. Kim, Christian R. Shelton, and Tomaso Poggio. 2002. *Modeling stock order flows and learning market-making from data*. AI Memo 2002-009. MIT.

[23] Michail G Lagoudakis and Ronald Parr. 2003. Least-Squares Policy Iteration. *JMLR* 4 (2003), 1107–1149.

[24] Blake LeBaron. 2000. Agent-Based Computational Finance: Suggested Readings and Early Research. *Journal of Economic Dynamics and Control* 24 (2000), 679–702.

[25] Ye-Sheen Lim and Denise Gorse. 2018. Reinforcement Learning for High-Frequency Market Making. In *Proc. of ESANN*.

[26] Javier Falces Marin, David Díaz Pardo de Vera, and Eduardo Lopez Gonzalo. 2022. A reinforcement learning approach to improve the performance of the Avellaneda-Stoikov market-making algorithm. *Plos one* 17, 12 (2022).

[27] Chris Mascioli, Anri Gu, Yongzhao Wang, Mithun Chakraborty, and Michael P. Wellman. 2024. A Financial Market Simulation Environment for Trading Agents Using Deep Reinforcement Learning. In *5th International Conference on Artificial Intelligence in Finance*.

[28] Kaj Nyström, Sidi Mohamed Ould Aly, and Changyong Zhang. 2014. Market making and portfolio liquidation under uncertainty. *International Journal of Theoretical and Applied Finance* 17, 05 (2014), 1450034.

[29] R. G. Palmer, W. Brian Arthur, John H. Holland, Blake LeBaron, and Paul Tayler. 1994. Artificial economic life: A simple model of a stockmarket. *Physica D: Nonlinear Phenomena* 75 (1994), 264–274.

[30] Jonathan Sadighian. 2019. Deep reinforcement learning in cryptocurrency market making. *arXiv preprint arXiv:1911.08647* (2019).

[31] Megan Shearer, Gabriel Rauterberg, and Michael P. Wellman. 2023. Learning to manipulate a financial benchmark. In *4th ACM International Conference on AI in Finance*. 592–600.

[32] Thomas Spooner, John Fearnley, Rahul Savani, and Andreas Koukorinis. 2018. Market Making via Reinforcement Learning. In *Proc. of AAMAS*. 434–442.

[33] Thomas Spooner and Rahul Savani. 2020. Robust Market Making via Adversarial Reinforcement Learning. *Proc. of IJCAI* (2020).

[34] Elaine Wah and Michael P. Wellman. 2013. Latency Arbitrage, Market Fragmentation, and Efficiency: A Two-Market Model. In *14th ACM Conference on Electronic Commerce*. 855–872.

[35] Elaine Wah, Mason Wright, and Michael P. Wellman. 2017. Welfare effects of market making in continuous double auctions. *Journal of Artificial Intelligence Research* 59 (2017), 613–650.

[36] Xintong Wang, Christopher Hoang, Yevgeniy Vorobeychik, and Michael P. Wellman. 2021. Spoofing the limit order book: A strategic agent-based analysis. *Games* 12, 2 (2021), 46.

[37] Xintong Wang and Michael P. Wellman. 2017. Spoofing the limit order book: An agent-based model. In *AAMAS*. 651–659.

[38] Michael P. Wellman. 2006. Methods for Empirical Game-Theoretic Analysis (Extended Abstract). In *Twenty-First National Conference on Artificial Intelligence*. Boston, 1552–1555.

[39] Michael P Wellman, Karl Tuyls, and Amy Greenwald. 2024. Empirical game-theoretic analysis: A survey. *Journal of Artificial Intelligence Research* (2024).

[40] Ziyi Xu, Xue Cheng, and Yangbo He. 2022. Performance of Deep Reinforcement Learning for High Frequency Market Making on Actual Tick Data. In *Proc. of AAMAS*. 1765–1767.

[41] Yueyang Zhong, YeeMan Bergstrom, and Amy R. Ward. 2020. Data-Driven Market-Making via Model-Free Learning. In *Proc. of IJCAI*. 4461–4468.