ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Reinforcement Learning for High-Frequency Market Making

Ye-Sheen Lim and Denise Gorse

University College London - Computer Science
Gower Street, London, WC1E 6BT - UK

Abstract. In this paper we present the first practical application of reinforcement learning to optimal market making in high-frequency trading. States, actions, and reward formulations unique to high-frequency market making are proposed, including a novel use of the CARA utility as a terminal reward for improving learning. We show that the optimal policy trained using Q-learning outperforms state-of-the-art market making algorithms. Finally, we analyse the optimal reinforcement learning policies, and the influence of the CARA utility from a trading perspective.

1 Introduction

The mathematical modelling of market dynamics is extremely challenging. Even in their full complexity, mathematical modelling of financial markets is insufficient in capturing the reality of financial systems. These issues are even more salient in *high-frequency trading* (HFT), which is a relatively new field of great interest to both financial institutions and market regulators.

HFT involves the use of high-speed communication technologies and accurate real-time market data to make trades in timescales of seconds down to microseconds. HF traders generally function as *market makers*. Market making refers to the act of *liquidity provision* by simultaneously quoting a *bid* (buy) price and an *ask* (sell) price on an asset. *Liquidity* is loosely defined as the quantity of asset available for trade. Profit is earned by the market maker in the form of the *spread* between the quoted price placed on the buy and sell prices.

Relatively little work has been done to model HF market making. The work done in [1, 2] assumes a market maker operating in a low-frequency and quote-driven environment reminiscent of pre-modern financial markets. The work closest in content to that presented here is that of [3]. The authors applied stochastic approximation to fit the *Avellaneda-Stoikov* model [4], which is well-known as an industrial standard.

In this paper we present a formulation of a discrete Q-learning algorithm for market making, test it against the model used in [3], and analyse the resulting optimal policy. We make a novel use of the CARA utility [5] to improve learning based on the measure of an agent's risk aversion. We demonstrate RL outperforms analytical models such as the one used in [3].

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

2 Background and Related Work

2.1 Limit Order Book

Modern financial markets are order-driven and decentralised, where any trader can submit bid or ask *limit orders* (LO) for a quantity of asset at a specific limit price. If the limit price of an order cannot be satisfied on arrival to an exchange, the LO is added to a *limit order book* (LOB) to await execution against subsequent orders arriving at the exchange. LOs in the LOB can also be cancelled. Traders can also submit a *market order* (MO), which has no limit price and is always immediately executed at the best price in the LOB.

Hence, a LOB is a continuous double auction where traders compete in buying and selling an asset by submitting bid and ask orders. Transactions can occur anytime an existing LO in the book is matched to an incoming order by a *matching engine*. Typically the matching engine of a LOB implements the *price-time priority rules* for matching an incoming order to LOs in the book. For priority of execution, the LOs in the book are ranked first by their price, and second by the time the order arrived at the exchange. Readers are directed to [6] for a comprehensive introduction to LOBs.

In this paper we are interested in simulating an environment which reproduces the stochastic dynamics of LOBs at HF for learning market making policies. The dynamics of LOBs are governed by the random arrivals of so-called *order book events* [7]: LO submission, LO cancellation and LO executions. To this end we use the Poisson model of order book event arrivals of [8] to simulate the limit order book. The model has been shown to reproduce important statistical properties of empirical order books, and more importantly is derived in a form that is suitable for use as a reinforcement learning environment.

2.2 High-Frequency Market Making

HF market makers provide liquidity by posting simultaneous bid and ask quotes, and making profit off the spread, while cancelling and resubmitting orders at high speed to react to minute changes in the market. The main objective of a market maker is to minimise inventory risk, i.e. to avoid being in possession of large amounts of inventory which might unexpectedly lose its value.

Market making control policies determine the offsets from the best bid and best ask in the LOB at which to post the bid and ask quotes. In the earlier years of HFT, market making strategies set zero tick (a *tick* is the lowest price change allowed by the exchange) offsets but this nowadays an unprofitable strategy. [9] proposed to determine the offset based on market conditions such as market volatility and the arrival rates of MOs. [10] used a so-called reservation spread determine whether or not to post LOs. [4] described a complex optimal control approach that became an industrial standard.

The cited papers above proposed strategies derived from mathematical models making strong assumptions about market behaviour. Also, the models rely heavily on parameters that have to be fitted, possible unreliably from recent

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

market data. In practice these parameters are subjected to regime changes. Finally, the inherently continuous and non-linear action space of the models proved to be very unstable; very small perturbations can lead to extreme control values. In order to develop more effective strategies, it is necessary to move beyond a reliance on such models into a framework that learns effective strategies from experience; this will be the subject of the work to be described below.

3 RL Formulation

3.1 States

In what follows, we will assume Markovian state transitions, and treat the partially observable environment as if it were fully observable. The ultimate test of these assumption will lie in the later empirical evaluation of the algorithm.

The two states of primary concern here are: the *inventory* i, and the *time remaining* τ. To reduce the computational complexity, the inventory states are binned to six states representing *small*, *medium* and *large* inventory imbalance for either direction, with an additional state representing zero inventory. The trading period T is discretised into k timesteps; hence we have the remaining time as $\tau \in \{\frac{1}{k}T, \frac{2}{k}T, \ldots, T\}$.

3.2 Actions

Since all exchanges in practice impose a minimum in the price change, called the *tick*, the action space is naturally discrete. We define the action of simultaneously quoting bid/ask limit orders at a given timestep as a tuple $a = (d_b, d_a)$, where $d_b \in \mathbb{Z}$ is the number of ticks lower than the best bid to quote a bid limit order, and $d_a \in \mathbb{Z}$ is the number of ticks higher than the best ask to quote an ask limit order. At every timestep, we cancel any unexecuted orders previously placed and submit new limit orders according to the action a selected from the optimal policy given the current states. All orders placed are for a unit size of 100 shares.

3.3 Rewards

Existing approaches to the market making problem generally seek to maximise the expected utility of some economic measure of the agent. These utilities are meant to evaluate the performance of the agent at the end of a trading period and do not accurately represent *immediate* rewards at each timestep t. We propose a more suitable utility through the reward function R_t described below

$$R_t = a(V_t - V_{t-1}) + e^{b\tau_t} sgn(|i_t| - |i_{t-1}|) \quad , \tag{1}$$

where a and b are constants, V_t is the *value* of the agent at time t, i_t is the *inventory* of the agent at time t, and τ_t is the remaining trading time at t.

After the end of a trading period, we introduce a terminal reward based on the *constant absolute risk aversion* (CARA) utility [5] to represent the attitude

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

of the agent to the gains or losses caused by having inventory i_T at the end of the trading period. The CARA utility is an exponential in the form of

$$R_T = \alpha - \exp(-r(C_T - i_T S_T)) \quad , \tag{2}$$

where α is a constant, r is the risk aversion parameter, C_T is the profit or loss made during the trading period, and S_T is the average price (including costs) at which we can immediately liquidate i_T shares.

4 Experiments

The discrete Q-learning algorithm [11] is used to find the optimal action-selection policy. Function approximation is not required, as the states and actions are naturally discrete. Since we do not have a good starting policy for an environment as complex as a high-frequency market, we use an off-policy algorithm in preference to an on-policy algorithm for better exploration.

During learning, the optimal actions are chosen ϵ-greedily. Both ϵ and the learning rate are set to diminish as more episodes are run. Each trading period is set to 120 seconds, with $k = 12$ timesteps and hence 12 time states. For the inventory state i, we define *small* inventory as $0 < i \leq 200$, *medium* inventory as $200 < i \leq 400$, and *large* inventory as $i > 400$. With this, there are in total 66 combinations of possible time and inventory states.

The Poisson model described previously is used to simulate a dynamic limit order book. The agent is trained for 10000 episodes. In every episode, the simulation is first run for 300 seconds to initialise the order book. Selecting the optimal offsets from Q, the agent submits simultaneous bid and ask limit orders every 10 seconds until the end of the trading period. Any previously submitted orders that have not been executed are cancelled. Q is then updated using the reward function R_t as described above. At the end of the trading period, the terminal reward R_T is used to update the value of the last encountered state, regardless of the action taken. By doing this the CARA utility is propagated through all previous states, letting the agent take into account its risk aversion in accumulating inventory throughout the trading period.

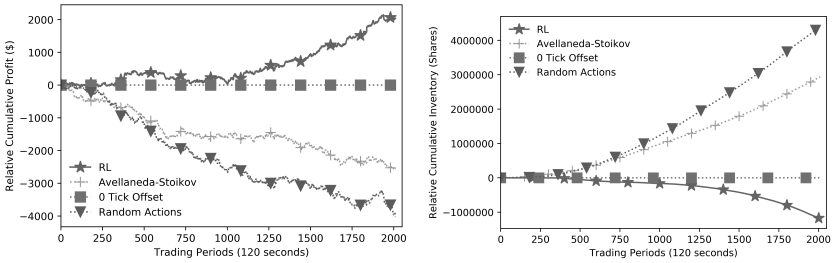
The performance of RL will be compared to other market making algorithms. In RL, the market maker can choose to quote an offset from the set $\{0, 1, 2\}$ ticks for each bid and ask side respectively, giving a total of 9 different tuples as actions. The Zero Tick Offset method is the simplest form of market making where the bid and ask prices of the limit orders are set to the best bid and best ask. We also include the Avellaneda-Stoikov [4] model, which is still regarded as state-of-the-art. Finally, we have Random Actions where the actions are randomly choosen from the action set available to RL.

4.1 Results

We simulate trading for 2000 trading periods. At the end of each period, the total inventory accumulated is immediately liquidated with a MO and the final

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

total profit obtained by the agent is then computed. Since the Zero Tick Offset algorithm is the simplest approach, for comparison we plot the difference between the cumulative profit of each market making algorithm and that of the Zero Tick Offset throughout the 2000 trading periods. This can be seen in Figure 1a, with Figure 1b the cumulative inventory relative to Zero Tick Offset.



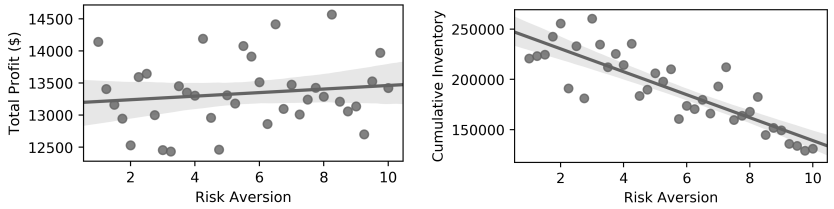
(a) Cumulative profit (b) Cumulative inventory

Fig. 1: Cumulative profit and cumulative inventory, relative to the Zero Tick Offset method, after 2000 trading periods for different market making algorithms

From Figure 1a we can see that RL clearly outperforms in terms of profit all the other methods, including the mathematical model of *Avellaneda-Stoikov*. In addition, Figure 1b demonstrates that RL remains the most inventory neutral.

4.2 Influence of the CARA Utility

The use of the CARA utility as the terminal reward enables the agent to learn an optimal policy that suits its attitude to the risk of inventory imbalance.



(a) Final cash vs risk aversion (b) Cumulative inventory vs risk aversion

Fig. 2: Final cash and cumulative inventory after 2000 periods for agents with increasing risk aversion, including a fitted line, with shaded standard deviation

Figure 2a shows the total profit earned after trading for 2000 periods for agents with increasing risk aversion. At the end of each trading period, the inventory is liquidated with a MO and the total profit is added to a cumulative sum. Figure 2b shows the cumulative sum of the inventory that has to be liquidated at the end of every trading period, as function of risk aversion. As

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

this is increased, the accumulated inventory decreases almost linearly, reducing the exposure of the agent to liquidation costs and costs due to natural price movements.

One can observe in Figure 2a that although a linearly increasing trend can be fitted to the total profit, multiple runs with the same risk aversion can have slightly varying total profit due to the stochastic nature of price movements. The liquidation cost of a unit of the asset may never be the same at two different points in time. The risk aversion parameter in the CARA utility represents the willingness of the agent to risk these natural price movements.

5 Discussion

The reinforcement learning formulation presented above has been shown to outperform the market making framework proposed by [4], which is still considered state-of-the-art in the literature. However, our use of the CARA utility has demonstrated that there is a lot of potential in the intersection between mathematical models and machine learning methods. An alternative to the methodology of this paper, to be considered in future work, would be the use of reinforcement learning to instead learn to tune the actions of the algorithm derived from [4], depending on the limit order book states; this would be a further example of the combination of reinforcement learning and mathematical modelling.

References

[1] Nicholas Tung Chan and Christian Shelton. An electronic market-maker. Technical report, MIT, 2001.

[2] Adlar J Kim, Christian R Shelton, and Tomaso Poggio. Modeling stock order flows and learning market-making from data. Technical report, MIT, 2002.

[3] Joaquin Fernandez-Tapia. High-frequency trading meets reinforcement learning: Exploiting the iterative nature of trading algorithms. 2015.

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

[5] Bruce A Babcock, E Kwan Choi, and Eli Feinerman. Risk and probability premiums for cara utility functions. *Journal of Agricultural and Resource Economics*, pages 17–24, 1993.

[6] Martin D Gould, Mason A Porter, Stacy Williams, Mark McDonald, Daniel J Fenn, and Sam D Howison. Limit order books. *Quantitative Finance*, 13(11):1709–1742, 2013.

[7] Nikolaus Hautsch and Ruihong Huang. Limit order flow, market impact and optimal order sizes: evidence from nasdaq totalview-itch data. 2011.

[8] Rama Cont, Sasha Stoikov, and Rishi Talreja. A stochastic model for order book dynamics. *Operations research*, 58(3):549–563, 2010.

[9] Irene Aldridge. *High-frequency trading: a practical guide to algorithmic strategies and trading systems*, volume 459. John Wiley and Sons, 2009.

[10] Thierry Foucault, Ohad Kadan, and Eugene Kandel. Limit order book as a market for liquidity. *The review of financial studies*, 18(4):1171–1217, 2005.

[11] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.