

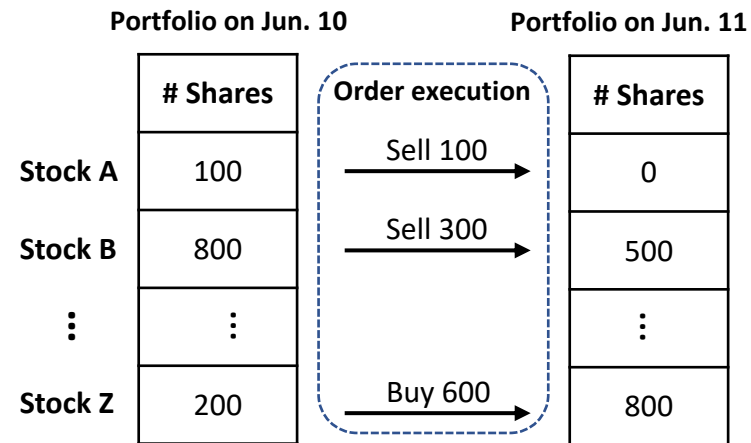
Imitation from Learning-based Oracle for Universal Order Execution in Quantitative Finance

Yuchen Fang, Kan Ren, Weiqing Liu,
Dong Zhou, Weinan Zhang, Yong Yu

Shanghai Jiao Tong University, Microsoft Research

Background

- Portfolio adjustment leads to order execution.



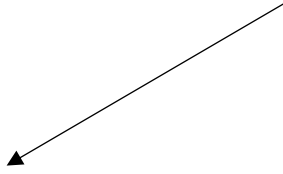
Optimal order execution

- Order execution is
 - during a trading time horizon $[0, T]$, to trade (**buy/sell**) the specific number of stock shares, at better price.

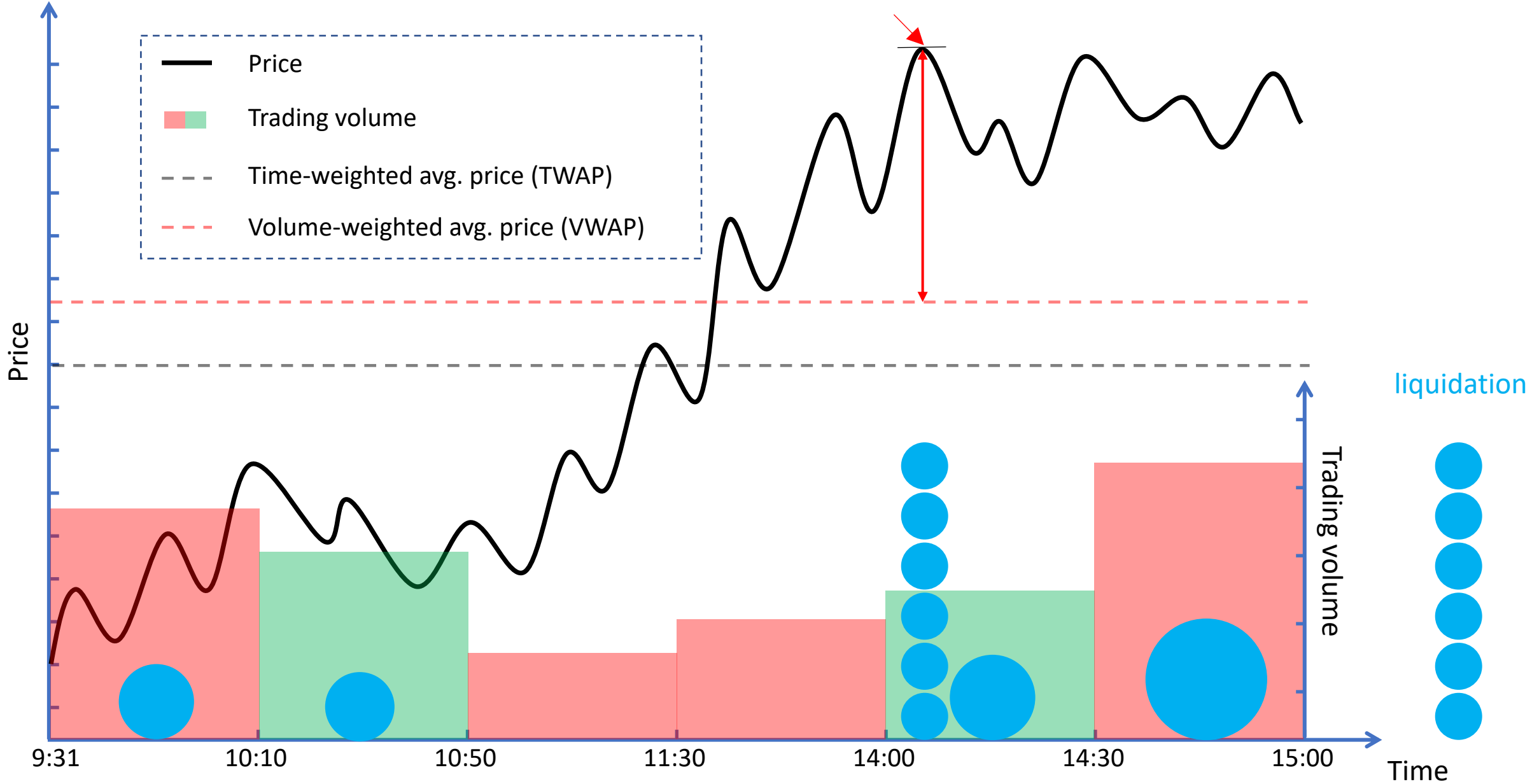
$$\min/\max \sum_{t=0}^T (p_t q_t)$$

$$\sum_{t=0}^T q_t = Q$$

Target order volume can be preset by hyper trading strategy

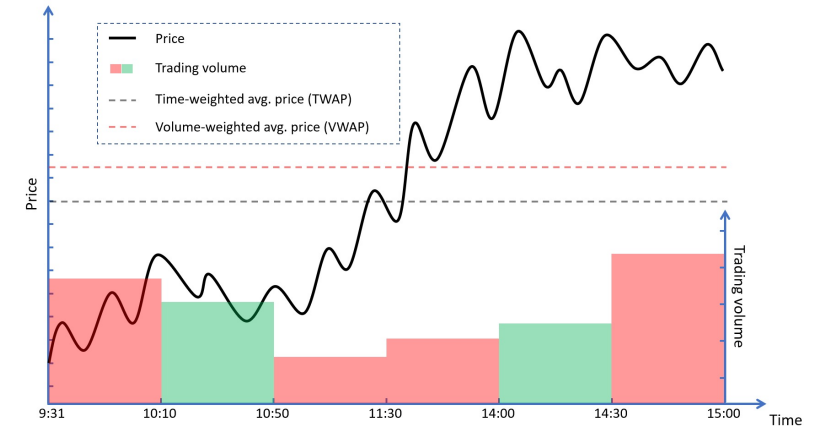


Background of **order execution**



Challenges of order execution

- Market forecasting is very hard
- Sequential trading decision making
- Higher frequency, more noise in data



Related works

Model-based analytical solution [1, 2, 3]

Practical issue

Giant gap between the practical situation and the theoretical analysis.

- First assumes market price following some process (model)
- Then solves it through stochastic control theory



$$p_t = p_{t-1} + \theta q_t + \epsilon_t, \quad \theta > 0, E[\epsilon_t | q_t, p_{t-1}] = 0.$$

White noise, i.i.d. following $N(0,1)$

[1] Bertsimas, Dimitris, and Andrew W. Lo. "Optimal control of execution costs." *Journal of Financial Markets* 1, no. 1 (1998): 1-50.

[2] Almgren, Robert, and Neil Chriss. "Optimal execution of portfolio transactions." *Journal of Risk* 3 (2001): 5-40.

[3] Cartea Á, Jaimungal S, Penalva J. *Algorithmic and high-frequency trading[M]*. Cambridge University Press, 2015.

Rethinking order execution

- It contains some *private information*, e.g., left time, left inventory, etc., to consider during decision making.
- There is a *final goal* to optimize after the whole episode, i.e., trading across the time horizon.
- It's an optimization problem with global *constraints*, i.e., the fulfillment of the target order.
- It's a typical problem of sequential decision optimization.
- Intuition: we may try direct *learning to trade* methodology.
 - A.k.a. **reinforcement learning to trade**.

Related works

Model-based analytical solution [1, 2, 3]

- First assumes market price following some process (model)
- Then solves it through stochastic control theory

Practical
issue

Giant gap between the practical situation and the theoretical analysis.

Reinforcement learning (RL) approaches

- Either extends to model-based solutions [4, 5]
- Or individually optimize for each instrument [6, 7]

Low SNR
problem

Fail to manage low signal-to-noise ratio (SNR) data

[1] Bertsimas, Dimitris, and Andrew W. Lo. "Optimal control of execution costs." *Journal of Financial Markets* 1, no. 1 (1998): 1-50.

[2] Almgren, Robert, and Neil Chriss. "Optimal execution of portfolio transactions." *Journal of Risk* 3 (2001): 5-40.

[3] Cartea Á, Jaimungal S, Penalva J. *Algorithmic and high-frequency trading*[M]. Cambridge University Press, 2015.

[4] Bao W, Liu X. Multi-agent deep reinforcement learning for liquidation strategy analysis[J]. arXiv preprint arXiv:1906.11046, 2019.

[5] Hendricks D, Wilcox D. A reinforcement learning extension to the Almgren-Chriss framework for optimal trade execution[C]//2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFER). IEEE, 2014: 457-464.

[6] Nevmyvaka Y, Feng Y, Kearns M. Reinforcement learning for optimized trade execution[C]//Proceedings of the 23rd international conference on Machine learning. 2006: 673-680.

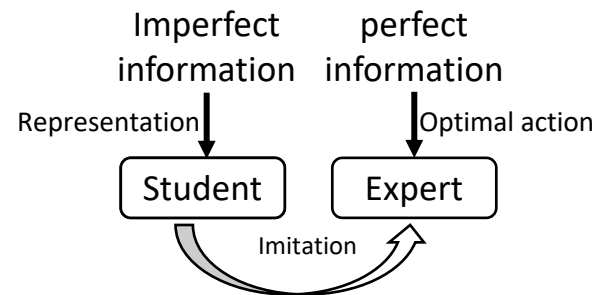
[7] Ning B, Ling F H T, Jaimungal S. Double deep q-learning for optimal execution[J]. arXiv preprint arXiv:1812.06600, 2018.

Our methodology

- Imitation from Learning-based Oracle for Universal Order Execution
- Universal trading strategy for optimal execution
 - More efficient than training over single instrument separately.
 - Learn general patterns from other instruments' data.

Our methodology

- Imitation from learning-based oracle
 - Bridge gap between representation learning and optimization decision making
 - Stabilize policy training and derive more reasonable trading strategy



Trading as a Markov decision process

- Take order execution as a direct sequential decision optimization.
- Markov decision process assumption

Notation	Markov decision process	Information
s_t	State	Private: left order to trade, timestep
		Public: market price & volume information
a_t	Action	The proportion of order to trade at the next timestep
$r_t(s_t, a_t)$	Reward	Weighted price advantage $\hat{R}_t^+ = \frac{q_{t+1}}{Q} \cdot \frac{p_{t+1} - \tilde{p}}{\tilde{p}} = a_t \left(\frac{p_{t+1}}{\tilde{p}} - 1 \right)$
		Large sub-order penalty $\bar{R}_t^- = -\alpha(a_t)^2$
γ	Discount rate	$\gamma = 1.0$

TWAP

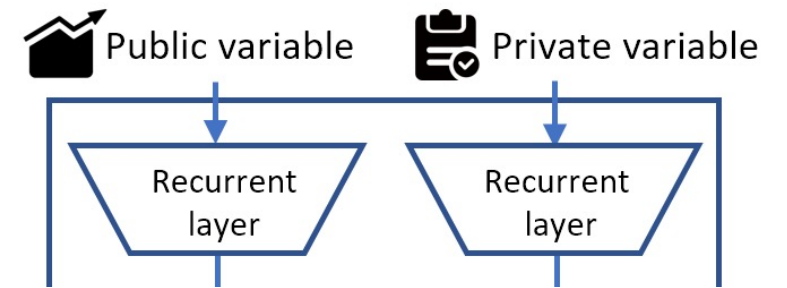
Policy Optimization for Order Execution

- Policy Optimization (PPO) algorithm (Schulman et al. 2017)
 - Decision loss for policy optimization

$$L_p(\theta) = -\mathbb{E}_t \left[\frac{\pi_{\theta}(a_t | \mathbf{s}_t)}{\pi_{\theta_{old}}(a_t | \mathbf{s}_t)} \hat{A}_{(\mathbf{s}_t, a_t)} - \beta \text{KL} [\pi_{\theta_{old}}(\cdot | \mathbf{s}_t), \pi_{\theta}(\cdot | \mathbf{s}_t)] \right]$$

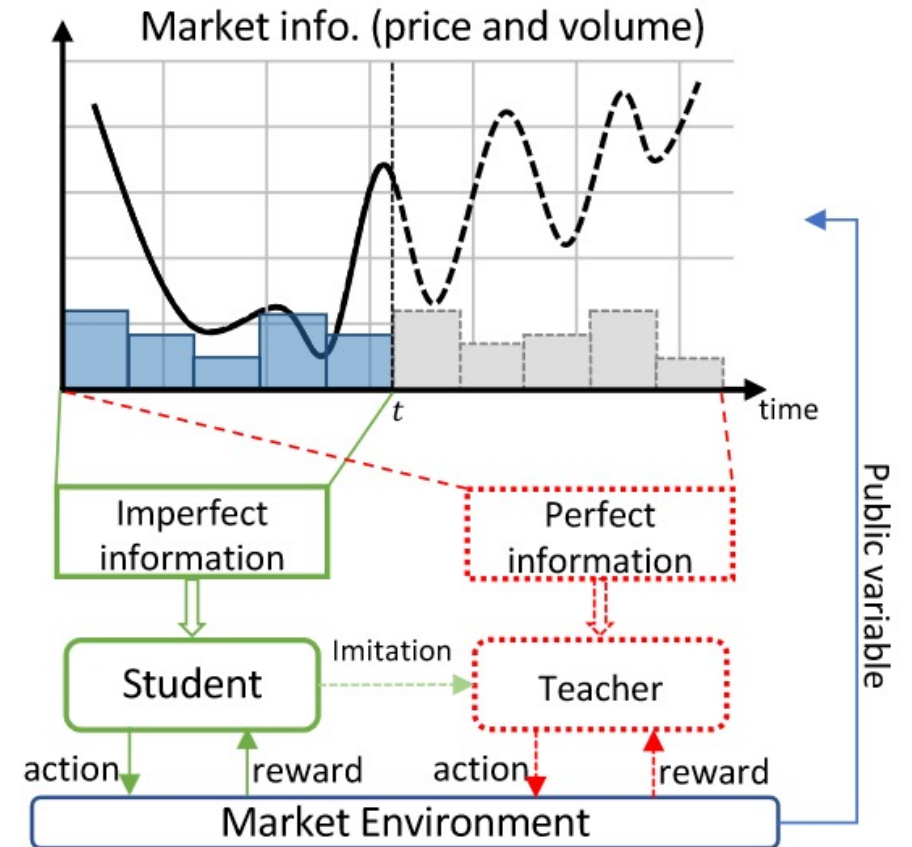
- Value function loss to improve training stability

$$L_v(\theta) = \mathbb{E}_t [\|V_{\theta}(\mathbf{s}_t) - V_t\|_2]$$



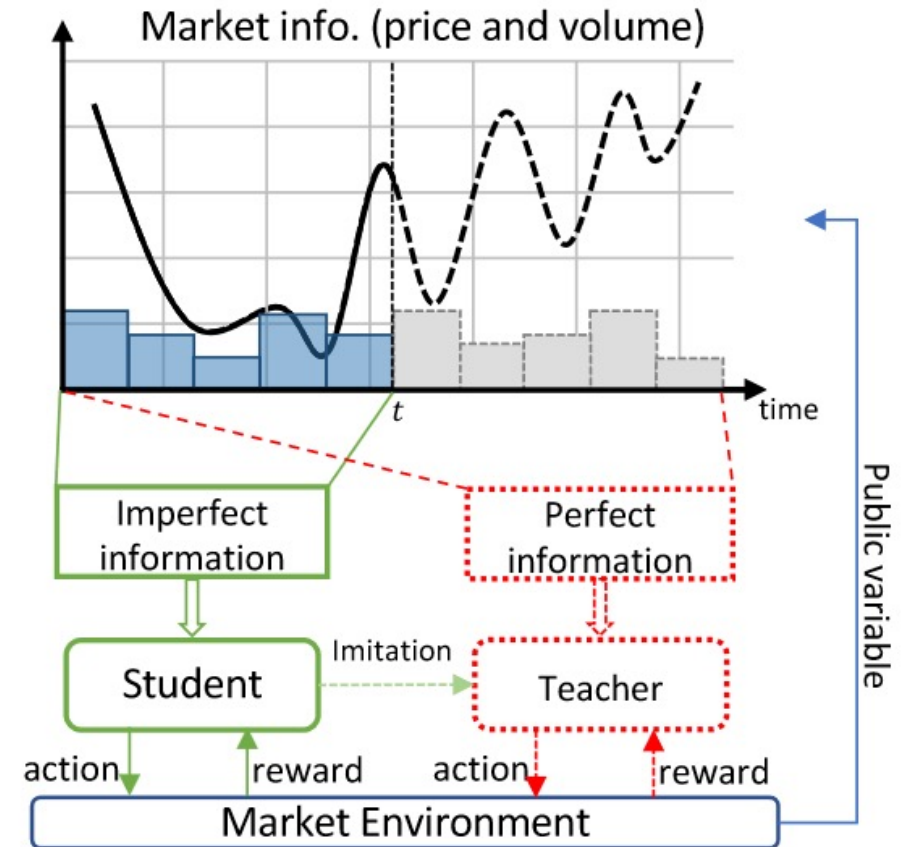
Imitation learning from oracle

- Teacher-student learning paradigm
- Oracle agent as teacher
 - *Perfect* observation
 - Interacts with the environment
 - *Approximate* the optimal trading strategy
- Common agent as student
 - Maps the imperfect market information to the optimal trading decision making



Imitation learning from oracle

- Oracle is only used for offline training
- Common agent is used for practical trading
- learning-based teacher vs searching-based teacher
 - Efficiency
 - Worse guidance
 - Extendibility



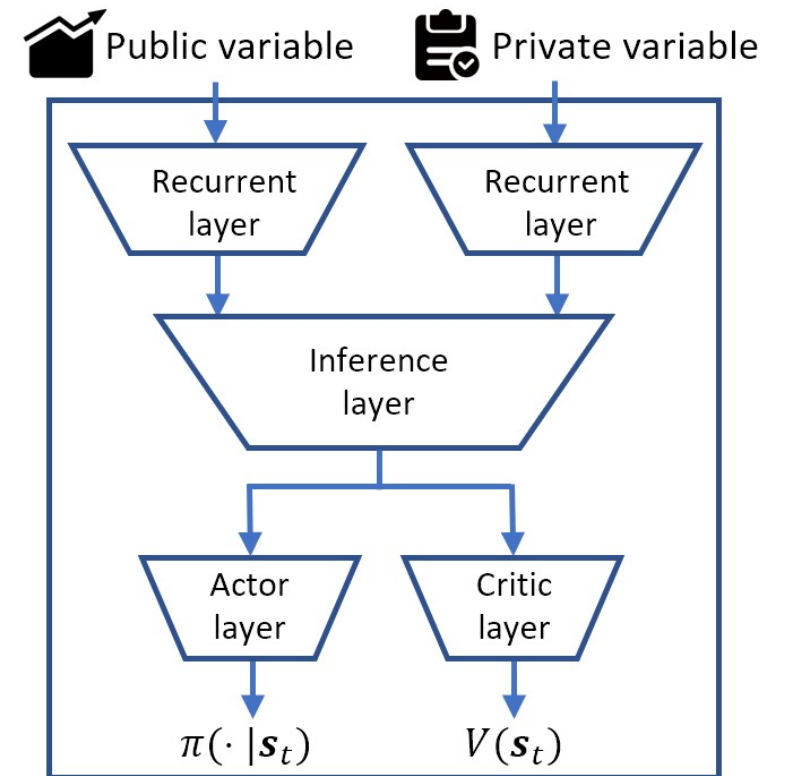
Imitation loss

- Minimize the log-likelihood loss measuring how well the student's decision matching teacher's action

$$L_d = -\mathbb{E}_t [\log \Pr(a_t = \tilde{a}_t | \pi_{\theta}, \mathbf{s}_t; \pi_{\phi}, \tilde{\mathbf{s}}_t)]$$

Student's action

Teacher's action



Experiments

- Dataset
- Compared methods
- Evaluation metrics
- Results

Dataset

- Chinese A-stock price-volume information from 2017.1.1 – 2019.6.30
- Split training and test datasets w.r.t. time.

	Training	Validation	Test
# instruments	3566	855 (CSI800)	855 (CSI800)
# order	1,654,385	35,543	33,176
Time period	1/2017 – 2/2019	3/2019 – 4/2019	5/2019 – 6/2019

Compared methods

- **TWAP** (Time-Weighted Average Price) is a strategy which splits the order into T pieces and executes the same amount of shares at each timestep. It has been proven to be the optimal strategy under the assumption that the market price follows Brownian motion (*Bertsimas and Lo 1998*).
- **AC** (Almgren-Chriss) is a model-based method (*Almgren and Chriss 2001*), which analytically finds the efficient frontier of optimal execution.
- **VWAP** (Volume-weighted Average Price) is another model-based strategy which distributes orders in proportion to the (empirically estimated) market transaction volume in order to keep the execution price closely tracking the market average price ground truth (*Kakade et al. 2004; Białkowski et al. 2008*)
- **DDQN** is a value-based RL methodology (*Ning et al. 2018*) for order execution.
- **PPO** is a policy-based RL method (*Lin and Beling 2020*) which utilized PPO with a sparse reward to train an agent with RNN for state feature extraction.
- **ILO** is our proposed methodology of policy optimization with imitation learning from oracle.
- Ablation study
 - **ILO^S** is the student policy without teacher guidance.
 - **ILO^T** is the teacher policy.

Evaluation metrics

- Reward $R = R^+ + R^-$

- Price advantage (PA) to TWAP strategy

$$\text{PA} = \frac{10^4}{N} \sum_{k=1}^{|D|} \left(\frac{\bar{p}_{\text{strategy}}^k}{\bar{p}^k} - 1 \right)$$

- Gain-loss ratio (GLR)

$$\text{GLR} = \frac{E[\text{PA} | \text{PA} > 0]}{E[\text{PA} | \text{PA} < 0]}$$

Overall performance

- The higher, the better. (* indicates p-value < 0.01)

Category	Strategy	Reward($\times 10^{-2}$)	PA	GLR
financial model- based	TWAP	-0.42	0	0
	AC	-1.45	2.33	0.89
	VWAP	-0.30	0.32	0.88
learning- based	DDQN	2.91	4.13	1.07
	PPO	1.32	2.52	0.62
	ILO ^S	3.24	5.19	1.19
	ILO	3.36*	6.17*	1.35

Table 2: Performance comparison; the higher, the better.

Learning curves

- Oracle guidance helps improve both performance and generalization.

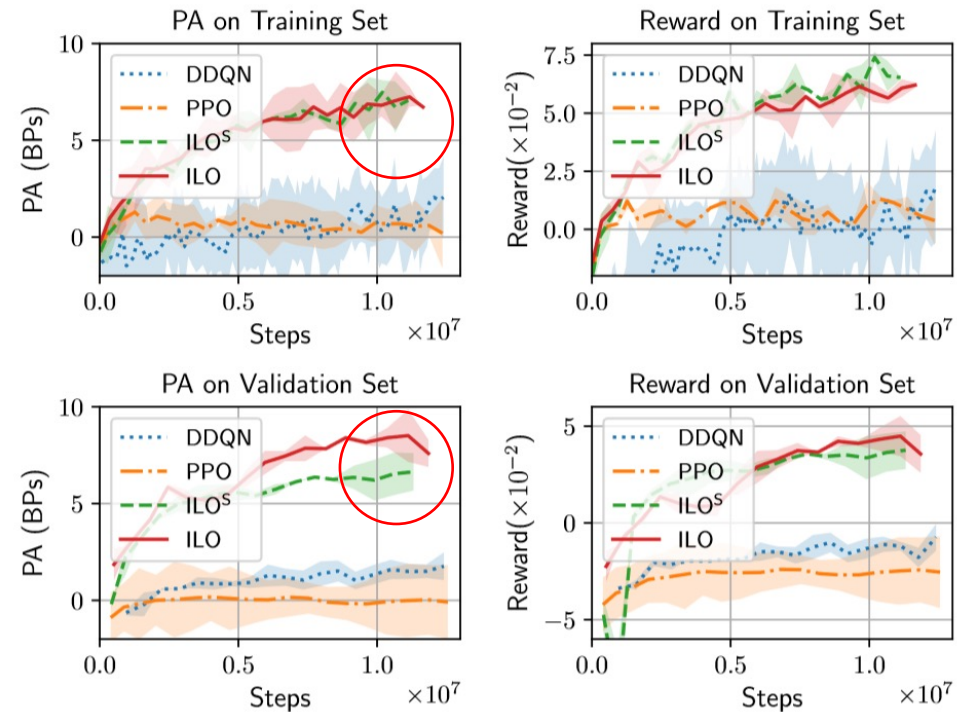


Figure 8: Learning curves (mean±std over 6 random seeds). Every step means one interaction with the environment.

Case study

- ILO illustrates the similar trading patterns to that of ILO^T (the teacher agent).

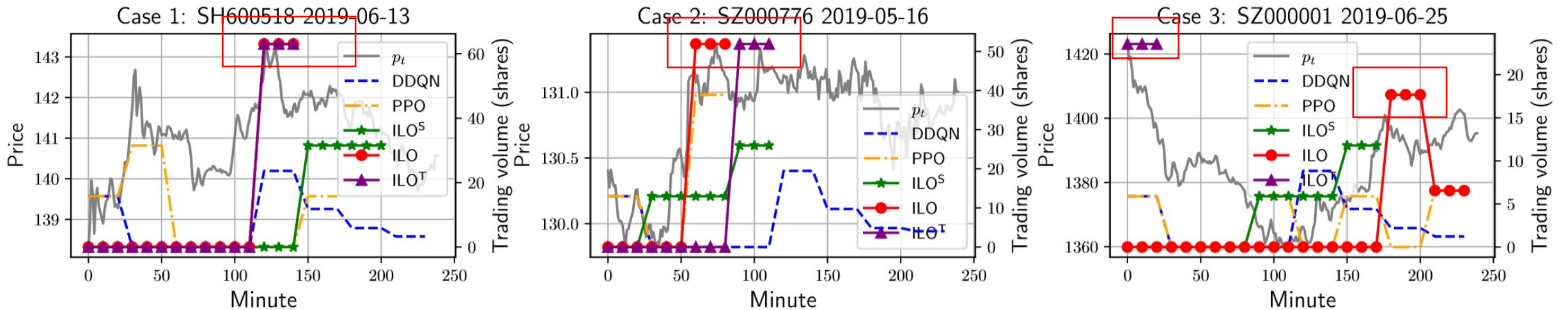


Figure 7: An illustration of execution details of different methods.

Efficiency of universal strategy

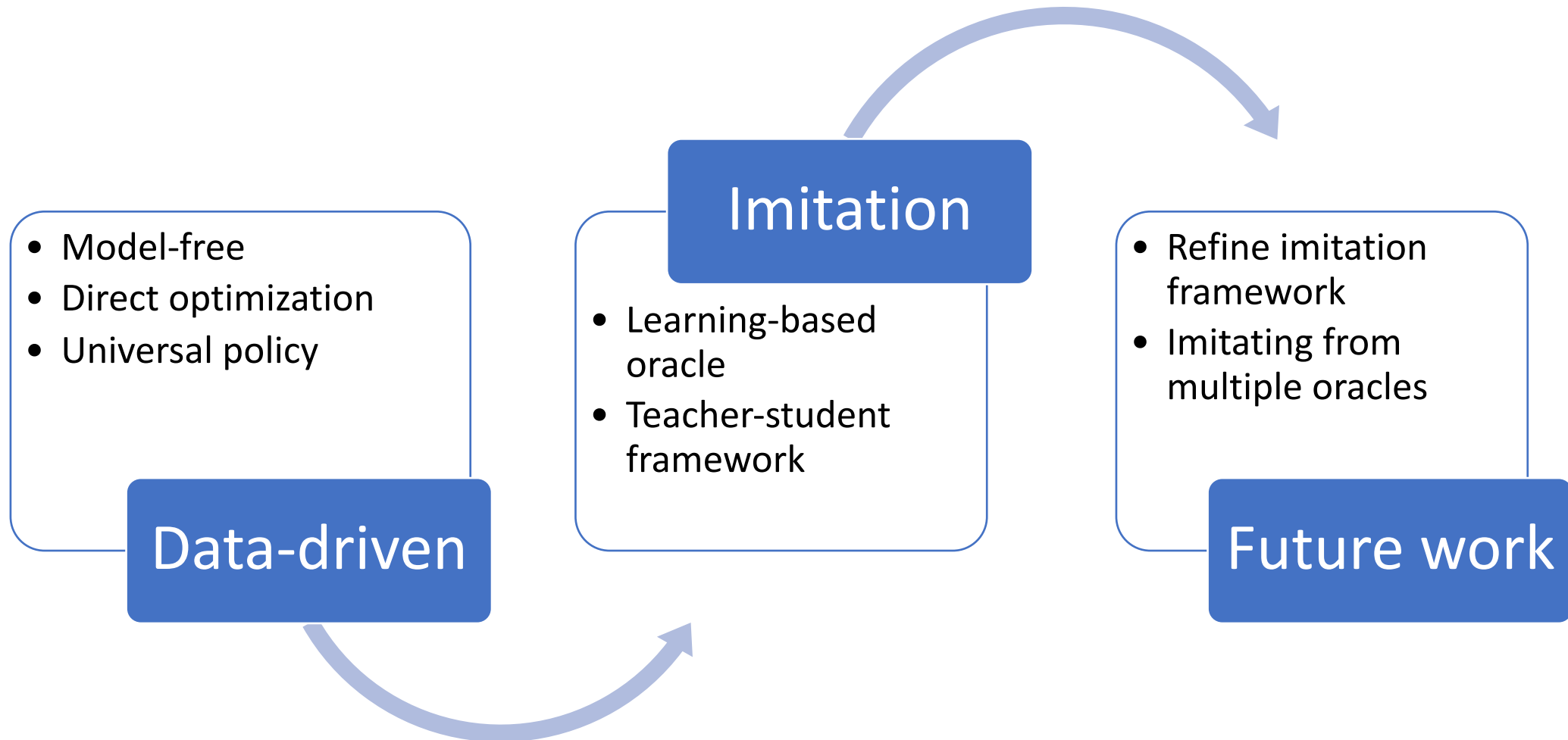
Algorithm	Policy	Reward $\times 10^{-2}$										
		Mean	SH600519	SH601318	SH600036	SH600276	SH601166	SH600030	SH600887	SH601328	SH600016	SH601288
DDQN	Single	-2.55	0.01	-3.94	0.32	-6.62	-2.13	-5.24	-5.00	0.24	-0.82	-2.33
PPO	Single	-20.62	-2.57	-12.28	-10.07	-29.19	-2.70	-3.62	-11.95	-48.70	-7.98	-77.12
ILO	Universal	0.34	7.79	2.23	1.39	-3.75	-2.28	3.28	9.57	-6.98	-4.45	-3.37
	Fine-tuned	1.82	14.30	-0.62	9.46	4.46	-7.24	4.84	2.55	-8.67	-1.13	0.21

Algorithm	Policy	PA										
		Mean	SH600519	SH601318	SH600036	SH600276	SH601166	SH600030	SH600887	SH601328	SH600016	SH601288
DDQN	Single	0.57	-2.82	1.31	3.61	-0.77	-0.36	-0.34	-2.19	5.41	1.60	0.28
PPO	Single	4.7916	12.12	9.474	-1.31	1.4	1.70	4.97	4.016	7.213	3.079	5.257
ILO	Universal	4.12	5.65	8.39	4.85	1.82	0.27	9.74	12.66	-1.79	-0.80	0.37
	Fine-tuned	6.63	12.89	6.42	13.67	11.03	-4.35	11.95	9.31	-2.66	3.31	4.78

Algorithm	Policy	GLR										
		Mean	SH600519	SH601318	SH600036	SH600276	SH601166	SH600030	SH600887	SH601328	SH600016	SH601288
DDQN	Single	1.02	0.66	0.84	1.07	0.90	0.93	0.91	0.53	2.10	1.10	1.04
PPO	Single	1.10	1.39	0.91	0.78	0.84	0.56	0.91	1.33	1.66	1.34	1.08
ILO	Universal	1.29	0.87	1.19	1.31	1.79	1.07	1.49	2.12	0.73	0.99	1.37
	Fine-tuned	1.11	1.18	0.89	1.82	0.93	0.76	0.88	1.18	0.64	1.39	1.49

Table 3: The test results over 10 selected instruments.

Conclusion and future work



Thank You

