IMM: An Imitative Reinforcement Learning Approach with Predictive Representation Learning for Automatic Market Making

Hui Niu[*1] , **Siyuan Li**[* 2] , **Jiahao Zheng**[3] , **Zhouchi Lin**[3] , **Bo An**[4,5] , **Jian Li**[†1] and **Jian Guo**[† 3]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[2]Faculty of Computing, Harbin Institute of Technology, Harbin, China
[3]International Digital Economy Academy, Shenzhen, China
[4]Skywork AI, Singapore
[5]School of Computer Science and Engineering, Nanyang Technological University, Singapore

Abstract

Market making (MM) via Reinforcement Learning (RL) has attracted significant attention in financial trading. Most existing RL-based MM methods focus on optimizing single-price level strategies which fail at frequent order cancellations and loss of queue priority. By comparison, strategies involving multiple price levels align better with actual trading scenarios. However, given the complexity that multi-price level RL strategies involve a comprehensive trading action space, the challenge of effectively training RL persists. Inspired by the effective workflow of professional human market makers, we propose Imitative Market Maker (IMM), a novel RL framework leveraging knowledge from both suboptimal signal-based experts and direct policy interactions. Our framework starts with introducing effective state and action formulations that well encode information about multi-price level orders. Furthermore, IMM integrates a representation learning unit capable of capturing both short- and long-term market trends to mitigate adverse selection risk. Subsequently, IMM designs an expert strategy based on predictive signals, and trains the agent through the integration of RL and imitation learning techniques to achieve efficient learning. Extensive experimental results on four real-world market datasets demonstrate the superiority of IMM against current RL-based market making strategies.

1 Introduction

Market making (MM) is a process where a market maker continuously places both buy and sell orders on the limit order book (LOB) of a given security with an aim of maximizing the risk-adjusted returns. The complexity of MM dynamics brings great challenges to market makers. During this process, market makers encounter various risks including inventory risk, adverse selection risk, and non-execution risk.

In contrast to traditional MM approaches [Avellaneda and Stoikov, 2008; Guéant *et al.*, 2012] which rely on mathematical models with strong assumptions, (deep) Reinforcement Learning (RL) has emerged as a promising approach for developing MM strategies capable of adapting to changing market dynamics. While there has been extensive research on the application of RL for MM, the majority of studies have focused on optimizing single-price level strategies [Spooner *et al.*, 2018; Sadighian, 2019; Xu *et al.*, 2022]. Moreover, most RL-based methods utilize the fluctuated mid-price as a reference price for defining actions related to order prices. Unfortunately, such single-price strategies with unstable reference prices result in frequent and unnecessary order cancellations, leading to the loss of order priority [Chung *et al.*, 2022]. In practice, traders prefer placing multi-price level orders in advance without frequent cancellations to reserve good queue positions. This raises the necessity for a novel MM formulation that accommodates order stacking and muli-price level strategies with a stable reference price.

Besides the formulation challenge, addressing the effective training of profitable RL policies for MM becomes a formidable task when dealing with multi-price level strategies, which involve a large, fine-grained trading action space. Learning from scratch proves to be particularly challenging for RL algorithms in such complex environments. A promising way to solve this problem is to extract trading knowledge from expert data and improve the exploration ability of RL algorithms. Furthermore, to tackle the primary concern of market makers in mitigating adverse selection risk and non-execution risk, agents must possess the capability for both long-term and short-term market condition predictions. This enables the balancing of long-term risk caused by unfavorable price movements and short-term execution probabilities.

With the above motivations, this paper proposes the Imitative Market Maker (IMM), an RL framework that integrates a state representation learning unit (SRLU) with an imitative RL unit (IRLU), to solve the optimal MM problem using multi-price level policies. IMM first gathers both micro and macro-level market information and predicts short-term and long-term market trends based on this information. Later on, it trades off between risks and profits based on risk appetite. Finally, it makes trading decisions to decide how wide and how asymmetrically w.r.t. the reference price it sets the quotes.

Our primary contributions can be summarized as follows:

- We formulate the MM process as a MDP with effective state, action, and reward definitions. This formulation is well-suited for realistic MM scenarios, enabling the implementation of multi-price level order stacking and customized risk-profit trade-offs.

- We propose SRLU that incorporates multi-granularity predictive signals as auxiliary variables, and employs a temporal convolution and spatial attention (TCSA) network to distill representations from noisy market data.
- We leverage IRLU to extract trading knowledge from experts, facilitating effective exploration in the complex trading environment.
- Extensive experiments on four real-world financial futures datasets demonstrate that IMM outperforms baseline methods in relation to both risk-adjusted returns and adverse selection ratios.

2 Related Work

Traditional approaches for MM in the finance literature [Amihud and Mendelson, 1980; Glosten and Milgrom, 1985; Avellaneda and Stoikov, 2008; Guéant *et al.*, 2012] consider MM as a stochastic optimal control problem that can be solved analytically. For example, the Avellaneda–Stoikov (AS) model [Avellaneda and Stoikov, 2008] assumes a drift-less diffusion process for the mid-price evolution, then uses the Hamilton-Jacobi-Bellman equations to derive closed-form approximations to the optimal quotes. On a related note, [Guéant *et al.*, 2012] consider a variant of the AS model with inventory limits. However, such methods are typically predicated on a set of strong assumptions, and employ multiple parameters that need to be laboriously calibrated on historical data. It seems promising to consider more advanced methods such as RL that enable learning directly from data in a model-free fashion.

Recent years have witnessed a strong popularity of (deep) RL in the field of quantitative trading [Chan and Shelton, 2001; Cartea *et al.*, 2015a; Patel, 2018; Zhong *et al.*, 2020; Fang *et al.*, 2021; Niu *et al.*, 2022; Sun *et al.*, 2023]. The majority of the RL-based MM approaches adopt a single-price level strategy. Among those studies, several methods proposed defining the action space in advance [Spooner *et al.*, 2018; Xu *et al.*, 2022; Sadighian, 2019]. For instance, the action space of [Spooner *et al.*, 2018] is an octuple of prespecified half-spread pairs. Several researchers utilized a "half-spread" action space which chooses a continuous half-spread on each side of the book [Glosten and Milgrom, 1985; Jumadinova and Dasgupta, 2010; Cartea *et al.*, 2015b; Lim and Gorse, 2018; Guéant and Manziuk, 2019]. Unfortunately, when actually implementing such a strategy, to change the half spread on each side of the book, it is necessary to actively cancel orders at each time step and place new orders at the new level. This results in frequent unnecessary order cancellations, leading to queue position losing [Jerome *et al.*, 2022]. To overcome this limitation, ladder strategies, which place a unit of volume at all prices in two price intervals, one on each side of the book, have been adopted [Chakraborty and Kearns, 2011; Abernethy and Kale, 2013]. The latest work uses variants of this strategy to construct multi-price level MM policies. The DRL_{OS} [Chung *et al.*, 2022] agent decides whether to retain one unit of volume on each price level and the beta policy [Jerome *et al.*, 2022] allows for flexibility in the distribution of order volumes. However, to provide more accurate trading decisions, a multi-price level strategy involves a much more complex fine-grained action space (multi-level price and quantity). Little attention has been paid to inefficient exploration problems due to the complex action space of multi-price strategies. While have shown great promise, these approaches might be limited to achieving efficient exploration, particularly in highly dynamic and complex market environments.

3 Problem Formulation

In this section, we formulate the MM problem with multi-price level strategies as a Markov Decision Process (MDP). We start with introducing a stable reference price of LOB and a novel state/action space. Subsequently, we illustrate the transition dynamics of MM procedure that accommodates multi-price level order stacking. Finally, we introduce a mixed reward function to mitigate diverse risks.

3.1 A Stable Reference Price

In practical MM scenarios, market participants analyze many quantities before sending orders, among which the most important one is the distance between their target price and the "reference market price" p_{ref}, typically the midprice [Huang *et al.*, 2013]. The LOB can be formulated as a $2K$-dimensional vector, where K denotes the number of available limits on each side. Notably, p_{ref} serves as the LOB's central point, thereby determining the positions of the 2K limits $Q_{\pm i}$, where $Q_{\pm i}$ represents the limit at the distance $i - 0.5$ ticks to the right $(+i)$ or left $(-i)$ of p_{ref}. It is assumed that buy limit orders are placed on the bid side, and sell ones on the ask side. The queue length at Q_i is denoted as l_i.

Most existing MM methods adopt the market midprice as p_{ref} to encode the LOB. However, the specific price linked with level i may experience frequent shifts due to the dynamic fluctuations in the midprice. Whenever alterations occur in p_{ref}, the corresponding l_i instantaneously transitions to the value of one of its adjacent neighbors. This hinders the extraction of valid micro-market information from LOB. The necessity arises to define a stable reference price.

To this end, this paper formulates p_{ref} in the subsequent manner: Firstly, we set up a reference price $\tilde{p}_{ref}$ that follows the midprice. Specifically, in instances where the market spread is odd, $\tilde{p}_{ref,t} := m_t = \frac{ask_t + bid_t}{2}$. When the spread is even, $\tilde{p}_{ref,t} := m_t \pm \frac{\text{price tick}}{2}$, with the selection of the sign based on proximity to the prior value. Afterward, at the beginning of an episode, we set $p_{ref,0} = \tilde{p}_{ref,0}$. Then when the midprice m_t increases (or decreases), only if $l_{-1,t} = 0$ (or $l_{1,t} = 0$), $p_{ref,t}$ is updated to $\tilde{p}_{ref,t}$. Consequently, changes of p_{ref} are possibly caused by one of the three following events: (1) The insertion of a buy/sell limit order within the bid-ask spread while Q_1/Q_{-1} is empty. (2) A cancellation of the last limit order at one of the best prices. (3) A market order that consumes the last limit order at one of the best offer

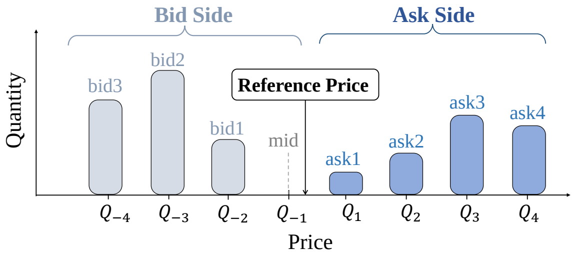


Figure 1: Limit order book and the reference price. In this instance, Q_{-1} is an empty limit.

queues. Note that within our framework, the LOB accommodates empty limits, as depicted in Figure 1. In this way, we obtain a more stable p_{ref}[1], enabling effective encoding of LOB and multi-price level orders.

3.2 MDP Formulation

With such a stable reference price, IMM then effectively encodes market information and multi-price level orders.

State Space

To mitigate adverse selection risk, the state space necessitates the inclusion of macro-level market information[1]. At time step t, the IMM agent observes the state formulated by:

$$\boldsymbol{s}_t = (\boldsymbol{s}_t^m, \boldsymbol{s}_t^s, \boldsymbol{s}_t^p), \tag{1}$$

where $\boldsymbol{s}_t^m$ denotes the *market variables* encoding the current market status; $\boldsymbol{s}_t^s$ denotes the *signal variables*, including multi-granularity auxiliary predictive signals; $\boldsymbol{s}_t^p$ denotes the *private variables*, including: the current inventory z_t, the queue position information and volume of the agent's orders that rests on the LOB, denoted by $s_t^q = (q_t^{-K} ..., q_t^{-1}, q_t^1, ..., q_t^K)$ and $s_t^v = (v_t^{-K} ..., v_t^{-1}, v_t^1, ..., v_t^K)$ respectively. Here q_i and v_i denote the queue position information and volume at price level i respectively. Suppose there are m_i orders resting at level i placed at different time steps. The queue position value of the j-th order at level i can be defined as $q_t^{i,j} = \frac{l_i^{front,j}}{l_t^i}$, where $l_i^{front,j}$ is the queue length in front of this order. Thus the queue position value at price level i can be defined as the volume weighted average of the queue position values of the m_i orders: $q_t^i = \sum_{j=1}^{m_i} \frac{l_t^{front,i,j}}{l_t^i} \frac{v_t^{i,j}}{l_t^i}$. Through covering the information of the current quotes in states, the IMM learns to avoid frequent order cancellations and replacements.

Action Space

Reserving good queue positions beyond best bid/ask levels holds the advantage of controlling adverse selection and non-execution risks. Therefore, practitioners tend to deploy order stacking strategies that place limit orders at multiple price levels in advance. IMM introduces an action encoding that expresses the complex multi-price level strategies within a low-dimensional space. At time step t, the action $\boldsymbol{a}_t$ is defined as

$$\boldsymbol{a}_t = (m_t^*, \delta_t^*, \boldsymbol{\phi}_t^{bid}, \boldsymbol{\phi}_t^{ask}), \tag{2}$$

where m_t^* and δ_t^* denote the desired quoted midprice w.r.t p_{ref} and spread respectively. This implies that the agent's target selling price is no lower than $m_t^* + \delta_t^*/2$, and the highest buying price is $m_t^* - \delta_t^*/2$. $\boldsymbol{\phi}_t^{bid}$ and $\boldsymbol{\phi}_t^{ask}$ represent parameter vectors that govern the volume distribution of the multi-level quotations. Several instances of diverse two-price level strategies and more-price level strategies are illustrated in Figure 2 and supplementary materials. By adopting such an action formulation, the agent gains the flexibility to determine both the width and asymmetry of the quotes with respect to the reference price.

Transition Dynamics

We formulate MM as an episodic RL task. The MM procedure allows for multi-price level order stacking, as specified below: (1) Choose a random start time for the episode and initialize the environment and the simulator. (2) Let the agent choose the desired volumes and price levels at which the agent would like to be positioned in the LOB. (3) Turn these desired positions into orders, including canceling orders from levels with too much volume and placing new limit orders. (4) Match the orders in the market-replay simulator according to the price-time priority. (5) Update the agent's cash and inventory of the traded asset and track profit and loss. (6) Repeat steps (2-4) until the episode terminates. A picture illustration can be found in Figure **??** in the supplementary material.

Reward Function for Diverse Utilities

The decision process of the market makers is subject to several trade-offs, including probability of execution and spread, inventory risk, and compensation from the exchange. To meet the diverse utilities of market makers, three factors are proposed to be considered:

Profit and loss (PnL) is a natural choice for the problem domain, comprising a realized PnL term (left part) and a floating PnL term (right part), given by:

$$PnL_t = \left(\sum_{i \in A_t} p_i^a \cdot v_i^a - \sum_{j \in B_t} p_j^b \cdot v_j^b \right) + (p_{t+1} - p_t) \cdot z_{t+1}, \tag{3}$$

where p denotes the market midprice; p^a, v^a, p^b, v^b represent the price and volume of the filled ask (bid) orders respectively; z signifies the current inventory, with $z > 0$ when the agent holds a greater long position than short.

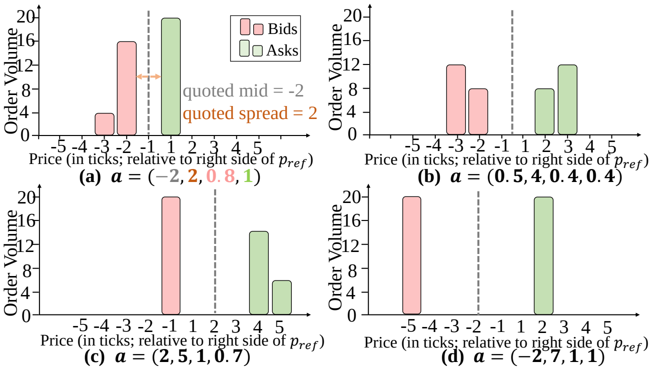
Truncated Inventory Penalty is used to mitigate inventory risk. Considering that advanced market makers may choose to hold a non-zero inventory to exploit clear trends while capturing the spread, an enhanced approach involves applying an additional inventory-dampening term solely to higher-risk inventory levels:

$$IP_t = -\eta |z_t| \cdot \mathbb{I}(|z_t| > C), \tag{4}$$

A penalty for inventory holding is applied solely when the inventory z_t surpasses a constant C.

The Market Makers' compensation from the exchange constitutes a primary revenue stream for numerous market makers [Fernandez-Tapia, 2015]. Therefore, ensuring a substantial volume of transactions to secure compensation holds significant importance for a variety of MM companies. To this

[1] See detailed explanations in the supplementary materials.



(a) $\boldsymbol{a} = (-2, 2, 0.8, 1)$

(b) $\boldsymbol{a} = (0.5, 4, 0.4, 0.4)$

(c) $\boldsymbol{a} = (2, 5, 1, 0.7)$

(d) $\boldsymbol{a} = (-2, 7, 1, 1)$

Figure 2: Illustration of action space. Consider a scenario where an agent positions two-level orders on both sides of the LOB, ϕ_t^{ask} (the green number) determines the ratio of volume placed at the two adjacent ask price levels (heights of the green bars).

end, a bonus term is incorporated to encourage transactions of the agent:

$$C_t = \beta\left(\sum_{i \in A_t} p_i^a \cdot v_i^a + \sum_{j \in B_t} p_j^b \cdot v_j^b\right), \tag{5}$$

Ultimately, by appropriately tuning the parameters η and β based on personalized utilities, IMM ensures alignment with the requirements of a broad spectrum of market makers, employing the combination of these three categories of rewards:

$$\mathcal{R}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}) = PnL_t + C_t + IP_t. \tag{6}$$

4 Imitative Market Maker (IMM)

This section introduces the proposed IMM Approach. As illustrated in Figure 3, IMM consists of two components: Section 4.1 elaborates on the SRLU that aims at forecasting multi-granularity signals while extracting valuable representations from the noisy market data. Section 4.2 outlines the MM policy learning approach, which incorporates the RL and imitation learning objectives.

4.1 State Representation Learning Unit (SRLU)

Signal Generation

To leverage the labels containing future information, IMM pretrains supervised learning (SL) models to generate both short- and long-term trend signals. The choice of the SL models is flexible. In this paper, we adopt LightGBM [Ke *et al.*, 2017], a highly robust ensemble model based on decision trees, to generate four multi-granularity trend signals denoted by $(y^{20}, y^{120}, y^{240}, y^{600})$, which are the labels of price movement trend after $1/6, 1, 2, 5$ minutes respectively. While training the RL policy, the parameters of the pre-trained predictors are frozen, and the outputs constitute the auxiliary signal variables $\boldsymbol{s}^s$.

Attention-based Representation Learning

Deep RL algorithms usually suffer from the low data-efficiency issue. Besides the auxiliary signal prediction, we propose a temporal convolution and spatial attention (TCSA) network to extract additional effective representations from the noisy market data $\boldsymbol{x}$. The structure of TCSA is depicted in Figure 3.

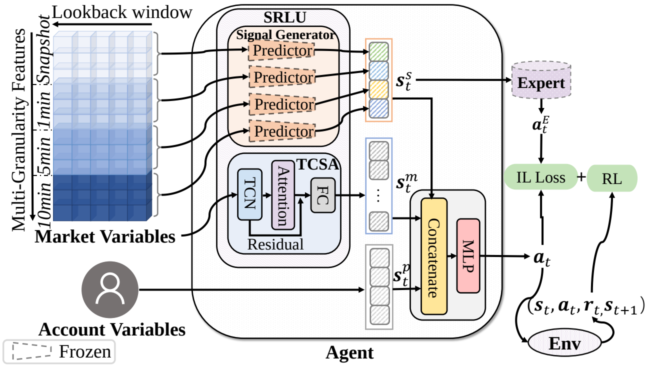


Figure 3: The proposed IMM learning framework.

IMM first utilizes a temporal convolution network (TCN) [Yu and Koltun, 2015] block to extract the time-axis relations in the data. Compared to recurrent neural networks, TCN has several appealing properties including parallel computation and longer effective memory. After conducting TCN operations on $\boldsymbol{x} \in \mathbb{R}^{F \times L}$ along the time axis, we obtain an output tensor denoted by $\hat{\boldsymbol{H}} \in \mathbb{R}^{F \times L}$, where F is the dimension of features, and L is the temporal dimension.

Afterward, IMM adopts an attention mechanism [Vaswani *et al.*, 2017] to handle the spatial relationships among different features. Given the output vector of TCN, we calculate the spatial attention weight as $\hat{\boldsymbol{S}} = \boldsymbol{V} \cdot \text{sigmoid}\big((\hat{\boldsymbol{H}}\boldsymbol{W}_1) \cdot (\hat{\boldsymbol{H}}\boldsymbol{W}_2)^T + \boldsymbol{b}\big)$, where $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^L$, and $\boldsymbol{V} \in \mathbb{R}^{F \times F}$ are parameters to learn, $\boldsymbol{b} \in \mathbb{R}^{F \times F}$ is the bias vector. The matrix $\hat{S} \in \mathbb{R}^{F \times F}$ is then normalized by rows to represent the correlation among features: $\boldsymbol{S}_{i,j} = \frac{\exp(\hat{\boldsymbol{S}}_{i,j})}{\sum_{u=1}^{F} \exp(\hat{\boldsymbol{S}}_{i,u})}, \forall 1 \leq i \leq F$.

We adopt the ResNet [He *et al.*, 2016] structure to alleviate the vanishing gradient problem in deep learning. The final representation abstracted from $\boldsymbol{x}$ is denoted by $H = S \times \hat{\boldsymbol{H}} + \boldsymbol{x}$, and it is then translated to a vector with dim F' using a fully connected layer: $\boldsymbol{s}^m = \text{sigmoid}(W_4 \cdot \text{ReLU}(\boldsymbol{H}W_3 + \boldsymbol{b}_3) + \boldsymbol{b}_4)$. The representation $\boldsymbol{s}^m$ is concatenated with the signal state $\boldsymbol{s}^s$ and private state $\boldsymbol{s}^p$.

4.2 Imitative Reinforcement Learning Unit(IRLU)

A Signal-Based Expert

To guide efficient exploration, we define a linear suboptimal rule-based expert strategy named Linear in Trend and Inventory with Inventory Constraints (LTIIC), which is commonly used by human experts. Readers might employ other effective expert strategies if available. $LTIIC(a, b, c, d)$ corresponds to a strategy where a market maker adjusts its quote prices based on both its inventory level and trend prediction signals. If $-d \leq z_t \leq d$ at time t, the ask and bid orders are placed with prices:

$$\begin{cases} ask_t^q = & m_t + a + b \cdot z_t + c \cdot \hat{y}_t, \\ bid_t^q = & m_t - a + b \cdot z_t + c \cdot \hat{y}_t, \end{cases} \tag{7}$$

where a, b, c and d are predetermined parameters, m_t stands for the midprice of the LOB, z_t represents the inventory, and

	RB			FU			CU			AG		
	EPnL[10^3]↑	MAP[unit]↓	PnLMAP↑	EPnL[10^3]↑	MAP[unit]↓	PnLMAP↑	EPnL[10^3]↑	MAP[unit]↓	PnLMAP↑	EPnL[10^3]↑	MAP[unit]↓	PnLMAP↑
FOIC	3.23 ± 4.35	255 ± 111	14 ± 22	-7.79 ± 9.25	238 ± 135	-43 ± 56	-33.05 ± 27.63	206 ± 141	-161 ± 224	-48.39 ± 28.83	189 ± 154	-250 ± 335
LIIC	2.26 ± 3.32	123±32	20 ± 29	-6.89 ± 6.66	115± 30	-66 ± 69	-24.19 ± 14.83	150 ± 20	-164 ± 513	-38.9 ± 26.2	142 ± 45	-302 ± 243
LTIIC	9.16 ± 4.87	65 ± 6	139 ± 68	8.26 ± 2.64	52 ± 3	160 ± 50	-16.74 ± 15.81	112 ± 109	-190 ± 203	-32.57 ± 22.8	128 ± 22	-264 ± 166
RL_{SD}	4.36 ± 1.64	**38 ± 4**	114 ± 38	7.31 ± 5.38	76 ± 29	90 ± 46	-19.7 ± 17	214 ± 109	-92 ± 298	-25.43 ± 23.83	107 ± 37	-237 ± 235
DRL_{OS}	8.22 ± 3.70	51 ± 4	156 ± 61	11.03 ± 13.87	**37 ± 3**	30 ± 36	-18.9 ± 18.02	647 ± 2367	-99 ± 147	-28.39 ± 27.92	169 ± 154	**-167 ± 135**
IMM	**16.46 ± 9.10**	96± 13	**165 ± 74**	**28.10 ± 10.27**	102 ± 14	**274 ± 89**	**-4.86 ± 10.17**	**111 ± 28**	**-43 ± 87**	**-14.5 ± 20.2**	**102 ± 14**	-274 ± 89

Table 1: The comparison results of the proposed method and the benchmarks.

$\hat{y}_t \in \{-1, 0, 1\}$ signifies a short-term predictive trend signal. The insights of LTIIC strategy lie in that during a short-term upward market trend, ask-side limit orders are more likely to be executed than bid-side ones. A logical approach involves implementing a narrow half-spread on the bid side and a broader half-spread on the ask side, thus reducing the risk exposure due to adverse selection. Simultaneously, the trader adjusts parameter b to regulate the inventory level, while a determines the quoted spread. In cases where $z_t >= d$ or $z_t <= -d$, only orders on the side opposite to the inventory are posted.

Policy Learning

We utilized the actor-critic RL framework [Konda and Tsitsiklis, 1999], where the critic evaluates the action taken by the actor by computing the value function, and the actor (policy) is optimized to maximize the value output by the critic. To improve the sample efficiency, we use the off-policy actor-critic method TD3 [Fujimoto *et al.*, 2018] as the base learner, and the policy $\pi : a = \mu_\theta(s)$ is updated with the deterministic policy gradient [Silver *et al.*, 2014]:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim\rho^\mu}\left[\nabla_\theta \mu_\theta(s)\nabla_a Q^\mu(s,a)\big|_{a=\mu_\theta(s)}\right], \tag{8}$$

where Q^μ is a value function approximating the expected cumulative reward of the policy $\mu_\theta(s)$.

In Equation (8), D denotes the replay buffer collected by a behavior policy, which is generated by adding some noise to the learned policy π. Following the TD3 method [Fujimoto *et al.*, 2018], the value function Q is optimized in a twin delayed manner with the data sampled from both the replay buffer D and expert dataset D_E.

Since the high-dimensional state space, the complex action space, and the stochastic trading environment induce a hard exploration problem, learning with a pure RL objective in Equation (8) is extremely difficult. To promote policy learning in such a complex trading environment, we propose to augment the RL method with the objective of imitating the quoting behavior in an expert dataset D_E as:

$$\pi = \arg\max_\pi \mathbb{E}_{(\boldsymbol{s},a)\sim D}\left[Q(\boldsymbol{s},\pi(\boldsymbol{s}))\right] - \mathbb{E}_{(\boldsymbol{s},\hat{a})\sim D_E}\left[\lambda\cdot(\pi(\boldsymbol{s})-\hat{a})^2\right], \tag{9}$$

where λ is a scaling coefficient that balances maximizing the Q values and minimizing the behavior cloning (BC) loss. We set λ to decrease with the growth of the training steps.

As the expert dataset contains reasonable suboptimal MM behaviors, the agent benefits from imitation learning techniques through abstracting advanced trading knowledge. Thus the proposed method could achieve more efficient exploration and policy learning in the highly stochastic market environment compared to the RL methods without imitation learning.

5 Experiments

5.1 Experimental Setup

We conduct experiments on four datasets comprised of historical data of the spot month contracts of the FU, RB, CU, and AG futures from the Shanghai Futures Exchange[2]. The data consists of the 5-depth LOB and aggregated trade information associated with a 500-millisecond real-time financial period. We use the data from July 2021 to March 2022 (126 trading days) for training with 20% as the validation set, and test model performance on April 2022 $\sim$ July 2022 (60 trading days). In each episode, the agent adjusts its 2-level bids and asks every 500 millisecond, with a fixed total volume of $N = 20$ units on each side. The episode length is set to 1.5 trading hours, with $T = 10800$ steps.

Benchmarks

We compare IMM with three rule-based benchmarks and two state-of-the-art RL-based approaches:

1. **FOIC** represents a Fixed Offset with Inventory Constraints strategy introduced by [Gašperov and Kostanjčar, 2021]. $FOIC(d)$ refers to the strategy that posts bid (ask) orders at the current best bid (ask) while adhering to the inventory constraint d.
2. **LIIC**. A Linear in Inventory with Inventory Constraints strategy [Gašperov and Kostanjčar, 2021] corresponds to the strategy where a market maker adjusts its quote prices based on its inventory level. The quotes of LIIC can be formulated as $LTIIC(a, b, c = 0, d)$ using Equation (7).
3. **LTIIC** is the expert adopted in IMM.
4. $\mathbf{RL_{DS}}$ refers to a RL-based single-price level strategy proposed by [Spooner *et al.*, 2018].
5. $\mathbf{DRL_{OS}}$ refers to a state-of-the-art RL-based multi-price level strategy proposed in [Chung *et al.*, 2022]. The agent decides whether to retain one unit of volume on each price level, not allowing for volume distribution across all price levels.

Evaluation Metrics

We adopt four financial metrics to assess the performance of an MM strategy:

- **Episodic PnL** is a natural choice to evaluate the profitability of a MM agent, since there is no notion of starting capital in MM procedure: $EPnL_T = \sum_{t=1}^{T} PnL_t$.

[2] Here RB, FU, CU, and AG refers to the Steel Rebar, Fuel Oil, Copper, and Silver Futures Contracts respectively.

- **Mean Absolute Position (MAP)** accounts for the inventory risk, defined as: $MAP_T = \frac{\sum_{t=1}^{T} |z_t|}{\sum_{t=1}^{T} 1\cdot\mathbb{I}(|z_t|>0)}$.
- **Return Per Trade (RPT)** evaluates the agent's capability of capturing the spread. It is normalized across different markets by the average market spread $\overline{\delta^m}$.

$$RPT_T = \left(\frac{\sum_{i\in A_T} p_i^a * n_i^a}{\sum_{i\in A_T} n_i^a} - \frac{\sum_{j\in B_T} p_j^b * n_j^b}{\sum_{j\in B_T} n_j^b}\right)\Big/\overline{\delta^m}.$$

- **PnL-to-MAP Ratio (PnLMAP)** simultaneously considers the profitability and the incurred inventory risk of a market making strategy: $PnLMAP_T = \frac{EPnL_T}{MAP_T}$.

5.2 Comparison Results with Baselines

For a fair comparison, we tune the hyper-parameters of these methods for the maximum PnLMAP$_T$ value on the validation dataset. The comparison results of IMM and the benchmarks on the four test datasets are given in Table 1 and supplementary materials. These comparison results indicate that the proposed approach significantly outperforms the benchmarks in terms of both profitability and risk management.

As demonstrated in Table 1, on the RB dataset, the proposed method attains the highest terminal wealth as well as risk-adjusted return, albeit with a slightly elevated MAP in comparison to the expert and the two RL-based methods. Besides, the two multi-price level RL-based agents DRL_{OS} and IMM outperform the single-price level method RL_{DS}, indicating the superiority of the mult-price level strategy. On the FU dataset, IMM not only achieves the highest terminal wealth but also demonstrates the most favorable return-to-risk performance and spread-capturing ability, while maintaining the second-lowest inventory level. The RL-based strategies achieve commendable performance compared to the rule-based strategies which fail to make profits in most trading days. Moreover, it is observed that IMM acquires a competitive MM strategy, which attains stable dividends with profits (the pink line) while sustaining the inventory at a tolerable low level (the compact blue region) in a violate market, as illustrated in the left part of Figure 4. The inventory fluctuates around zero, which is a desirable behavior. Remarkably, since IMM does not force the agent to place opposite-side orders to clear its inventory, it is quite an appealing result to see IMM accomplish automatic inventory control based on state-derived information. Even in the challenging MM tasks on CU and AG markets which show lower market liquidity, the proposed method significantly outperforms the benchmarks in terms of terminal wealth and risk-adjusted return.

5.3 Model Component Ablation Study

To investigate the effectiveness of the model component, we compare the proposed IMM with its five variations summarized in Table 2, and the results are listed in Table 3. In Table 2, the notions "QuotesInfo", "Signals", "TCSA", "RL", and "IL" refer to whether to include "quote state s_t^q and s_t^v", "signal state s_t^s", "TCSA network", "RL objective $E_{(s,a)\sim D}[...]$ in Eq.5", and "IL objective $E_{(\boldsymbol{s},\hat{a})\sim D_E}[...]$ in Eq.5" in the model. In Table 3, "SR" refers to the Sharpe Ratio calculated as $\frac{\mathbb{E}(EPnL_T)}{\sigma(EpnL_T)}$.

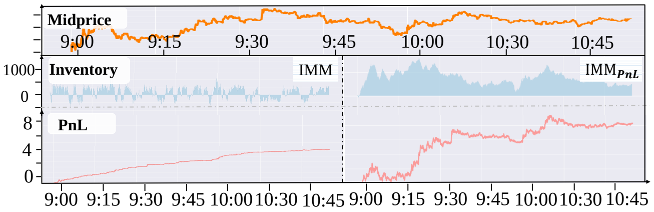
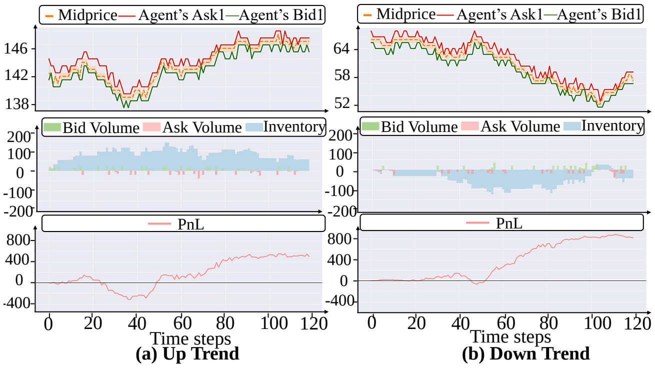


Figure 4: Performance of IMM and IMM$_{-}PnL$ on FU on Jun. 14th, 2022.



(a) Up Trend

(b) Down Trend

Figure 5: 1-minute Cases. IMM performs well in both up and downtrend markets.

Effectiveness of the State Representations

To examine the efficacy of the proposed state representations, we analyze the performance of three IMM$_{SL(\cdot)}$ models. Based on Table 3, it is evident that introducing multigranularity signals as auxiliary observations holds significant importance in enhancing the MM strategy's performance. Figure 5 visualizes the behavior of IMM during two 1-minute periods with different trends on FU test dataset. As demonstrated in Figure 5(a), benefiting from the auxiliary signals, the IMM agent anticipates an ascending price trend and proactively maintains a long position prior to the onset of a short-term bullish trend (steps 0-40). With the trend terminates (step 70-120), the agent gradually reduces its inventory through placing orders with narrow ask-side half-spread and broad bide-side one w.r.t the market midprice. Similarly, the IMM agent demonstrates proficient behavior in downside markets as shown in 5(b). Combined with the adverse select ratio depicted in Figure 6, it can be concluded that IMM has learned to mitigate adverse selection risk.

For a deeper investigation into the role of auxiliary signals in improving performance, we calculated the adverse selection ratio as $adv\_ratio = \frac{\text{\# adverse fills}}{\text{\# fills in last interval}}$. Here adverse fills refer to limit bid (ask) orders which are executed shortly before a downward (upward) movement of the best bid (ask) price. As if the best bid price had gone down, it might have been better to wait for the next bid price level [Chung *et al.*, 2022]. Based on Figure 6(a), we can deduce that the multigranularity predictive signals play a vital role in mitigating adverse selections. That might because they provide effective information about market conditions, which enables a more flexible trade-off between spread-capturing and trend-

Models	QuotesInfo	Signals	TCSA	RL	IL
IMM$_{SL(m)}$	O	O	X	O	O
IMM$_{SL(s)}$	O	X	O	O	X
IMM$_{SL(q)}$	X	O	O	O	O
IMM$_{BC(0)}$	O	O	O	O	X
IMM$_{BC(1)}$	O	O	O	X	O

Table 2: Five variations of the proposed IMM.

	EPnL[10^3]↑	MAP[unit]↓	PnLMAP↑	SR↑
IMM$_{SL(m)}$	10.57 ± 8.63	74 ± 41	142 ± 39	1.22
IMM$_{SL(s)}$	7.83 ± 3.64	**49 ± 5**	159 ± 46	2.15
IMM$_{SL(q)}$	10.20 ± 9.72	74 ± 47	104 ± 56	1.05
IMM$_{BC(0)}$	14.67 ± 5.11	85 ± 5	172 ± 57	**2.87**
IMM$_{BC(1)}$	8.22 ± 3.70	51 ± 4	156 ± 61	2.22
IMM	**28.097 ± 10.27**	103 ± 15	**274 ± 89**	2.80

Table 3: Comparison results of ablation study on FU dataset.

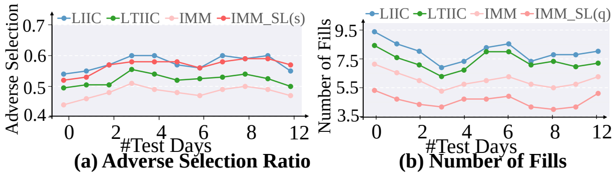
chasing. Moreover, Figure 6(b) demonstrates that the information regarding multi-price level orders additionally contributes to enhancing the fill count by minimizing frequent cancellations and preserving queue positions.

Effectiveness of the IRLU

The comparison results between IMM$_{BC(\cdot)}$ and IMM provide empirical evidence of the importance of extracting additional knowledge from the expert and conducting efficient exploration, particularly for challenging financial tasks. Besides, IMM outperforms the expert LTIIC strategy a lot. As the RL agent faces challenges in identifying a viable trading approach and sustaining it over multiple steps during the initial training phase. Training while pursuing the imitation learning objective facilitates the agent in obtaining favorable rewards and drawing valuable lessons from these experiences.

Effects of Different Reward Functions

To investigate whether the proposed reward function meets different utilities, we train IMM that with three different rewards: The IMM$_{PNL}$ method trains IMM using the PnL reward ($\eta, \beta = 0$); The IMM$_{PNL+C}$ method trains IMM with the combination of the PNL and compensation reward ($\eta = 0, \beta > 0$); and The IMM$_{PNL+IP}$ method trains IMM with the combination of the PNL and truncated inventory penalty reward ($\eta > 0, \beta = 0$). We select the hyperparameters that result in the maximum PnLMAP value on the validation dataset for these models. The hyperparameters resulting in the maximum PnLMAP value on the validation dataset are chosen for these models. The results on the FU dataset are outlined in Table 4. Here, the metric #T signifies the number of fills normalized by the episode length.



(a) Adverse Selection Ratio **(b) Number of Fills**

Figure 6: The daily adverse selection ratios and normalized number of fills.

	EPnL[10^3]↑	MAP[unit]↓	PnLMAP↑	#T
IMM$_{PnL}$	58.76 ± 94.43	2156 ± 655	31 ± 48	4.43 ± 0.94
IMM$_{PnL+C}$	42.86 ± 123.04	2041 ± 465	27 ± 68	4.85 ± 1.09
IMM$_{PnL+IP}$	**73.07 ± 53.83**	756 ± 289	90 ± 46	4.42 ± 0.96
IMM	28.097 ± 10.27	**103 ± 15**	**274 ± 89**	**5.15 ± 1.19**

Table 4: IMM variations trained with different rewards on FU.

The experimental results validate the effectiveness of different reward formulations. As shown in Table 4, the IMM$_{PNL}$ strategy tends to have the most substantial inventory risk exposure. We depict an example of the intra-day performance of the IMM$_{PNL}$ policy in the right part of Figure 4. The performance varies from day to day. We observe that the IMM$_{PNL}$ agent learns to chase trends through maintaining a large inventory (> 1000). This results in poor out-of-sample performance with large variance. Therefore, the truncated inventory penalty term proves crucial in curbing blind trend-chasing tendencies.

The IMM$_{PnL+C}$ strategy also grapples with elevated inventory risk, yet it achieves a greater number of transactions #T compared to IMM$_{PnL}$. The IMM$_{PnL+IP}$ strategy achieves the largest average terminal wealth and the return per trade metric while having the lowest #T, which is unfavorable for risk-averse market makers. attains the highest average terminal wealth and the return per trade metric, but concurrently records the lowest #T, a circumstance that may be less advantageous for risk-averse market makers. The strategy trained with the proposed reward significantly improved the return-to-risk performance with the lowest MAPs, as well as a larger #T, compared to the IMM$_{PnL+IP}$ strategies. Despite having the lowest average terminal wealth, the proposed IMM strategy acts very stably and might be the most favorable policy among these four policies for a risk-averse market maker. Besides, note that the proposed strategy has the largest #T, so it could receive more compensation from the exchange.

6 Conclusion

In this paper, we propose IMM, a novel RL-based approach aimed at efficiently learning multi-price-level MM policies. IMM first introduces efficient state and action representations. Subsequently, it pre-trains an SL-based prediction model to generate multiple trend signals as effective auxiliary observations. Furthermore, IMM utilizes a TCSA network to handle the temporal and spatial relationships in noisy financial data. Through abstracting trading knowledge from a sub-optimal expert, while interacting with the environment, IMM explores the state and action spaces efficiently. Experiments on four futures markets demonstrate that IMM outperforms the benchmarks, and further ablation studies verify the effectiveness of the components in the proposed method. Since it is important for market makers to preempt order priorities in inactive markets, we would like to take the order cancellation into account and investigate the automatic MM strategies for less liquid markets in the future.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and helpful suggestions. This work is supported by the National Natural Science Foundation of China (Grant No. 62306088 and Grant No.62161146004) and Songjiang Lab (Grant No.SL20230309).

Contribution Statement

Author 1 (First Author): Conceptualization, Methodology, Software, Investigation, Formal Analysis, Writing (Original Draft); **Author 2 (First Author)**: Data Curation, Methodology, Software, Visualization, Writing (Original Draft); **Author 3 & 4**: Data Curation, Conceptualization, Investigation, Validation; **Author 5**: Validation, Writing (Review & Editing); **Author 6 (Corresponding Author)**: Methodology, Funding Acquisition, Resources, Supervision, Writing (Review & Editing); **Author 7 (Corresponding Author)**: Data Curation, Resources, Supervision, Writing (Review & Editing).

References

[Abernethy and Kale, 2013] Jacob D. Abernethy and Satyen Kale. Adaptive market making via online learning. In *NIPS*, 2013.

[Amihud and Mendelson, 1980] Yakov Amihud and Haim Mendelson. Dealership market: Market-making with inventory. *Journal of Financial Economics*, 8:31–53, 1980.

[Avellaneda and Stoikov, 2008] Marco Avellaneda and Sasha Stoikov. High frequency trading in a limit order book. *Quantitative Finance*, 8:217–224, 04 2008.

[Cartea *et al.*, 2015a] Álvaro Cartea, Ryan Francis Donnelly, and Sebastian Jaimungal. Enhancing trading strategies with order book signals. *Applied Mathematical Finance*, 25:1 – 35, 2015.

[Cartea *et al.*, 2015b] Álvaro Cartea, Sebastian Jaimungal, and Jose S. Penalva. Algorithmic and high-frequency trading. 2015.

[Chakraborty and Kearns, 2011] Tanmoy Chakraborty and Michael Kearns. Market making and mean reversion. In *ACM Conference on Economics and Computation*, 2011.

[Chan and Shelton, 2001] Nicholas Tung Chan and Christian R. Shelton. An electronic market-maker. 2001.

[Chung *et al.*, 2022] Gu Yoon Chung, Munki Chung, Yongjae Lee, and Woo Chang Kim. Market making under order stacking framework: A deep reinforcement learning approach. *Proceedings of the Third ACM International Conference on AI in Finance*, 2022.

[Fang *et al.*, 2021] Yuchen Fang, Kan Ren, Weiqing Liu, Dong Zhou, Weinan Zhang, Jiang Bian, Yong Yu, and Tie-Yan Liu. Universal trading for order execution with oracle policy distillation. In *AAAI*, 2021.

[Fernandez-Tapia, 2015] Joaquin Fernandez-Tapia. *Modeling, optimization and estimation for the on-line control of trading algorithms in limit-order markets*. PhD thesis, 09 2015.

[Fujimoto *et al.*, 2018] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[Gašperov and Kostanjčar, 2021] Bruno Gašperov and Zvonko Kostanjčar. Market making with signals through deep reinforcement learning. *IEEE Access*, 9:61611–61622, 2021.

[Glosten and Milgrom, 1985] Lawrence R. Glosten and Paul R. Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14:71–100, 1985.

[Guéant and Manziuk, 2019] Olivier Guéant and Iuliia Manziuk. Deep reinforcement learning for market making in corporate bonds: Beating the curse of dimensionality. *Applied Mathematical Finance*, 26:387 – 452, 2019.

[Guéant *et al.*, 2012] Olivier Guéant, Charles-Albert Lehalle, and Joaquin Fernandez-Tapia. Dealing with the inventory risk: a solution to the market making problem. *Mathematics and Financial Economics*, 7(4):477–507, sep 2012.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Huang *et al.*, 2013] Weibing Huang, Charles-Albert Lehalle, and Mathieu Rosenbaum. Simulating and analyzing order book data: The queue-reactive model. Papers 1312.0563, arXiv.org, December 2013.

[Jerome *et al.*, 2022] Joseph Jerome, Gregory Palmer, and Rahul Savani. Market making with scaled beta policies. *Proceedings of the Third ACM International Conference on AI in Finance*, 2022.

[Jumadinova and Dasgupta, 2010] Janyl Jumadinova and Prithviraj Dasgupta. A comparison of different automated market-maker strategies. 2010.

[Ke *et al.*, 2017] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3149–3157, Red Hook, NY, USA, 2017. Curran Associates Inc.

[Konda and Tsitsiklis, 1999] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

[Lim and Gorse, 2018] Ye-Sheen Lim and Denise Gorse. Reinforcement learning for high-frequency market making. In *ESANN*, 2018.

[Niu *et al.*, 2022] Hui Niu, Siyuan K Li, and Jian Li. Metatrader: An reinforcement learning approach integrating diverse policies for portfolio optimization. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022.

[Patel, 2018] Yagna Patel. Optimizing market making using multi-agent reinforcement learning. *ArXiv*, abs/1812.10252, 2018.

[Sadighian, 2019] Jonathan Sadighian. Deep reinforcement learning in cryptocurrency market making. *arXiv: Trading and Market Microstructure*, 2019.

[Silver *et al.*, 2014] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. PMLR, 2014.

[Spooner *et al.*, 2018] Thomas Spooner, John Fearnley, Rahul Savani, and Andreas Koukorinis. Market making via reinforcement learning. In *AAMAS*, 2018.

[Sun *et al.*, 2023] Shuo Sun, Xinrun Wang, Wanqi Xue, Xiaoxuan Lou, and Bo An. Mastering stock markets with efficient mixture of diversified trading experts. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 2109–2119, New York, NY, USA, 2023. Association for Computing Machinery.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[Xu *et al.*, 2022] Ziyi Xu, Xue Cheng, and Yangbo He. Performance of deep reinforcement learning for high frequency market making on actual tick data. In *AAMAS*, 2022.

[Yu and Koltun, 2015] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[Zhong *et al.*, 2020] Yueyang Zhong, YeeMan Bergstrom, and Amy R. Ward. Data-driven market-making via model-free learning. In *IJCAI*, 2020.