制約を課した深層強化学習による最適注文執行手法の検討

A Study on Optimal Order Execution Method Using Deep Reinforcement Learning with Constraints

保住 純[1*]　和泉 潔[1]
Jun Hozumi[1]　Kiyoshi Izumi[1]

[1] 東京大学大学院工学系研究科 システム創成学専攻
[1] Department of Systems Innovation, School of Engineering, the University of Tokyo

Abstract: For traders, it is important to minimize execution costs and achieve more efficient order execution. Since the mechanisms for incurring costs are unclear, being able to properly account for them will lead to lower execution costs and higher revenues. In order to achieve order execution with minimal costs, methods that model and infer market principles have been used. In recent years, model-free offline reinforcement learning methods have widely been utilized. However, the data on financial instruments contains a lot of noise, which makes learning hard and makes it difficult to converge to the optimal trading method. In this paper, we propose an optimal order execution method that improves performance by imposing constraints on the model. Through experiments, we have found that by imposing appropriate constraints, we can improve the performance of the optimal order execution method. We show that by setting appropriate constraints, we can achieve improved order execution compared to conventional methods.

1　はじめに

金融商品の取引を行うすべてのトレーダーにとって，執行コストを抑え，より効率的な注文執行を実現することは重要である．執行コストには手数料のような明示的なものがあれば，マーケット・インパクトやタイミングコストのような非明示的なものもあり，特に非明示的なコストの発生メカニズムは不明確であるため，これらを適切に考慮できるようになることが執行コストの低減および最終的な収益の増加につながる．これらのコストを抑えた注文を実現するために，従来より執行コスト等を踏まえた市場の原理をモデル化して推測する手法が用いられてきたが，近年では深層強化学習技術が発展したことにより，意思決定を深層学習によって実施するモデルフリーの強化学習の手法が用いられる事例が増えつつある．

一方で，金融商品に関するデータはノイズが多く含まれることから一般的な強化学習アルゴリズムを適用するだけでは学習に困難を伴い，最適な取引手法へ収束しづらいことから，従来手法に比べて適切な執行ができなくなる可能性があるという問題を抱えている．このため，このようなデータに適した強化学習の手法を開発することを検討する余地があるといえる．

そこで，本論文ではより性能を向上させた注文執行手法を実現するために，これまでの強化学習手法の学習時に新たに制約等を課すことでより性能を向上させる手法を提案する．改良のベースとする強化学習手法には，オラクル方策蒸留 (Oracle Policy Distillation; OPD)[1] を採用する．改良した手法と従来手法との比較実験によって，学習時に適宜制約等を設けることで，より改善された注文執行が実現する可能性があることを示す．

2　提案方法

本研究ではオラクル方策蒸留 [1] を用いて学習したエージェントをベースとし，さらにその学習時に制約等を付与した手法を用いることで性能を改善することを目的とする．本章ではオラクル方策蒸留や，それに追加する制約の要点を説明する．各手法の詳細については，提案されている各論文を参照されたい．なお，以降では一般的な強化学習の記法に従い，計 T ステップからなるタイムステップを t，行動は a，状態を s，報酬を R，方策を π，期待値を E で表す．

*連絡先：東京大学大学院工学系研究科システム創成学専攻
〒 113-8656 東京都文京区本郷 7-3-1 工学部 8 号館 5 階 531 号室
E-mail: hozumi@socsim.org

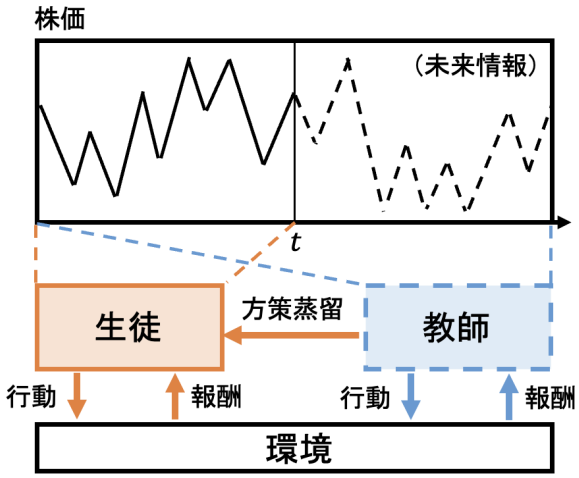

図 1: オラクル方策蒸留の概念図.

2.1 オラクル方策蒸留

オラクル方策蒸留とは，[1] で提案されている執行コストを抑えた取引を行う深層強化学習エージェントを学習するための手法である．まず，教師エージェントと生徒エージェントの 2 つを用意し，教師エージェント側を学習用データ全体を参照し，収束するまで学習する．教師エージェントは学習時に未来のデータを参照できるため，オラクル（預言者）とも呼ばれる．その後，生徒エージェントを，実際の条件に即して逐次的に現在の価格情報を参照しながら収束するまで学習する．教師エージェントと生徒エージェントの形状や学習方法は同じであるが，生徒エージェントの学習時に用いる誤差関数にのみ以下の L_d で表される教師エージェントの方策との差異をペナルティとして新たに加える．

$$L_d = -E_t[\log Pr(a_t = \tilde{a}_t|\pi_\theta, s_t)] \tag{1}$$

ただし，a_t は生徒モデルの方策 $\pi_\theta(\cdot|s_t)$ から導かれた行動（π は生徒モデルに関連するパラメータ），$\tilde{a}_t$ は教師モデルの方策 $\pi_\phi(\cdot|\tilde{s}_t)$ から導かれた行動（ϕ は教師モデルに関連するパラメータ），Pr は確率，E_t は全タイムステップを通じた期待値を表す．教師エージェントは学習時に未来の情報を参照しており，そのまま用いるとデータ漏洩となるので，訓練時の L_d の計算時の利用にとどめ，生徒エージェントのみを最終的なテスト時に使用する．学習メカニズムの全体像を図 1 に記す．

2.2 スペクトル正規化

敵対的生成ネットワーク (Generative Adversarial Networks; GAN) の学習を安定化させるための手法として，スペクトル正規化 (Spectral Normalization; SN) が知られている [2]．そして近年，[3] にてそのスペクトル正規化を強化学習に適用することによっても学習性能を高めることが示された．具体的には，層 W に以下の式を通じて与えられる 1-リプシッツ制約を加え $W_{SN}(W)$ とする手法である．

$$W_{SN}(W) = \frac{W}{\sigma(W)} \tag{2}$$

ただし，$\sigma(W)$ は W の最大特異値である．このような制約はエージェント内の各層に課すことができるが，[3] の報告では各層に複数課すなど，制約のかけ具合によっては逆に学習性能が悪くなることが記されている．そのため，本研究では [3] での報告に従い，出力層の一つ前の層のみに課すことにした．

2.3 方策エントロピー

エージェントの学習時により探索を多様化させるために，[4] の Soft Actor-Critic(SAC) にあるような方策エントロピー項を導入する．具体的には，以下の式で表されるような学習時に最大化する目的関数に方策エントロピー項を加える．

$$E_\pi[\sum_{t=0}^{T-1} \gamma^t \left(R_t(s_t, a_t) + \beta H(\cdot|s_t)\right)] \tag{3}$$

ただし $H(\cdot|s_t)$ はエントロピー関数，γ は割引率，β は温度パラメータである．方策エントロピーを報酬和とのトレードオフをとりつつ最大化する目的関数に加えているため，SAC は方策エントロピーの正則化を指向していると捉えることができる．

3 実験

3.1 問題設定

本研究の問題設定は従来手法との比較のために，[1] 内の実験での問題設定に従っている．このため，本研究では株式 $Q = 1$ 単位を規定期間内に売却し，最大の収益を得ることを目的として設定した．各タイムステップごとの行動 a_t は株式の売却量とし，0〜1 の範囲の値をとるものとする．これより $\Sigma a_t = 1$ であり，$a = 0$ はそのタイムステップでは何も売却しない (Hold) ことを意味する．環境 s_t として Apple(AAPL) の分足株価データの Close の値を用い，過去の情報も参照するために株価の現在値に加え，その 2 分平均，5 分平均，10 分平均，および 20 分平均を作成した．Train 期間を 2020 年 12 月 1 日〜10 日，Validation 期間を 12 月 11 日〜15 日，Test 期間を 12 月 16 日〜20 日に設定している．

報酬 $R_t(s_t, a_t)$ は，以下の式で表される．

$$\hat{R}_t^+(s_t, a_t) = \left(\frac{p_{t+1}}{\tilde{p}} - 1\right) a_t \tag{4}$$

$$\hat{R}_t^-(s_t, a_t) = -\alpha(a_t)^2 \tag{5}$$

$$R_t(s_t, a_t) = \hat{R}_t^+(s_t, a_t) + \hat{R}_t^-(s_t, a_t) \tag{6}$$

$\hat{R}_t^+$ はそのタイミングでの価格 p_{t+1} に連動した収益を意味し，全体を通じた平均価格 $\tilde{p}$ で割ることによって正規化された値である．$\hat{R}_t^-$は 多くの取引量を執行するほどマーケットインパクトなどの執行コストの影響を受けてしまうペナルティを意味し，本研究でも [1] と同じく $\alpha = 100$ に設定している．最終的には，それぞれのエージェントは以下の式で表される累積報酬和の最大化問題を解くことで学習される．取引終了時の累積収益が高くなることを目的としているため，一般的な強化学習で考慮される報酬の割引率を考慮しなくてよいことから，本研究では $\gamma = 1$ と設定される．

$$\arg\max E_\pi \left[\sum_{t=0}^{T-1} \gamma^t R_t\left(s_t, a_t\right)\right] \tag{7}$$

比較対象としては深層強化学習を用いた代表的なトレード手法である Double-DQN(DDQN) や Proximal Policy Optimization(PPO) を設定し，それらと本論文で提案しているオラクル方策蒸留 (OPD) およびそれらに各制約を加えたエージェントで学習を行った．各手法の評価指標として，[1] と同じく報酬の総和 (Reward)，Price Advantage(PA)，GLR(Gain-Loss Ratio) を採用し，性能を比較することとした．PA 及び GLR は以下の式で表される．ただし，$\bar{P}^k_{strategy}$ は a により重みづけされた平均執行価格，$\tilde{p}^k$ は平均市場価格，$|D|$ はデータセットのサイズである．

$$PA = \frac{10^4}{|D|} \sum_{k=1}^{|D|} \left(\frac{\bar{P}^k_{strategy}}{\tilde{p}^k} - 1\right) \tag{8}$$

$$GLR = \frac{E[PA|PA > 0]}{E[PA|PA < 0]} \tag{9}$$

3.2 実験結果

結果を表 1 に記す．実験の結果，OPD にスペクトル正規化を適用したエージェントが最も性能が高いという結果を得られた．一方，方策エントロピーを導入したエージェントでも，それを適用しないエージェントより性能が改善された．

また，DDQN や PPO に比べ，OPD やそれらに制約を加えたエージェントでは，価格が高いタイミングで売却を行っていることが確認されている．もともとの OPD と OPD に制約を加えたモデル間では，行動の挙動に大きな差がなかったため，制約の付与によってエージェントの学習が大きく変更されずに性能が改善されたことが推察される．

表 1: 実験結果一覧（最も良い結果を太字で表示）．

手法	Reward($*10^{-2}$)	PA	GLR
DDQN	2.61	4.11	0.96
PPO	1.17	2.44	0.66
OPD	3.22	5.19	1.18
OPD+SN	**3.38**	**6.14**	**1.35**
OPD+SAC	2.93	5.49	1.24

4 むすび

本論文ではエージェントの学習時に制約などを課すことでより性能を向上させた最適注文執行手法を提案した．オラクル方策蒸留という既存の深層強化学習トレード手法をもとに，新たにスペクトル正規化と方策エントロピーという 2 種類の制約を加えたエージェントを設計した．そして，検証実験を通じて，このような制約を設けることによって従来の手法よりも改善された注文執行が可能なエージェントが実現される可能性があることが示された．

今回は使用できたデータや時間の都合，分足レベルのデータを用いて実験を行ったが，執行コストの影響をより正確に捉えられるようにするためにも，今後は秒足やティックレベルといった，より粒度の小さいデータを用いて実験を進めたい．また，今回はまだ一種類の金融データのみでしか実験ができていないが，複数の銘柄や異なる金融商品によっても同様の実験を実施し，それらの時系列の特徴と実験結果を照らし合わせることを通じて，金融データのどのような性質に対し，どのような制約をかけることがより高収益を得られるエージェントの学習に効果的であるのかを検証していきたいと考えている．

他にも，今回の手法においては，取引量に応じた負の報酬を設定するなど，執行コストをモデル化した上で実験を行っている部分が残されている．このような執行コストの影響を一部であってもモデル化して反映させていることは，従来の研究より残され続けている課題である．この課題を克服するために，十分な取引履歴データがある際にはオフライン強化学習によるエージェントの学習を検討するなど，執行コストをモデル化しない，より厳密にモデルフリーな取引エージェントの設計を検討したい．

謝辞

本研究は大和証券株式会社および株式会社大和総研との共同研究の一部であり，大和証券グループの支援を受けて行われました．大和証券グループ各社の多くの方にご支援をいただきました．この場を借りて厚く御礼申し上げます.

参考文献

[1] Fang, Y., et al.: Universal Trading for Order Execution with Oracle Policy Distillation, *In Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 107–115 (2021)

[2] Miyato, T., et al.: Spectral Normalization for Generative Adversarial Networks., *In Proceedings of the International Conference on Learning Representations*, (2018)

[3] Gogianu, F., et al.: Spectral Normalisation for Deep Reinforcement Learning: An Optimisation Perspective, *In Proceedings of the 38th International Conference on Machine Learning*, pp. 3734–3744 (2021)

[4] Haarnoja, T., et al.: Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor, *In Proceedings of the 35th International Conference on Machine Learning*, pp. 1861–1870 (2018)