Robust Market Making via Adversarial Reinforcement Learning

Extended Abstract

Thomas Spooner
Department of Computer Science
University of Liverpool
t.spooner@liverpool.ac.uk

Rahul Savani
Department of Computer Science
University of Liverpool
rahul.savani@liverpool.ac.uk

ABSTRACT
We show that adversarial reinforcement learning can be used to develop market marking strategies that are robust to adversarial and adaptively chosen market conditions.

KEYWORDS
Adversarial Reinforcement Learning; Robustness; Market Making

ACM Reference Format:

1 INTRODUCTION
Market makers provide liquidity by offering to buy and sell a financial instrument, and generate profits when they manage to buy and sell evenly. However, a market maker (MM) faces inventory risk, where toxic agents exploit their informational or technological advantages to preempt or cause prices to move unfavourably for the MM as it inadvertently builds a (positive or negative) inventory. This phenomenon has been the subject of much research in the optimal control [4], artificial intelligence [1], and reinforcement learning (RL) literature [10].

In this paper, we develop market making agents that are robust to adversarial and adaptively chosen market conditions by applying adversarial RL (ARL). We start from a well-known model of market making [2], which has been used extensively in quantitative finance [3–6]. We convert this model to a discrete-time game, with a "market player," an adversary that can be thought of as a proxy for other market participants that would like to profit at the expense of the market maker. The adversary controls the dynamics of the market environment in a zero-sum game against the MM.

We evaluate the sensitivity of RL-based strategies to three parameters of the model dynamics affecting price and execution for the MM, each of which naturally vary over time in real markets; this feature has not received much attention so far as existing works consider a single instantiation of parameters for the underlying model. We go beyond a fixed parametrisation — henceforth called the fixed setting — with two extended learning settings: the random setting initialises each instance of the model with (bounded) uniformly random values for the three parameters; the strategic setting features the previously mentioned "market player", a learner whose objective is to minimise cumulative reward in a zero-sum game with the market maker. The random and strategic settings are more realistic than the fixed setting and pose a significantly more difficult challenge for the market making agent. We show that MM strategies trained in each of these settings yield significantly different behaviour, and, moreover, that adversarial training has benefits beyond its immediately obvious implications.

2 TRADING MODEL
We consider a standard model of market making [2, 3] in which an MM trades a single asset of price $Z_{t+1} = Z_t + b_n\Delta t + \sigma_n W_n$, where $b_n$ and $\sigma_n$ are the drift and volatility coefficients. Randomness in this model derives from a sequence of independent Normal random variates, $W_n$, each with zero mean and variance $\Delta t$. The process begins with initial value $Z_0 = z$ and continues until step $N$ is reached. At each step, the MM places limit orders about $Z_n$ at which it is willing to buy (bid) and sell (ask), denoted by $p_{n}^b$ and $p_{n}^a$, respectively; these may be updated at each timestep at no cost.

The probability of orders being executed is dictated by market liquidity and the prices $p_{n}^\pm$. These interactions are modelled by independent Poisson processes, denoted by $N_{n}^\pm$ and $N_{n}^\mp$ for the bid/ask sides, respectively, with intensities $\lambda_{n}^\pm$. The agent’s inventory is then captured by the difference between these two terms, $H_{n} = (N_{n}^+ - N_{n}^-) \in [H, H]$, where $H_0$ is known and the values of $H_{n}$ are constrained such that trading stops on the opposing side of the book when either limit is reached. The arrival intensities are defined by $\lambda_{n}^\pm = A_{n}^\pm e^{-k_{n}^\pm |p_{n}^\pm - Z_{n}|}$, where $A_{n}^\pm$, $k_{n}^\pm > 0$ describe the rate of market order arrivals and distribution of volume in the book.

In this framework, the evolution of the market maker’s cash is given by the difference relation, $X_{n+1} = X_n + p_{n}^a \Delta N_{n}^+ - p_{n}^b \Delta N_{n}^-$, where $\Delta N_{n}^\pm = N_{n+1}^\pm - N_{n}^\pm$. We have that the cash flow is a combination of the profit due to executing at prices away from $Z_n$; and the change in value of the MM’s inventory. The total value accumulated by the MM by timestep $n$ may thus be expressed as the sum of the cash held and value invested $\Pi(X, H, Z) = X + HZ$. This is known as the mark-to-market (MtM) value of the MM’s portfolio.

Game formulation. The proposed model can be used to define a zero-sum stochastic game between an MM and an adversary, which acts as a proxy for all other market participants.

Definition 2.1. [Market Making Game] The game has $N$ stages. At each stage, MM chooses $p^\pm$ and the adversary $b$, $A^\pm$, and $k^\pm$. The stage payoff is given by expected change in MtM value of the MM’s portfolio, i.e., $E[\Pi]$. The total payoff paid by the adversary to MM is the sum of the stage payoffs.

In the full paper [11] we theoretically analyse Nash equilibria (NE) of the single-stage game ($N = 1$) for the two cases of variable $\{b\}$ with fixed $A^\pm$ and $k^\pm$, and variable $b$, $A^\pm$ and $k^\pm$, and empirically...
show a correspondence between these equilibria and the solutions to the multi-stage game that we find with adversarial RL.

3 ADVERSARIAL TRAINING

ARL is used to adaptively penalise policies of the MM that are not robust, i.e., those that are susceptible to exploitation by the adversary. While there are no guarantees our ARL will reach an NE, we show that ARL: consistently converges to an approximate NE, and outperforms past approaches in terms of raw performance and robustness to model ambiguity. This approach is adapted from [8] to support incremental actor-critic methods and asynchronous training. The adversary is trained in parallel with the market maker, and uses the same state and learning algorithm:

**States.** The state of the environment $s = (t_n, H_n)$ contains only the current time $t_n = \frac{t_n}{\delta t} = n\Delta t$ and the MM’s inventory $H_n$. Transitions are governed by the model dynamics in Section 2.

**Rewards.** The reward function is adapted from the optimisation objective pioneered by Cartea, Jaimungal, and others, to RL, where, unlike e.g. [3], the source and adaptive learning environment (bounded as in [8]), Strategic time $t \in [0, \Delta t]$.

(1) $R_n = \Delta \Pi_n - \xi H_n^2 - \left\{ \begin{array}{ll} 0 & \text{for } t < T, \\
\eta H_n^2 & \text{otherwise} \end{array} \right. \quad (1)$

Depending on $\eta$ and $\xi$, this formulation can give risk-neutral or risk-adverse preferences, e.g. if $\eta > 0$ and $\xi = 0$, then the MM is risk averse and is punished if the terminal inventory $H_T$ is non-zero.

**Algorithms.** Both agents use NAC-$\lambda$([13]) to learn stochastic policies, using semi-gradient SARSA($\lambda$) ([9]) for policy evaluation. The value functions are represented by compatible [7] radial basis function networks with accumulating eligibility traces [12].

**Learning settings.** We investigate three multi-stage variants of the Market Making Game, each with different restrictions on the adversary’s strategy. The following three types of adversary in turn increase the freedom to control the market’s dynamics:

**FIXED.** The adversary sets $b_n = 0$, $A_n^k = 140$ and $k_n^+ = 1.5$; these are chosen to match settings in [2]. This amounts to a single-agent learning setting with stationary transition dynamics.

**RANDOM.** Each episode has parameters chosen independently and uniformly at random from: $b_n = b \in [-5, 5]$, $A_n^k = A \in [105, 175]$ and $k_n^+ = k \in [1.125, 1.875]$. These are chosen at the start of each episode and remain fixed until termination. This is analogous to single-agent RL with varying transition dynamics.

**STRATEGIC.** The adversary chooses $b_n, A_n^k, k_n^+$ at each intra-episode step of the game (bounded as in RANDOM). This is a fully adversarial and adaptive learning environment where, unlike e.g. [3], the source of risk is exogenous and reactive to the policy of the MM.

4 EXPERIMENTS

Training was conducted by initialising each episode with a starting time to choose uniformly at random from $[0.0, 0.95]$, starting price $Z_0 = 100$, and inventory $H_0 \in [H = -50, H = 50]$. Prices $Z_n$ have fixed volatility $\sigma = 2$ between $[0.6, 1.1]$ with increment $\Delta t = 0.005$.

Table 1. Performance of MM policies learnt in the three different settings. Reported means and standard deviations were computed from $10^7$ test episodes against a FIXED adversary.

(a) Market makers trained against the FIXED adversary.

<table>
<thead>
<tr>
<th>$\eta$</th>
<th>$\xi$</th>
<th>Term. wealth</th>
<th>Sharpe</th>
<th>Term. inventory</th>
<th>Avg. spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>66.7 ± 11.8</td>
<td>7.65</td>
<td>0.38 ± 7.14</td>
<td>1.36 ± 0.05</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0</td>
<td>59.4 ± 7.6</td>
<td>7.79</td>
<td>0.02 ± 1.09</td>
<td>1.87 ± 0.02</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0</td>
<td>62.9 ± 8.4</td>
<td>7.51</td>
<td>-0.94 ± 1.21</td>
<td>1.68 ± 0.02</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0</td>
<td>65.8 ± 9.3</td>
<td>7.07</td>
<td>-0.18 ± 1.57</td>
<td>1.46 ± 0.03</td>
</tr>
<tr>
<td>0.0</td>
<td>0.01</td>
<td>64.0 ± 6.7</td>
<td>9.54</td>
<td>-0.06 ± 1.31</td>
<td>1.60 ± 0.02</td>
</tr>
<tr>
<td>0.0</td>
<td>0.001</td>
<td>65.9 ± 7.2</td>
<td>9.11</td>
<td>-0.07 ± 2.62</td>
<td>1.44 ± 0.02</td>
</tr>
</tbody>
</table>

(b) Market makers trained against the RANDOM adversary.

<table>
<thead>
<tr>
<th>$\eta$</th>
<th>$\xi$</th>
<th>Term. wealth</th>
<th>Sharpe</th>
<th>Term. inventory</th>
<th>Avg. spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>65.1 ± 6.7</td>
<td>9.78</td>
<td>-0.05 ± 1.94</td>
<td>1.44 ± 0.02</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0</td>
<td>60.5 ± 6.8</td>
<td>8.88</td>
<td>-0.02 ± 0.97</td>
<td>1.75 ± 0.02</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0</td>
<td>63.3 ± 6.8</td>
<td>9.32</td>
<td>-0.07 ± 1.05</td>
<td>1.60 ± 0.01</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0</td>
<td>64.8 ± 6.7</td>
<td>9.72</td>
<td>-0.06 ± 1.37</td>
<td>1.49 ± 0.02</td>
</tr>
<tr>
<td>0.0</td>
<td>0.01</td>
<td>62.9 ± 6.7</td>
<td>9.43</td>
<td>-0.03 ± 1.19</td>
<td>1.65 ± 0.02</td>
</tr>
<tr>
<td>0.0</td>
<td>0.001</td>
<td>64.3 ± 6.5</td>
<td>9.85</td>
<td>0.00 ± 1.71</td>
<td>1.44 ± 0.01</td>
</tr>
</tbody>
</table>

(c) Market makers trained against a STRATEGIC adversary.

<table>
<thead>
<tr>
<th>$\eta$</th>
<th>$\xi$</th>
<th>Term. wealth</th>
<th>Sharpe</th>
<th>Term. inventory</th>
<th>Avg. spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>65.1 ± 6.7</td>
<td>9.78</td>
<td>-0.05 ± 1.94</td>
<td>1.44 ± 0.02</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0</td>
<td>60.5 ± 6.8</td>
<td>8.88</td>
<td>-0.02 ± 0.97</td>
<td>1.75 ± 0.02</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0</td>
<td>63.3 ± 6.8</td>
<td>9.32</td>
<td>-0.07 ± 1.05</td>
<td>1.60 ± 0.01</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0</td>
<td>64.8 ± 6.7</td>
<td>9.72</td>
<td>-0.06 ± 1.37</td>
<td>1.49 ± 0.02</td>
</tr>
<tr>
<td>0.0</td>
<td>0.01</td>
<td>62.9 ± 6.7</td>
<td>9.43</td>
<td>-0.03 ± 1.19</td>
<td>1.65 ± 0.02</td>
</tr>
<tr>
<td>0.0</td>
<td>0.001</td>
<td>64.3 ± 6.5</td>
<td>9.85</td>
<td>0.00 ± 1.71</td>
<td>1.44 ± 0.01</td>
</tr>
</tbody>
</table>

as in Table 1a, but with half the standard deviation. Interestingly, we note that in Table 1c the MM achieves this with much lower variance on terminal inventory, even in the risk-neutral case. These observations are shown to hold when the MMs are evaluated in the RANDOM and STRATEGIC environments, suggesting improved invariance to model specification. We also verify, using empirical best response computation, that the solutions found by ARL are (approximately) Nash equilibria of the corresponding game.

5 CONCLUSIONS

We have introduced a new approach for deriving trading strategies with ARL that are robust to the discrepancies between the market model in training and testing. We show that our approach leads to strategies that outperform previous methods in terms of PnL and Sharpe ratio, and have comparable spread efficiency. This is shown to be the case for out-of-sample tests in all three of the proposed settings. In other words, our MMs are not only more robust to misspecification, but also dominate in overall performance, regardless of the reward function used. We verify empirically that ARL finds equilibria of the multi-stage stochastic game, and that in some special cases that these correspond to equilibria in the corresponding single-stage game.

2015
REFERENCES