An Application of Deep Reinforcement Learning for Order Execution

with a supporting analysis of Optimal Liquidation under a Transient Price Impact Model

Matthew B. Reiter

Division of Engineering Science, University of Toronto

May 1, 2020

Supervised by Dr. Y. Lawryshyn, University of Toronto
With guidance from Dr. M. Fukasawa, Osaka University

Agenda

Preamble

The following resources are available to accompany this presentation:

- The complete thesis ...
  https://mbreiter.github.io/doc/thesis.pdf
- A summary of the thesis ...
  https://mbreiter.github.io/doc/thesis_summary.pdf
- A copy of these slides ...
  https://mbreiter.github.io/doc/thesis_slides.pdf

Introduction

A trader who needs to sell or purchase a large quantity of shares has to decide on an execution strategy to follow.

- Order Execution formalizes the models and optimization objectives which arise from profit maximizing trading incentives.
- Deep Reinforcement Learning methods offer an attractive models-free approach to Optimal Order Execution, where the complicated factors that drive asset prices can be handled under reduced modelling assumptions.
- A notable academic contribution in the space of Deep Reinforcement Learning comes from the work of [Ning et al., 2018], for adapting a *Double-Deep Q-Network* that trains under an Optimal Order Execution directive.

Objectives and Recommendations (1 of 3)

The primary objective is to develop an Asynchronous Advantage Actor-Critic (A3C) model for Optimal Order Execution, within the context of Volume-Weighted Average Price (VWAP) execution.

- **Recommendation**: The A3C model proves able to modestly track VWAP. The limitations of the modelling scheme introduce an opportunity to consider an enhanced reward signal, whereby an *Advice & State-Driven* learning model is a prime candidate for supplemental training.

Objectives and Recommendations (2 of 3)

A supporting objective is to demonstrate that a Teacher/Student learning framework can train a reference model, referred to as a *Student*, at an accelerated rate when compared to a baseline model.

- **Recommendation**: For the Teacher/Student learning problem, it is effective to initialize the Student model with the network parameters from the Teacher model. Additionally, a significant reduction in training time is observed under the *Advice & State-Driven* decisioning approach – that is, when the Student model receives both Teacher-recommended advice and an observation on the state as feature inputs.

Objectives and Recommendations (3 of 3)

The final supporting objective is to explore a models-dependent approach to Order Execution and discover insights into how economic factors such as market depth, resilience, tightness and bias influence optimal trading decisions.

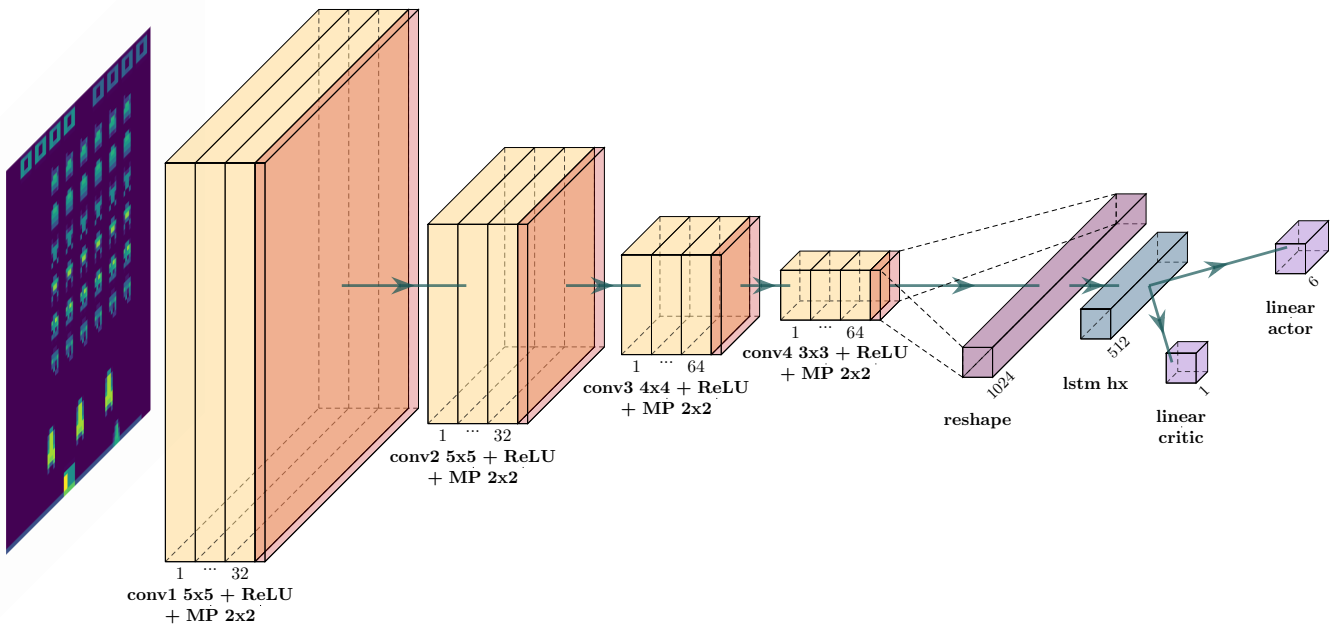
- **Recommendation**: The block shaped limit order book is a convenient setting to define a transient price impact model on and outline an Optimal Liquidation objective to. The take-away from completing this work is a characterization of the richly diverse trading policies under varying agent- and market-specific parameters.

Teacher/Student Framework: Overview

The Teacher/Student learning framework is approached from the perspective of efficiently training a Student model by processing advice from a Teacher model, as facilitated by:

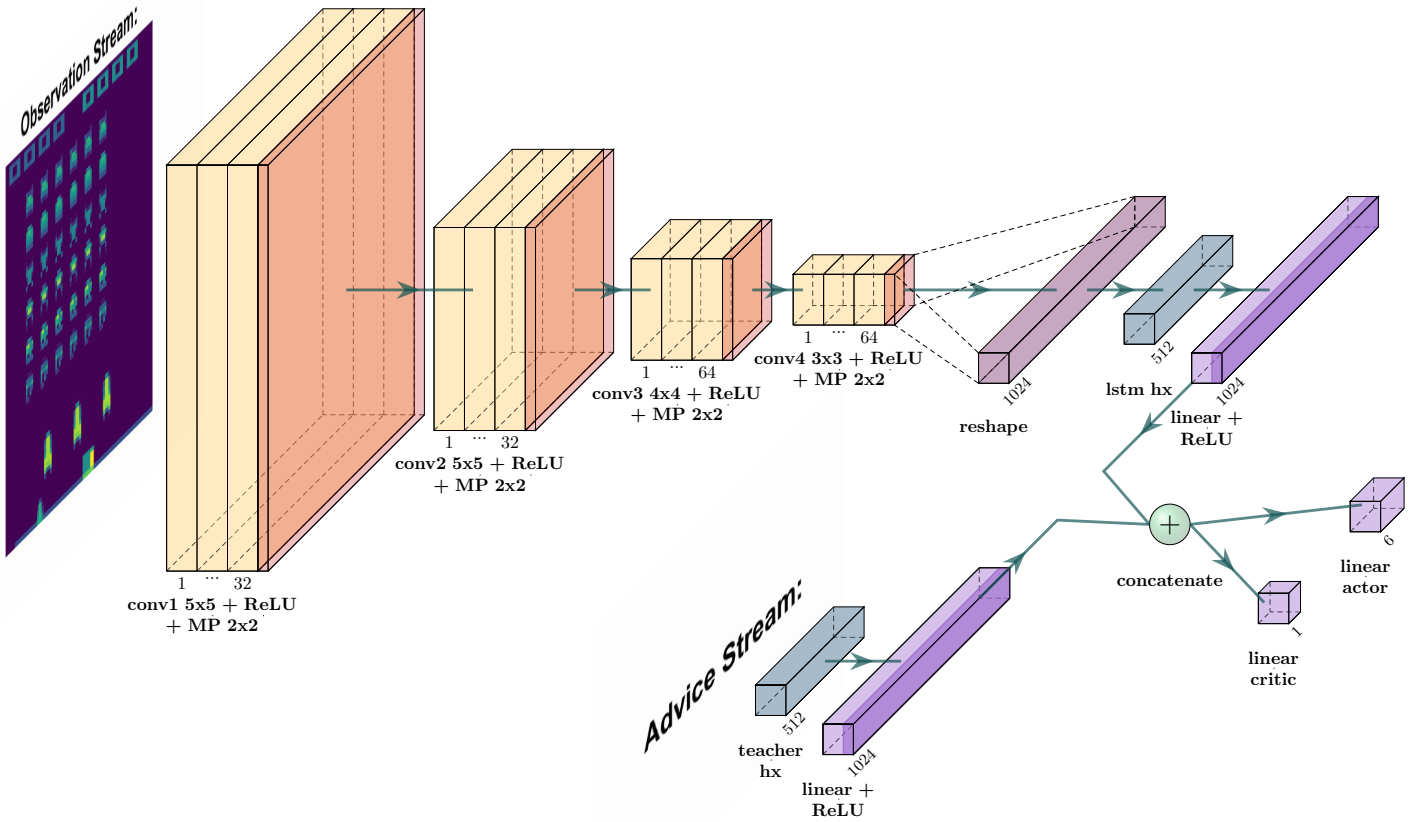
- **Advice-Driven (AD)** decisioning, which takes only Teacher-fed advice as an input.
- **Advice & State-Driven (ASD)** decisioning, which incorporates both an advice-stream and an observation-stream on the environment.

Teacher/Student Framework: Baseline



- Based on: https://github.com/dgriff777/rl_a3c_pytorch

Teacher/Student Framework: ASD Decisioning



Teacher/Student Framework: Teacher Scheme

The Teacher is incentivized to *stay alive for as long as possible*; reinforced for survival with the following signal

$$r(t) = \frac{t\eta}{t_{\max}}\left(1 + \mathbb{1}_{\{\texttt{done} \wedge \texttt{cleared}\}}\right) - \kappa \mathbb{1}_{\{\texttt{done} \wedge \neg \texttt{cleared}\}}$$

- The parameters η, κ are interpreted respectively as the survival bonus and the hit penalty.
- The performance metric to evaluate the success of the Teacher model is the average episode length over a training session.

Teacher/Student Framework: General Comments

- Training took place on three Atari 2600 games: *Breakout*, *Beamrider* and *Space Invaders*. The Student models are trained to the specifications of these original Atari 2600 games.
- The A3C algorithm is adapted to train the Teacher/Student learning framework.
- Consult the thesis document for details pertaining to the preprocessing of the state observations and the configuration of the hyperparameters.

Transient Price Impact Model: Overview

Given an inventory of shares, how should an agent trade in order to maximize her risk-adjusted utility, while balancing tradeoffs from ...

- **Impact Risk**: Trading too quickly or in large block sizes carries a penalty as per the laws of supply/demand.
- **Uncertainty Risk**: Trading too slowly compromises forecast accuracy and faces pressure from risk aversion.

Seminal work from [Robert Almgren and Neil Chriss, 1999] introduced temporary and permanent price impacts induced by the rate of trade, extended to a *block shaped limit order book* with transient price dynamics by [Anna A. Obizhaeva and Jiang Wang, 2013] and further analyzed by [Moritz Voß, 2017].

Transient Price Impact Model: Price Dynamics

Consider the trading policy γ_t defined by the non-decreasing cádlág functions for the cumulative volume of purchases $(+)$ and sales (-)

$$\gamma_t = \gamma_{0-} + \gamma_t^+ - \gamma_t^-$$

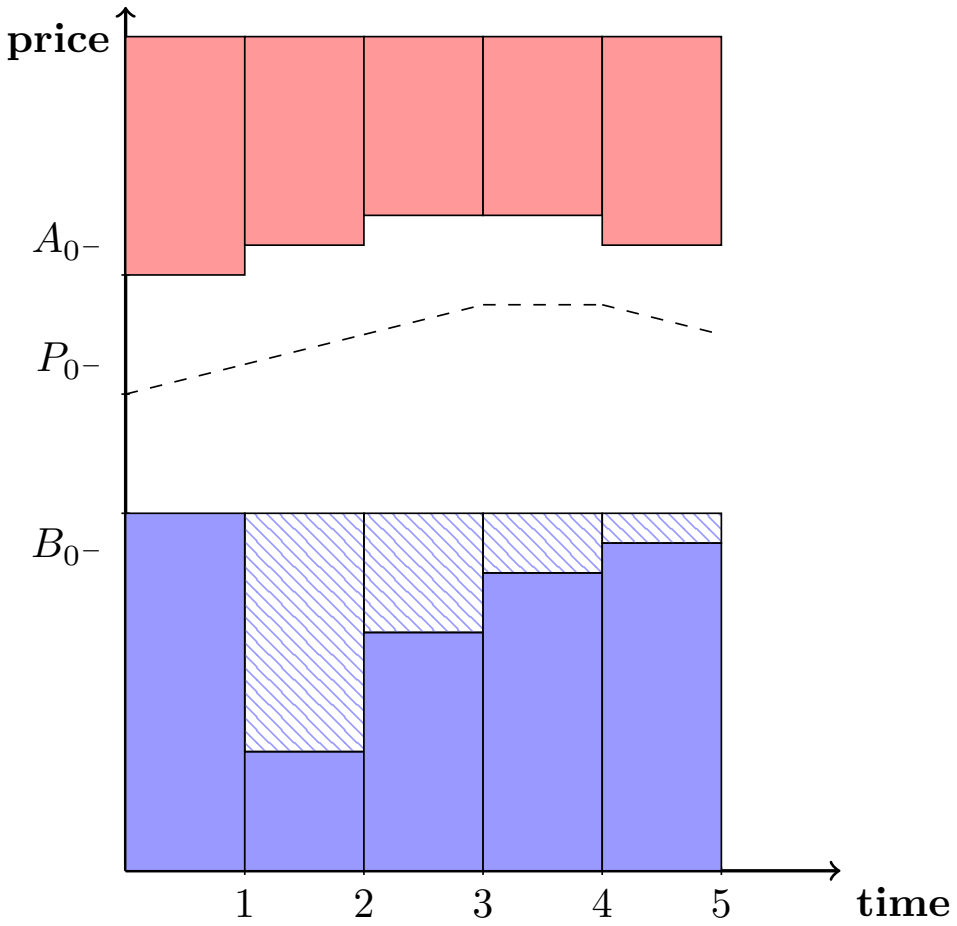
Trading activity impacts the bid- and ask-price according to

$$dA_t = dP_t + \lambda d\gamma_t^+ - \alpha(A_t - P_t)dt$$
$$dB_t = dP_t - \lambda d\gamma_t^- - \alpha(B_t - P_t)dt$$

where

- $\lambda > 0$ is the trading impact coefficient $:\rightarrow$ **"Depth"**
- $\alpha > 0$ is the market resilience rate $:\rightarrow$ **"Resilience"**
- $(P)_{t\geq 0}$ is the fundamental price process, taken in the simplest form $dP_t = \mu dt + \sigma dW_t$

Transient Price Impact Model: Price Dynamics



Transient Price Impact Model: Spread Dynamics

The bid/ask spread $S = A - B$ and the mid-quote $Q = (A + B)/2$ evolve according to

$$dQ_t = dP_t + \frac{1}{2}\lambda d\gamma_t - \alpha(Q_t - P_t)dt$$
$$dS_t = \lambda |d\gamma|_t - \alpha S_t dt$$

for $|d\gamma|_t = d\gamma_t^+ + d\gamma_t^-$. Defining $R = 2(Q - P)$,

$$R_t = R_{0-}\, e^{-\alpha t} + \lambda \int_{[0,t]} e^{-\alpha(t-u)} d\gamma_u :\to \textbf{“Bias”}$$
$$S_t = S_{0-}\, e^{-\alpha t} + \lambda \int_{[0,t]} e^{-\alpha(t-u)} |d\gamma|_u :\to \textbf{“Tightness”}$$

Remark: Under [Moritz Voß, 2017], $dQ_t = dP_t + \frac{1}{2}\lambda d\gamma_t$.

Transient Price Impact Model: Objective Statement

For some risk tolerance $\beta \geq 0$, the objective is to maximize the trader's mean-variance adjusted liquidation wealth W_τ

$$\mathbb{E}[W_\tau] - \frac{\beta}{2}\mathrm{Var}[W_\tau] \to max!$$

Transient Price Impact Model: Objective Statement

For some risk tolerance $\beta \geq 0$, the objective is to maximize the trader's mean-variance adjusted liquidation wealth W_τ

$$\mathbb{E}[W_\tau] - \frac{\beta}{2}\mathrm{Var}[W_\tau] \to max!$$

When $(P_t)_{t\geq 0}$ is an ABM with constant exogenously determined coefficients, the objective is equivalent to finding the minimizer of $J(\cdot)$

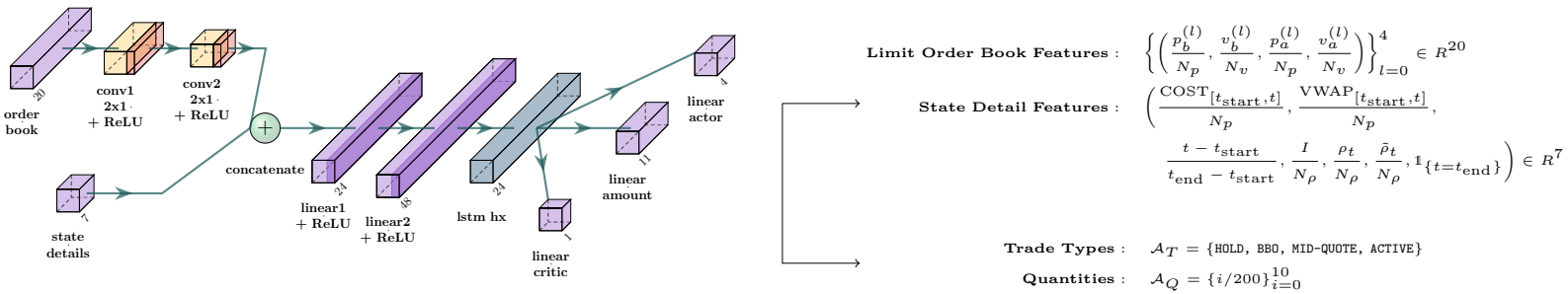
$$\begin{aligned} J(\gamma) = \frac{\beta\sigma^2}{2}\int_0^\tau \gamma_t^2 dt + \mu\int_{[0,\tau]} t d\gamma_t \\ + \frac{\lambda}{2}\left[\frac{R_{0-}}{\lambda}\int_{[0,\tau]} e^{-\alpha t} d\gamma_t + \frac{S_{0-}}{\lambda}\int_{[0,\tau]} e^{-\alpha t}|d\gamma|_t \right. \\ \left. + \int_{[0,\tau]}\int_{[0,\tau]} e^{-\alpha|t-s|} d\gamma_s^- d\gamma_t^- + \int_{[0,\tau]}\int_{[0,\tau]} e^{-\alpha|t-s|} d\gamma_s^+ d\gamma_t^+\right] \to min! \end{aligned}$$

A3C for VWAP Execution: Training Environments

Three training environments are developed that each prove conducive for Reinforcement Learning under a VWAP execution directive:

- The first environment is based on the block shaped limit order book from [Anna A. Obizhaeva and Jiang Wang, 2013]: https://github.com/mbreiter/drloe_block
- The second is based on historical limit order book data from the Nasdaq stock exchange: https://github.com/mbreiter/drloe_lobster
- The third is adapted from a simulated order-driven market maintained by the Rotman School of Management: https://github.com/mbreiter/drloe_rit

A3C for VWAP Execution: Model Details



- With order book $\{(p_b^{(k)}, v_b^{(k)}, p_a^{(k)}, v_a^{(k)})\}_{k=0}^{N}$; inventory I; quantity executed ρ_t; quantity pending $\tilde{\rho}_t$; and normalization factors $N_{(\cdot)}$.
- As per [Zhang et al., 2019], convolutional networks apply a feature-mapping to limit order book data that measures a micro-price based on order book imbalance

$$\tilde{p}^{(k)} = \frac{v_b^{(k)}}{v_b^{(k)} + v_a^{(k)}} p_a^{(k)} + \frac{v_a^{(k)}}{v_b^{(k)} + v_a^{(k)}} p_b^{(k)}$$

A3C for VWAP Execution: Reward Signal

$$r^{\pm}(t) = \underbrace{\mathbb{1}_{\left\{\left|\frac{\rho_t}{I} - \frac{t - t_{\text{start}}}{t_{\text{end}} - t_{\text{start}}}\right| \leq \epsilon\right\}} - 0.5}_{\text{incentive to track TWAP}} + \underbrace{\mathbb{1}_{\{\rho_{t-1} \neq \rho_t\}} \mathbb{1}_{\left\{\mp\left(\text{COST}_{[t_{\text{start}},t]} - \text{VWAP}_{[t_{\text{start}},t]}\right) \geq 0\right\}}}_{\text{incentive to outperform VWAP}}$$

- The novel reward signal is based on an adaptation from [Cartea and Jaimungal, 2016], which demonstrates that a closed-form VWAP strategy intrinsically tracks the Time-Weighted Average Price (TWAP) with appropriate offsets.

A3C for VWAP Execution: General Comments

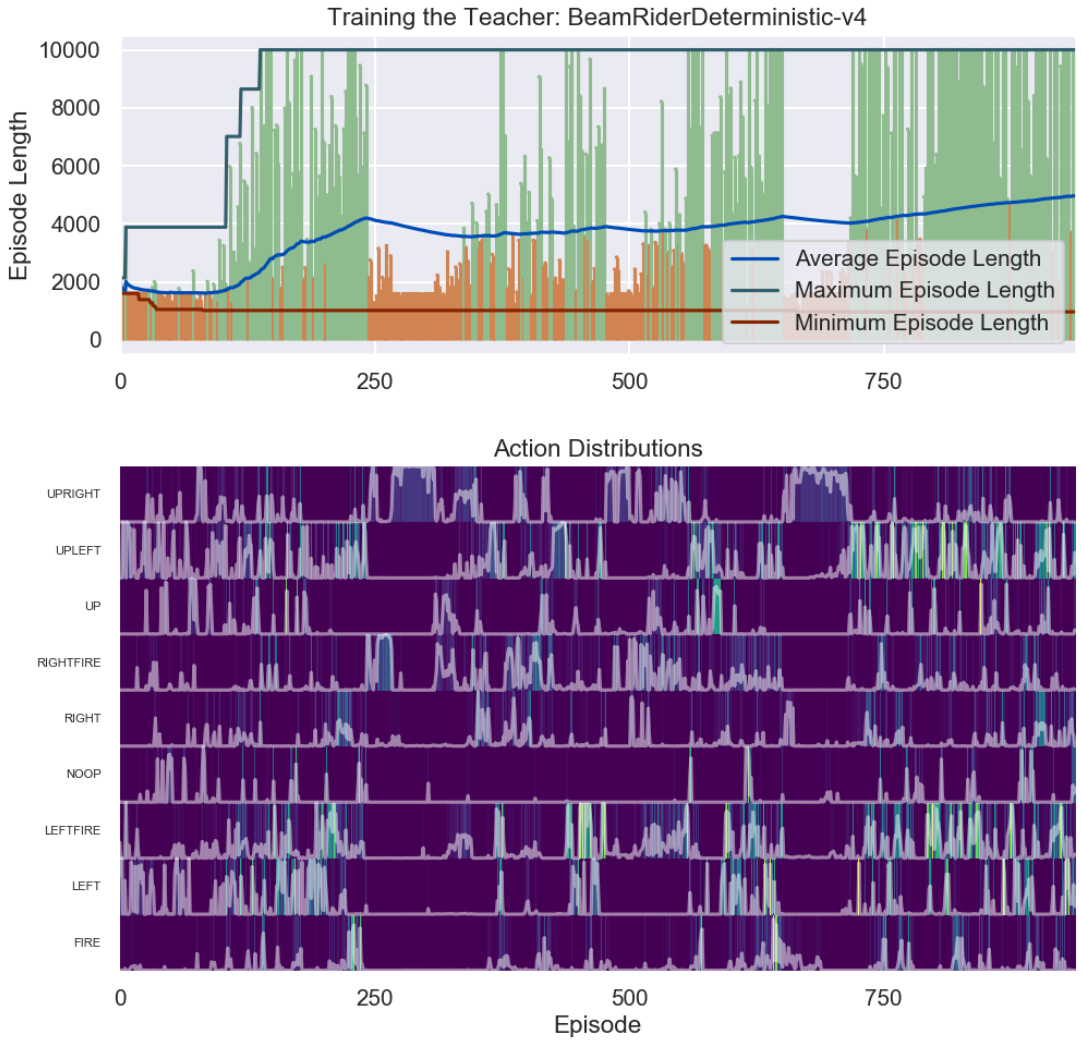
- The agent trades by specifying an action tuple $(a_t^{(T)}, a_t^{(Q)})$, where $a_t^{(T)} \in \mathcal{A}_T$ gives the trade type and $a_t^{(Q)} \in \mathcal{A}_Q$ gives the quantity to be traded.
- Trading is restricted to be unidirectional and a `MARKET-ORDER` is imposed to clear unspent inventory at the end of the horizon.
- The A3C model is executed with 16 training processes. Each training agent is assigned, uniformly at random, either a liquidation or acquisition directive. Consult the thesis for exact specifications.

Teacher/Student Framework: Baseline & Teacher Results

	Baseline Model			Teacher Model		
	Number of Episodes	Average Episode Reward	Average Episode Length	Number of Episodes	Average Episode Reward	Average Episode Length
▶ Breakout	304	202	3553	278	110	1693
▶ Beamrider	278	6793	5093	938	901	4968
▶ Space Invaders	2256	2525	2895	4101	394	1683

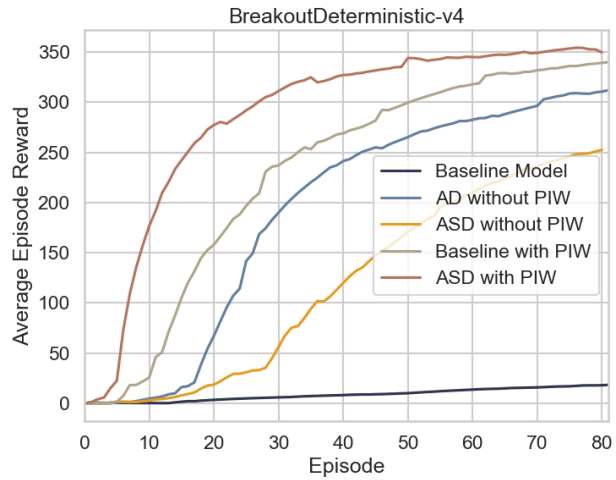
- The Baseline appears more effective, both in terms of average episode reward and episode length.
- Teacher models trained on Breakout and Beamrider were eventually able to demonstrate a policy that achieves maximum survival in their respective environment. The same result was not secured for Space Invaders.

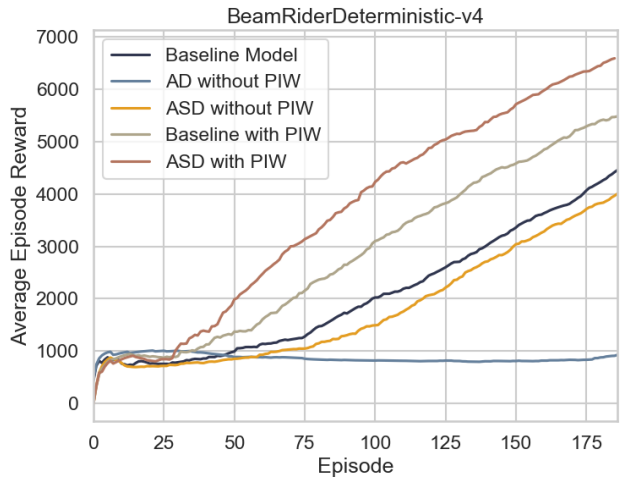
Teacher/Student Framework: Beamrider Teacher

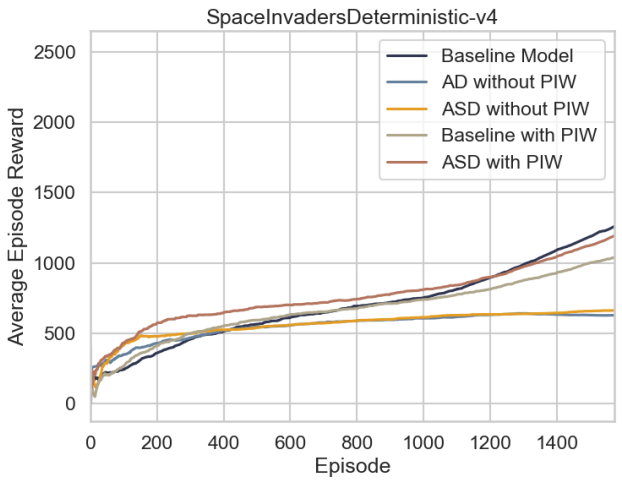


- The Teacher learns several effective survival policies.
- Consult the thesis for a detailed analysis of the performances for the Teacher models.

Teacher/Student Framework: Student Performance







- PIW stands for *preinitialized weights*, and describes the practice where network parameters from the Teacher are given to the Student.

Teacher/Student Framework: Take-aways

- The success of the Student models was achieved under different conditions in light of the environmental complexity, the relevance of the Teacher's advice, and the extent of the Teacher's own mastery of her advice.
- Such work is extendable outside the context of a Teacher/Student learning framework and offers relevance when designing a Deep Reinforcement Learning application from scratch.

Transient Price Impact Model: Recall the Objective

For some risk tolerance $\beta \geq 0$, the objective is to maximize the trader's mean-variance adjusted liquidation wealth W_τ

$$\mathbb{E}[W_\tau] - \frac{\beta}{2}\text{Var}[W_\tau] \rightarrow max!$$

When $(P_t)_{t\geq 0}$ is an ABM with constant exogenously determined coefficients, the objective is equivalent to finding the minimizer of $J(\cdot)$

$$J(\gamma) = \frac{\beta\sigma^2}{2}\int_0^\tau \gamma_t^2 dt + \mu\int_{[0,\tau]} t d\gamma_t$$
$$+ \frac{\lambda}{2}\left[\frac{R_{0-}}{\lambda}\int_{[0,\tau]} e^{-\alpha t} d\gamma_t + \frac{S_{0-}}{\lambda}\int_{[0,\tau]} e^{-\alpha t}|d\gamma|_t\right.$$
$$\left. + \int_{[0,\tau]}\int_{[0,\tau]} e^{-\alpha|t-s|} d\gamma_s^- d\gamma_t^- + \int_{[0,\tau]}\int_{[0,\tau]} e^{-\alpha|t-s|} d\gamma_s^+ d\gamma_t^+\right] \rightarrow min!$$

Transient Price Impact Model: Discretization Approach

Assumption

For an agent looking to maximize her risk-adjusted liquidation wealth, trading is only allowed over an evenly spaced time grid $\Xi = \{0, h, 2h, \ldots, \tau\}$ for $h = \tau/n$.

Transient Price Impact Model: Discretization Approach

Assumption

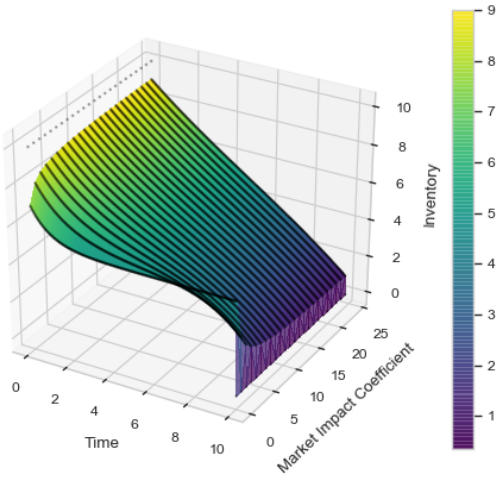
For an agent looking to maximize her risk-adjusted liquidation wealth, trading is only allowed over an evenly spaced time grid $\Xi = \{0, h, 2h, \ldots, \tau\}$ for $h = \tau/n$.

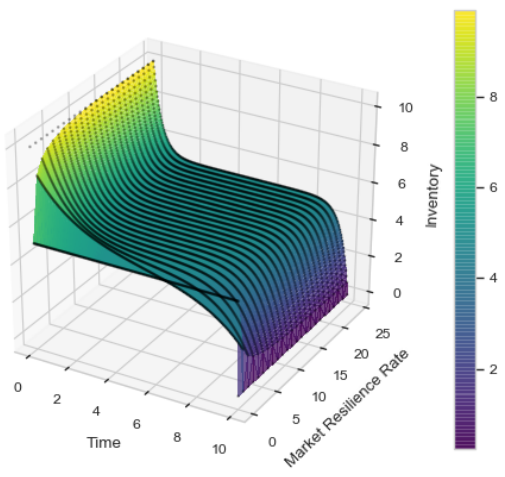
Under this framework, the minimization is now of $J_n(\gamma^n)$

$$J_n(\gamma^n) = \frac{\beta\sigma^2 h}{2} \sum_{i=0}^{n} \Big[\sum_{j=0}^{i} (\Delta\gamma_{jh}^{+} - \Delta\gamma_{jh}^{-}) \Big]^2 + \mu h \sum_{i=0}^{n} i \Big(\Delta\gamma_{jh}^{+} - \Delta\gamma_{jh}^{-} \Big) + C_n(\gamma) \to min!$$

with the cost functional $C_n(\cdot)$ and $\Delta\gamma_t^{\pm} = \gamma_t^{\pm} - \gamma_{t-}^{\pm}$.

Transient Price Impact Model: Investigating Optimality





- In the limit where $\lambda \to 0$, the agent can afford to trade in larger block sizes due to minimal market impacts.
- In the limit where $\alpha \to 0$, the agent is compelled to avoid permanent price dislocations by continuously trading.

Transient Price Impact Model: A Convexity Argument

Owing to the convexity of the problem, for any two trading policies ξ and γ having $\xi_{0-} = \gamma_{0-}$, the following holds given $\epsilon \in (0, 1]$

$$\begin{aligned} J_\tau(\xi) - J_\tau(\gamma) &\geq \lim_{\epsilon \to 0} \frac{J_\tau(\epsilon\xi + (1-\epsilon)\gamma) - J_\tau(\gamma)}{\epsilon} \\ &\geq \int_{[0,\tau]} \nabla_t^+ J_\tau(\gamma)\ (d\xi_u^+ - d\gamma_u^+) \\ &\quad + \int_{[0,\tau]} \nabla_t^- J_\tau(\gamma)\ (d\xi_u^- - d\gamma_u^-) \end{aligned}$$

First Order Optimality Condition

An absolutely continuous trading policy $\hat{\gamma}$ is optimal when $\nabla_t^{\pm} J_\tau(\hat{\gamma}) = 0$.

Transient Price Impact Model: A Variational Approach

In the spirit of [Moritz Voß, 2017], calculate the infinite dimensional buying- and selling-subgradients for $J_\tau(\gamma)$

$$\nabla_t^{\pm} J_\tau(\gamma) = \nabla_t^{\pm} D_\tau(\gamma) + \nabla_t^{\pm} L_\tau(\gamma) + \nabla_t^{\pm} Q_\tau(\gamma)$$

where

$$\nabla_t^{\pm} D_\tau(\gamma) = \begin{cases} \pm\beta\sigma^2 \int_t^\tau (\gamma_u - \frac{\mu}{\beta\sigma^2}) du & \text{if } \beta > 0 \\ \mu & \text{otherwise} \end{cases}$$

$$\nabla_t^{\pm} L_\tau(\gamma) = \frac{1}{2}(S_{0-} \pm R_{0-}) e^{-\alpha t}$$

$$\nabla_t^{\pm} Q_\tau(\gamma) = \frac{1}{2} e^{-\alpha(\tau - t)} \Big[(S_\tau \pm R_\tau) - (S_{0-} \pm R_{0-}) e^{-\alpha\tau} \Big] + \alpha \int_t^\tau \Big[(S_u \pm R_u) - (S_{0-} \pm R_{0-}) e^{-\alpha u} \Big] e^{-\alpha(u-t)} du$$

Transient Price Impact Model: Continuous-time Dynamics

Assume that γ is absolutely continuous over $(a, b) \subset (0, \tau)$. Following a routine of setting $\nabla_t^{\pm} J_\tau(\gamma) = 0$ and differentiating, as per [Moritz Voß, 2017], it can be shown

$$\ddot{\gamma}_t^{-} = \frac{\alpha^2 \beta \sigma^2}{2\alpha\lambda + \beta\sigma^2} \left(\frac{\mu}{\beta\sigma^2} - \gamma_t \right)$$

With the general form

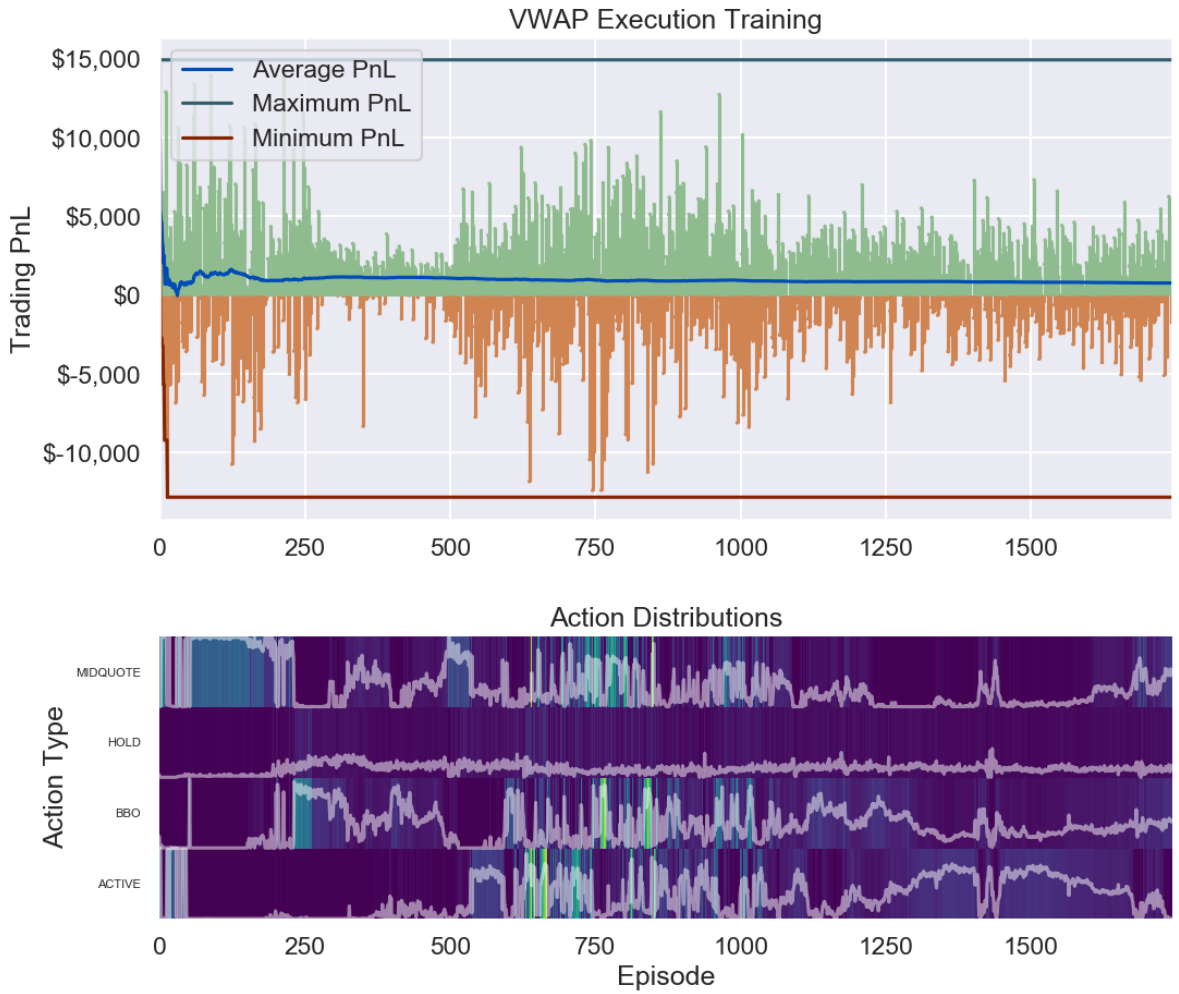
$$\gamma_t^{-} = c_+ e^{\theta t} + c_- e^{-\theta t} + \gamma_0 - \frac{\mu}{\beta\sigma^2}$$

Having a signed difference for γ_t^{+}.

A3C for VWAP Execution: The Motivation

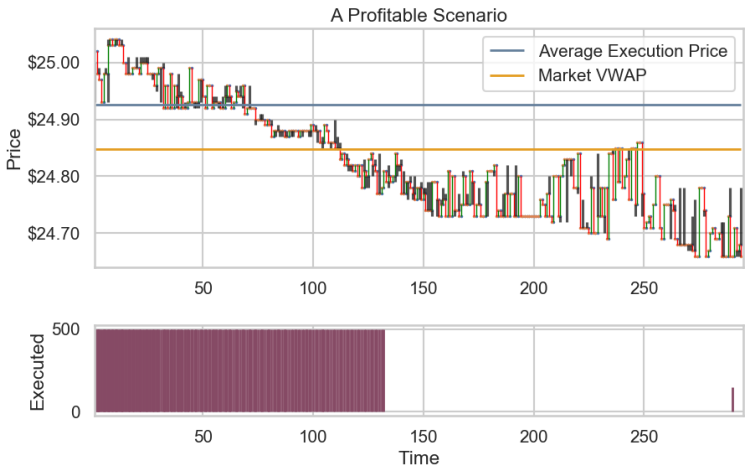
- The Teacher/Student learning framework makes familiar the Deep Reinforcement Learning approaches and gives an efficient training scheme leveraging past domain-specific knowledge.
- The transient impact model builds a relevant background for the Optimal Order Execution objective.

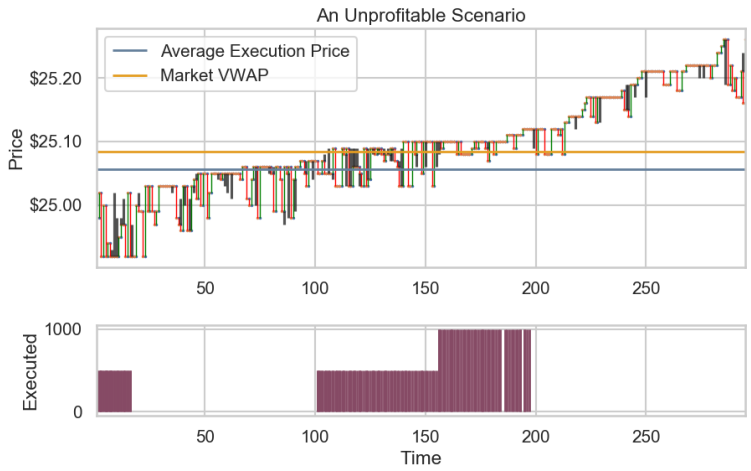
A3C for VWAP Execution: Training Results



- Training elapsed 1744 episodes on the *Rotman Interactive Trader*; each episode elapsed in real-time, over a 5 minute period.
- Profitable on average with a one-cent discount to VWAP.

A3C for VWAP Execution: Investigating the Decisioning





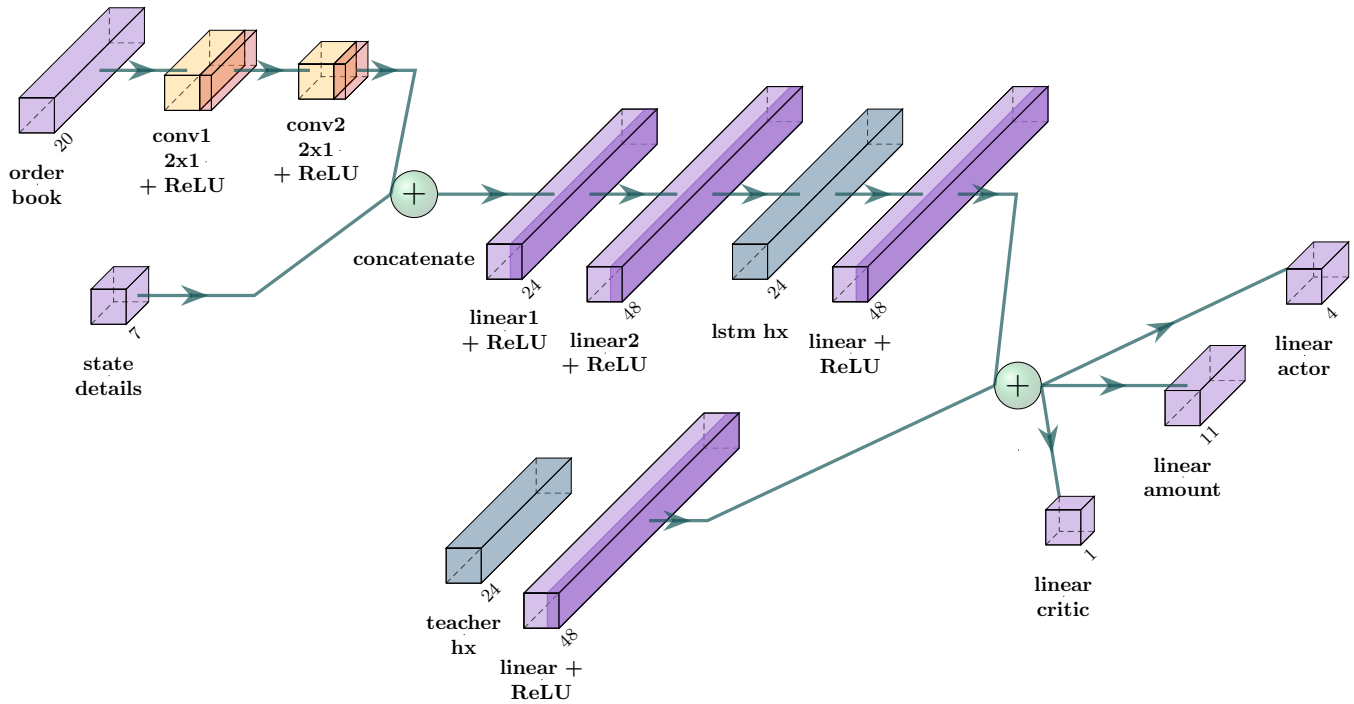
- The agent demonstrates a weak tracking of TWAP, and appears to recognize price movements.
- The agent struggles with the timing of trades.

A3C for VWAP Execution: Take-aways

- A3C for VWAP Execution is profitable *on average*, but inconsistently tracks VWAP across epsiodes.
- There is an opportunity to strengthen the outperformance incentive to VWAP with the following reward signal

$$r^{\pm}(t) = \underbrace{\mathbb{1}_{\left\{\left|\frac{\rho_t}{I} - \frac{t - t_{\text{start}}}{t_{\text{end}} - t_{\text{start}}}\right| \leq \epsilon\right\}} - 0.5}_{\text{incentive to track TWAP}} \mp \underbrace{100 \times \mathbb{1}_{\{\rho_{t-1} \neq \rho_t\}} \left(\text{COST}_{[t_{\text{start}}, t]} - \text{VWAP}_{[t_{\text{start}}, t]}\right)}_{\textbf{stronger} \text{ incentive to outperform VWAP}}$$

A3C for VWAP Execution: ASD Extensions



Concluding Remarks: Teacher/Student Learning (1 of 2)

- The success of the Teacher/Student learning framework depends on the complexity of the environment, the relevance of the Teacher's advice for the Student's objective, and the level of mastery achieved by the Teacher.
- Future work would benefit from further demonstrating the flexibility of reward signal engineering, and rigorously validating the performance of the Student learning models.

Concluding Remarks: Teacher/Student Learning (2 of 2)

- **Detailed Recommendation:** The Teacher/Student framework is evidenced to train a reference Deep Reinforcement Learning model more efficiently. To leverage this framework, the parameters for the Student model should be preinitialized from the Teacher's network, and the structural component for the Student's network should be under the *Advice & State-Driven* decisioning approach. The resulting ASD learner would be equipped to learn enriched generalizations from the Teacher's advice, and efficiently reconcile this input with the environmental observation-stream.

Concluding Remarks: Transient Price Impact Model (1 of 2)

- The transient price impact model provides an accessible entry point to the study of Optimal Order Execution. The accessibility comes from a solution method which avoids a complicated Hamilton-Jacobi-Bellman type equation, and instead uses Quadratic Programming and Variational Calculus.
- Continued work under this framework may be motivated by expanding the dimensionality of the model to cover multiple assets or by introducing a game-theoretic approach to study a Nash equilibrium for a market with competing agents.

Concluding Remarks: Transient Price Impact Model (2 of 2)

- **Detailed Recommendation:** The block shaped limit order book is a convenient setting to define a transient price impact model on and outline an Optimal Liquidation objective to. The solution modelling in discrete-time is easily achieved as a Quadratic Program, and the dynamics allow for insightful continuous-time representation. The take-away from completing this work is a characterization of the richly diverse trading policies under varying agent-specific and market-specific parameters.

Concluding Remarks: A3C for VWAP Execution (1 of 2)

- The A3C model for VWAP execution is constructed from the ground-up, and involves the careful implementation of trading environments which are conducive for Reinforcement Learning applications.
- While the A3C model for VWAP execution remains simple, and the accompanying results prove modest, the extensive detailing of the setup offers ample room for improvements. Immediately, evidence shows that the ASD learning framework would be well suited for supplemental training efforts.

Concluding Remarks: A3C for VWAP Execution (2 of 2)

- **Detailed Recommendation:** The A3C model is able to modestly track VWAP. The limitations of the modelling scheme introduce an opportunity to consider an enhanced reward signal, whereby an *Advice & State-Driven* learning model is a prime candidate for supplemental training.

Conclusion

Optimal Order Execution has challenged those who study, or practice, the subject to understand the wonderfully complicated world of market microstructure. Just as Deep Reinforcement Learning will continue to develop in maturity, so too will the connection strengthen for applications to trading.

Thank you for listening!

Acknowledgements

To the following, I offer my sincere gratitude:

Dr. Y. Lawryshyn, for continually supporting my ambitions and committing your time to ensure the success of this project.

Dr. M. Fukasawa, for introducing me to the subject of Order Execution and graciously hosting me in your laboratory group.

Mr. C. Geoffrey, for providing assistance and guidance when I needed to navigate the RIT environment.

David S, for outlining a challenging problem and offering insights to overcome the many challenges.

Zarir G, for being a steadfast mentor and friend.

My family, for your endless support. **Lauren**, I am ever thankful for your help with editing.

References I


Anna A. Obizhaeva and Jiang Wang (2013).
Optimal trading strategy and supply/demand dynamics.
Journal of Financial Markets, 16(1):1–32.


Cartea, A. and Jaimungal, S. (2016).
A closed-form execution strategy to target volume weighted average price.
SIAM J. Financial Math., 7(1):760–785.


Moritz Voß (2017).
Dynamic Hedging in Illiquid Financial Markets.
PhD thesis, Berlin Institute of Technology.


Ning, B., Ling, F. H. T., and Jaimungal, S. (2018).
Double deep q-learning for optimal execution.
CoRR, abs/1812.06600.

References II


Robert Almgren and Neil Chriss (1999).
Optimal Execution of Portfolio Transactions.


Zhang, Z., Zohren, S., and Roberts, S. J. (2019).
Deeplob: Deep convolutional neural networks for limit order books.
IEEE Trans. Signal Process., 67(11):3001–3012.