Soft Computing (2021) 25:7877–7885
https://doi.org/10.1007/s00500-021-05801-6

FOCUS

Multi-agent reinforcement learning approach for hedging portfolio problem

Uyen Pham[1] · Quoc Luu[2] · Hien Tran[3]

Accepted: 5 April 2021 / Published online: 19 April 2021

Abstract
Developing a hedging strategy to reduce risk of losses for a given set of stocks in a portfolio is a difficult task due to cost of the hedge. In Vietnam stock market, cross-hedge is involved hedging a long position of a stock because there is no put option for the stock. In addition, only VN30 stock index futures contracts are traded on Hanoi Stock Exchange. Inspired by recently achievement of deep reinforcement learning, we explore feasibility to construct a hedging strategy automatically by leveraging cooperative multi-agent in reinforcement learning techniques without advanced domain knowledge. In this work, we use 10 popular stocks on Ho Chi Minh Stock Exchange, and VN30F1M (VN30 Index Futures contracts within one month settlement) to develop a stock market simulator (including transaction fee, tax, and settlement date of transactions) for reinforcement learning agent training. We use daily return as input data for training process. Results suggest that the agent can learn trading and hedging policy to make profit and reduce losses. Furthermore, we also find that our agent can protect portfolios and make positive profit in case market collapses systematically. In practice, this work can help Vietnam's stock market investors to improve performance and reduce losses in trading, especially when the volatility cannot be controlled.

Keywords Deep reinforcement learning · Hedging · Trading · Portfolio

1 Introduction

Hedging a position in stock is an attractive topic for both academics and practitioners. The objective of hedging is to minimize market risk due to price fluctuation, maximize profit by speculation on the basis, and construct a portfolio with reduced risk Floros and Vougas (2004). Portfolio managers have used stock index futures as a means to adjust desired return of a portfolio and potential loss since the 1980s. Main advantage of index futures as a major hedging tool is liquidity and lower transaction costs Ghosh (1993). However, hedge strategies are not always effective as expected because relationship of cash price and price of a future contract is usually not perfect, or hedged position in stock is different from the underlying portfolio for the index contract Figlewski (1984), Floros and Vougas (2004). It is possible to increase risk of potential loss that leads to negative return. Hence, hedgers have to determine the optimal hedge ratio to control the risk of the portfolio.

In contrast to supervised learning and unsupervised learning, reinforcement learning mainly relies on experience of repeated interaction to learn optimal policy in order to make sequential decisions to maximize rewards in a given environment Sutton and Barto (2018). In a complex and dynamic environment, it may require huge amounts of computational power over a long period of time to train. With revolution of deep learning techniques and computer hardware, reinforcement learning has become more feasible by using deep neural network as a functional estimator. From long time horizons with high-dimensional observation and action spaces in real-time strategy games, to self-driving vehicles, data center cooling systems, deep reinforcement learning has been more

Communicated by Vladik Kreinovich.

✉ Uyen Pham
uyenph@uel.edu.vn

Quoc Luu
quoc.luu2015@qcf.jvn.edu.vn

Hien Tran
hien.tran@ttu.edu.vn

[1] Economic Mathematics, University of Economics and Law, Ho Chi Minh City, Vietnam

[2] Quantitative and Computational Finance, John von Neumann Institute, Ho Chi Minh City, Vietnam

[3] School of Engineering, Tan Tao University, Long An, Vietnam

and more applied to solve many complex real-world challenges Berner et al. (2019), Evans and Gao (2016), O'Kelly et al. (2018). In finance, deep reinforcement learning is also widely adopted. Zhang et al. Zhang et al. (2020) uses various RL algorithms including deep Q-learning, policy gradients, and Advantage Actor–Critic (A2C) to design trading strategies for continuous futures contracts. They use technical indicators such as moving average convergence/divergence (MACD) and relative strength index (RSI) as a part of input features. The agent shows that it can deliver profits even under heavy transaction costs. Ganesh et al. Ganesh et al. (2019) develop a multi-agent dealer market for market marking with different competitive scenarios and market price conditions. The research suggests that trained agent can learn to manage inventory and its competitor's policy for pricing.

However, investigating feasibility of stock and futures trading to hedge portfolio at the same time using deep reinforcement learning is still a topical and interesting problem. In this study, we selected 10 popular stocks on HSX, and one stock index futures contract on HNX to build a simulation of stock market environment with real market data to study learning performance of our agent. Our objective in this work is to investigate whether cooperation of multi-agent to determine optimal hedging strategy to protect stock portfolio is achievable.

2 Related work

2.1 Some basic background about hedging

Hedging is a finance strategy to reduce risk in investments by taking an opposite position in a related asset to offset losses. Basically, before making any investments, investors have to balance between profit and risk, for example, expected returns and variance of returns. It is the fact that a dollar of loss can cost the investor or the company more than a dollar of high profit. Hence, the reduction in risk provided by hedging also typically results in reduction in potential profits. The trade-off between profits and risks is the basic problem in finance.

Hedging generally involves the use of financial instruments known as derivatives. The two most common derivatives are options (such as call option, namely the right to buy an asset at the fixed strike price by the predetermined time t in the future, or the put option, i.e., the right to sell) and futures or future contracts. (The buyer must purchase or the seller must sell the underlying asset at the set price at the expiration date.)

How much should an open or spot position be hedged? Fixed or "obvious" hedge ration may increase rather than decrease risk (McDonald (2006)). It depends on which kinds of risk investors consider, and then, the optimal hedge ratio would be obtained accordingly. For example, if the variance of returns is used as a risk measure of a portfolio, the optimal hedge ration would be the minimum variance hedge ratio (MVHR).

2.2 Hedging effectiveness of stock index futures

Motivated by risk reduction, hedging a stock portfolio with index futures has been an active research topic since it was introduced Figlewski (1984). A hedger supposes that return of a hedged position (e.g., stock portfolio) can be closed to risk-free interest rate. In terms of optimal hedge ratio hr, there are many methods used to estimate the ratio. For instance, one-to-one hedge, the beta hedge, and the MVHR are some of these methods Brown (1985), Ederington (1979). Butterworth et. al. Butterworth and Holmes (2001) evaluated hedging effectiveness of stock index futures with four different strategies (i.e., the traditional hedge, MVHR, least trimmed squares (LTS), and beta ratio of cross-hedge) with two daily and weekly hedge durations in the UK market. The results suggest that MVHR and LTS methods are robust to estimate the ratio. With cash prices and futures moving closely together assumption, one-to-one hedge strategy suggests $hr = -1$. Beta hedge strategy uses negative of the beta cash portfolio as hr. The hedger expects the overall beta of the portfolio is zero. However, in practice, change of prices of spot and futures is imperfectly correlated. Particularly in case of cross-hedge (namely the use of a derivative on one asset to hedge another asset), one-to-one and beta hedge may not reduce risk. In contrast, futures hedging can lead to unexpected loss.

The MVHR was introduced to work around for the problem by taking the imperfect relationship of prices into account and determine the optimal ratio hr. Let R_s, R_f, and R_h are returns of spot position (e.g., open portfolio), futures positions (e.g., index futures for hedging), and the hedged portfolio with futures, respectively, then we get

$$R_h = R_s + hR_f \tag{1}$$

$$Var(R_h) = Var(R_s) + h^2 Var(R_f) + 2hCov(R_s, R_f) \tag{2}$$

The optimal ratio h (or hr) to minimize the $Var(R_h)$ is:

$$hr = -\frac{Cov(R_s, R_f)}{Var(R_f)} \tag{3}$$

Furthermore, by using ordinary least squares (OLS) regression to estimate minimum risk hedge, Figlewski Figlewski (1984) found that hedging effectiveness of a large capitalization portfolio can yield "fairly good" for a one-week holding period (p. 663). However, with diversified portfolio of small stocks, the effectiveness is reduced significantly. Basis risk

Fig. 1 Daily movement of VN30 Index from December 19, 2019, to April 16, 2020

is also not negligible even if the spot is hedged with index futures itself. When basis risk arises, it can generate profit or loss. It is suggested that one-day holding hedge positions strategy can potentially increase basis risk and reduce risk effectiveness than one-week hedge.

Stating that traditional methods to estimate optimal hedge ratio are misspecified, error correction model (ECM) was proposed to estimate optimal hedge ratio and forecast out of sample for evaluation as in Ghosh (1993). Firstly, it carries out cointegration test. Secondly, it use OLS regression to estimate error correction model. The model incorporates relationship of the long-run equilibrium as well as the short-run dynamics. The result shows that optimal hedge ratio is significantly improved with adjusted R^2 from ECM which is higher than traditional methods. Also, by comparing root-mean-squared error (RMSE), out-of-sample forecasts from the ECM are found to be better than other methods.

Beyond variance and standard deviation, value at risk (VaR) and conditional value at risk (CVaR) are extensively applied to measure market risk for hedging strategies of portfolio Cao et al. (2010), Huggenberger et al. (2016). VaR was introduced by J.P. Morgan in the 1990s and widely adopted to summarize risk of an entire portfolio at the end of each day Miller (2018). However, VaR is not a coherent risk measure. To be coherent, it must be monotonicity, positive homogeneity, translation invariance, and subadditivity Artzner (1999), Artzner et al. (1997). CVaR was constructed with these properties as a new valid practical alternative to VaR Acerbi and Tasche (2002). Espeholt et al. Alexander et al. (2003) show that CVaR is applicable to a wide range of derivatives portfolio including American options and exotic options. In addition, it is found that CVaR risk metric is suitable for asymmetric return distributions and expected loss of portfolio can be minimized in many circumstances Topaloglou et al. (2002).

2.3 Deep Reinforcement Learning in Trading

Reinforcement learning was proposed to train trading systems to make profit and to adjust risk Moody et al. (1998), Moody et al. (1998). Recurrent learning and Q-learning with neural networks were used to optimize financial performance functions including risk-adjusted return and immediate utility for online learning Moody et al. (1998). Furthermore, portfolios with continuous quantities of multiple assets were considered. The result shows that reinforcement learning can avoid large losses when market crashed. Basis risk hedging strategy was developed using reinforcement learning as in Watts (2015). Without assets modeling requirements, state–action–reward–state–action (SARSA)-based algorithm was applied to find an optimal trading policy to hedge a non-traded asset. Q-learning is proposed to extend Black–Scholes–Merton (BSM) model for option pricing and hedging in Halperin (2017). In an attempt to escape Greeks and complete market assumptions in risk management, by leveraging deep reinforcement learning, a Greek-free approach is proposed to focus on realistic market dynamics and out-sample testing performance for optimiz-

ing hedging of a portfolio of derivatives Buehler et al. (2019). Deep reinforcement learning is further investigated for hedging a portfolio of over-the-counter derivatives under generic market frictions as in Buehler et al. (2019). Trading costs and liquidity constraints are considered in the approach.

3 Multi-Agent Reinforcement Learning Approach

3.1 Deep Reinforcement Learning

3.1.1 Single-Agent Reinforcement Learning

For a given stochastic environment ε, an agent interacts with the environment by choosing to take a legal action a_t from many actions at time step t, $a_t \in A \equiv \{1, \ldots, L\}$. Action space can be discrete or continuous. When the selected action is passed to the environment ε, internal state s_t is switched to another state in many states S. In other words, the process of sequential interactions between the agent and the environment is result of mapping from perceived states s_t to actions a_t by policy π. For instance, in Dota 2 game, internal state can be all the available information for human player including positions, health, map Berner et al. (2019). In this research, internal state is asset return in percentage, position of each asset. In return, the agent receives reward r_t of the passed action as feedback, and new internal state s_{t+1} for each time step until reaching terminate state. The ultimate goal of deep reinforcement learning is to find an policy π that can select optimal action to maximize reward signal for each state s_t. Value of a state measures total expected return by predicting future reward with discount rate $\gamma \in [0, 1]$. The total accumulated discounted return G_t from time t with k time steps in the future is defined as:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{4}$$

The state value $V_\pi(s)$ is defined as in Sutton and Barto (2018):

$$\begin{aligned} V_\pi(s) \doteq & \mathbb{E}_\pi[G_t | S_t = s] \\ = & \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s \right] \end{aligned} \tag{5}$$

Similarly, action value $Q_\pi(s, a)$ is the expected return for state s from selecting action a following policy π.

$$\begin{aligned} Q_\pi(s, a) = & \mathbb{E}_\pi[G_t | S_t = s, A_t = a] \\ = & \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a \right] \end{aligned} \tag{6}$$

In value-based reinforcement learning, off-policy Q-learning was introduced to estimate the action value function $Q_\pi(s, a)$, defined as Watkins (1989).

$$\begin{aligned} Q(S_t, A_t) \leftarrow & Q(S_t, A_t) \\ & + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right] \end{aligned} \tag{7}$$

The algorithm directly approximates the optimal action value function $Q_*(s, a)$. By extending neural network as a function approximator, the function can be estimated as $Q_*(s, a) \approx Q(s, a, \theta)$. The approach is referred as Q-network with weights θ Mnih et al. (2013).

In contrast to value-based methods, policy-based can select actions directly by parameterizing the policy $\pi(a|s, \theta)$ and using gradient ascent to optimize $\mathbb{E}[R_t]$ to find the best θ that can produce the highest reward. In terms of probability, we can express the policy as $\pi(a|s, \theta) = Pr\{A_t = a | S_t = s, \theta_t = \theta\}$ for the probability of a given environment ε in state s at time t with parameter θ to take action a. Actor–critic algorithms use both value and policy functions to learn approximations Konda and Tsitsiklis (2000). To improve performance, the critic learns a value function (e.g., state value) and is used to update policy parameters of actor.

3.1.2 Multi-Agent Reinforcement Learning

Extent from single agent, multi-agent learning is considered n agents interacting with the environment ε. At state s_t of time step t, each agent selects action a_t^i to react to the state and receive reward r_t^i, where $i \in \{1, \ldots n\}$. Hence, for any given joint policy $\pi(a|s) \doteq \prod_i^n \pi^i(a^i|s)$ with state $s \in S$, state value function can be defined as in Zhang et al. (2019):

$$V^i_{\pi^i, \pi^{-i}}(s) \doteq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^i_{t+1} \middle| a^i_t \sim \pi^i(\cdot|s_t), s_0 = s \right] \tag{8}$$

where $-i$ indicates that all n agents except the ith agent.

3.2 High Throughput Architecture with Importance Weighted Actor–Learner Architecture

Importance weighted actor–learner architecture (IMPALA) is a decoupled actor–critic style learner with introduction of V-trace off-policy to learn a policy π and a baseline function V^π that achieves stability, high data throughput, and efficiency for agent training Espeholt et al. (2018). Moreover, deep neural networks can be trained efficiently with IMPALA as suggested in Fig. 2. Suppose at time t, a given local actor policy μ generates trajectory $(s_t, a_t, r_t)_{t=k}^{t=k+n}$. The n-step V-trace target for value approximation $V(s_k)$ at state s_k is defined as:

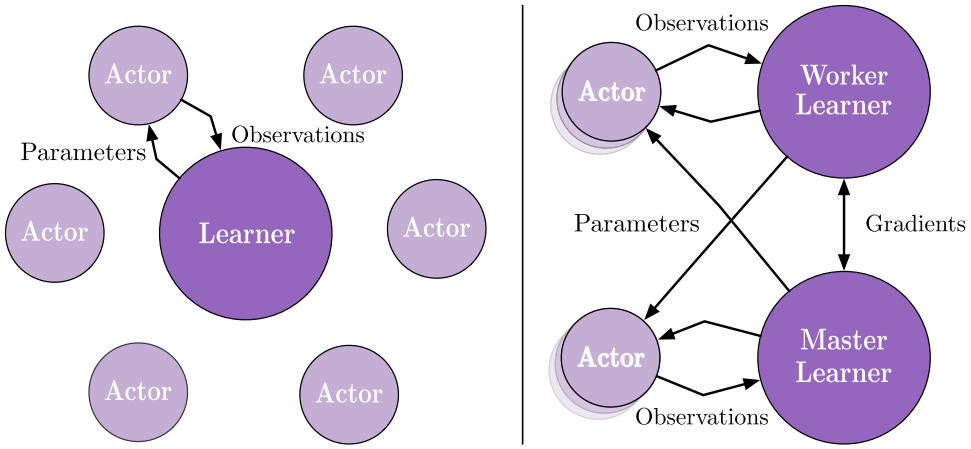

Fig. 2 Left: Single learner. Right: multiple synchronous learners. Adopted from Espeholt et al. (2018)

$$
\begin{aligned}
v_k &\doteq V(s_k) + \sum_{t=k}^{k+n-1} \gamma^{t-k} \Big(\prod_{i=k}^{t-1} c_i \Big) \delta_t V \\
\delta_t V &\doteq \rho_t (r_t + \gamma V(s_{t+1}) - V(s_t)) \\
\rho_t &\doteq \min \left(\bar{\rho}, \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} \right) \\
c_i &\doteq \min \left(\bar{c}, \frac{\pi(a_i|s_i)}{\mu(a_i|s_i)} \right)
\end{aligned}
\tag{9}
$$

where $\delta_t V$ is temporal difference for V, and ρ_t and c_i are truncated importance sampling. It is worth to note that the truncation levels are assumed $\bar{c} \leq \bar{\rho}$.

Furthermore, value function V_θ and policy π_ω with θ and ω parameters, respectively, can be updated in the direction of:

$$
\begin{aligned}
\Delta\theta &= (v_k - V_\theta(s_k)) \nabla_\theta V_\theta(s_k) \\
\Delta\omega &= \rho_k \nabla_\omega log \pi_\omega(a_k|s_k) \\
&(r_k + \gamma v_{k+1} - V_\theta(s_k)) - H(\omega) \\
H(\omega) &= \nabla_\omega \sum_a \pi_\omega(a|s_k) log \pi_\omega(a|s_k).
\end{aligned}
\tag{10}
$$

Entropy $H(\omega)$ is added to avoid immature convergence and encourage exploration in agent training process. IMPALA algorithm can be used to concurrently train for multiple tasks with one set of weights due to efficiency of the architecture.

Table 1 Evaluation Periods

	First trading date	Last trading date
1	2019-05-17	2019-06-20
2	2019-06-21	2019-07-18
3	2019-07-19	2019-08-15
4	2019-08-16	2019-09-19
5	2019-09-20	2019-10-17
6	2019-10-18	2019-11-21
7	2019-11-22	2019-12-19
8	2019-12-20	2020-01-16
9	2020-01-17	2020-02-20
10	2020-02-21	2020-03-19
11	2020-03-20	2020-04-16
12	2020-04-17	2020-05-21

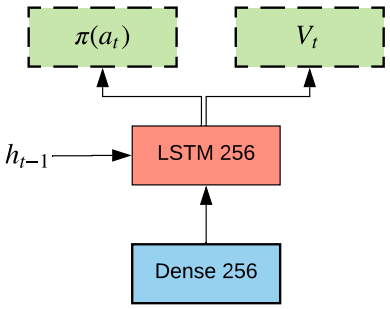

Fig. 3 Model architecture for policy and value networks

4 Experiments

4.1 Data

We collect daily historical stock prices and volumes data from Ho Chi Minh Stock Exchange (HSX) for equity and Ha Noi Stock Exchange (HNX) for derivatives. We use data of the stock markets from September 25, 2017, to May 21, 2020, for our training and evaluating purposes. The time range covers high fluctuation of price periods as impact of market events.

4.2 Data prepossessing and network architecture

Training data from raw inputs rather than handcrafted features are often recommended for feeding data to deep neural networks to achieve higher performance Krizhevsky et al. (2012). Likewise, researches show that reinforcement learn-

Table 2 Market returns of selected stocks in percentage

	FPT	GAS	SSI	MSN	NVL	VN30F1M
Period 1	− 0.15%	− 7.41%	− 3.49%	− 6.04%	− 1.88%	− 4.14%
Period 2	4.79%	5.01%	2.82%	− 4.36%	5.4%	2.22%
Period 3	12.54%	− 9.09%	− 15.86%	− 4.82%	− 0.83%	− 0.8%
Period 4	9.06%	6.43%	1.16%	7.72%	5.83%	4.95%
Period 5	1.42%	− 1.65%	3.51%	− 3.83%	− 2.36%	1.32%
Period 6	− 1.05%	2.67%	− 3.02%	− 6.17%	− 5.81%	− 0.65%
Period 7	− 2.29%	− 7.69%	− 11.96%	− 23.29%	− 5.65%	− 6.1%
Period 8	2.7%	− 2.08%	0%	0.89%	− 0.18%	2.78%
Period 9	− 1.58%	− 6.38%	− 3.81%	− 9.91%	− 1.45%	− 2.14%
Period 10	− 15.15%	− 36.36%	− 24.3%	− 4.72%	− 5.9%	− 20.99%
Period 11	4.41%	18.93%	14%	22.89%	1.37%	4.53%
Period 12	15.23%	11.86%	9.47%	7.05%	2.9%	12.85%

Table 3 Performance of proposed RL trading system in percentage

	FPT	GAS	SSI	MSN	NVL	VN30F1M	Portfolio
Period 1	2.5%	− 3.48%	− 2.99%	− 5.33%	− 2.67%	6.54%	2.07%
Period 2	1.06%	3.02%	1.35%	− 3.76%	2.89%	1.65%	1.28%
Period 3	11.47%	− 2.01%	− 11.91%	− 1.11%	− 1.61%	1.9%	0.43%
Period 4	5.11%	0.06%	2.44%	6.84%	5.32%	− 1.35%	1.3%
Period 5	− 1.66%	− 5.18%	− 0.16%	0.29%	− 2.85%	4.38%	1.23%
Period 6	− 0.92%	2.66%	− 0.2%	0.6%	− 4.04%	1.63%	0.06
Period 7	− 0.72%	− 0.9%	− 6.72%	− 10.38%	− 0.95%	6.33%	1.2%
Period 8	0.85%	− 2.03%	− 0.65%	6.99%	5.32%	0.74%	1.42%
Period 9	2.16%	1.79%	− 1.98%	− 9.91%	− 0.75%	2.93%	0.6%
Period 10	− 12.37%	− 21.19%	− 2.83%	− 0.34%	− 3.99%	19.17%	5.53%
Period 11	11.03%	11.19%	6.45%	17.61%	− 0.84%	− 0.74%	4.17%
Period 12	4.35%	3.82%	11.83%	2.63%	0.36%	11.2%	7.89%

ing can exceed human capabilities without human expert data or domain knowledge Mnih et al. (2013), Silver et al. (2017). As a result, in this study, instead of applying advanced quantitative finance theories to develop trading and hedging strategy, daily return data of each asset collected from HSX and HNX exchanges are used as main components of environment observation. Specifically, we conducted a set of 12 different periods for out-sample evaluation. We use data from September 25, 2017, to May 16, 2019, for training, and from May 17, 2019, to May 21, 2020, for evaluation (see Table. 1). We also provide position and unrealized profit of each asset (P&L) in portfolio to our neural network.

Actor–critic-based algorithms use policy and value networks. We can use policy network and value network separately or combine these two networks. We choose the combined architecture due to improvement of computational efficiency. Furthermore, we use LSTM Hochreiter and Schmidhuber (1997) beside dense layer in the shared network.

As suggested in Fig. 3, we use a shallow network architecture for this study. In detail, in terms of shared network, a trajectory is feed to the first fully connected hidden layer with 256 units and applies *tanh* activation function. The next hidden layer is stateful LSTM with 256 unit. The tanh function is also applied to the LSTM layer. Finally, policy and value heads are fully connected linear layer for single output of each action and state value, respectively.

4.3 Result

We use same network architecture and hyper-parameters all trading agent without tweaking. The agent learns to decide to long buy or short sell assets on its own. For instance, agent can cut loss or hold positions overnight without any constraint.

Our experiment uses discrete actions. For equity, trading agent can hold, buy, or sell stocks without considering amount of volume. Stocks are only sold after T+2 settlement. Likewise, derivatives trading agent can hold, long, or short futures contracts. However, the agent can trade continuously

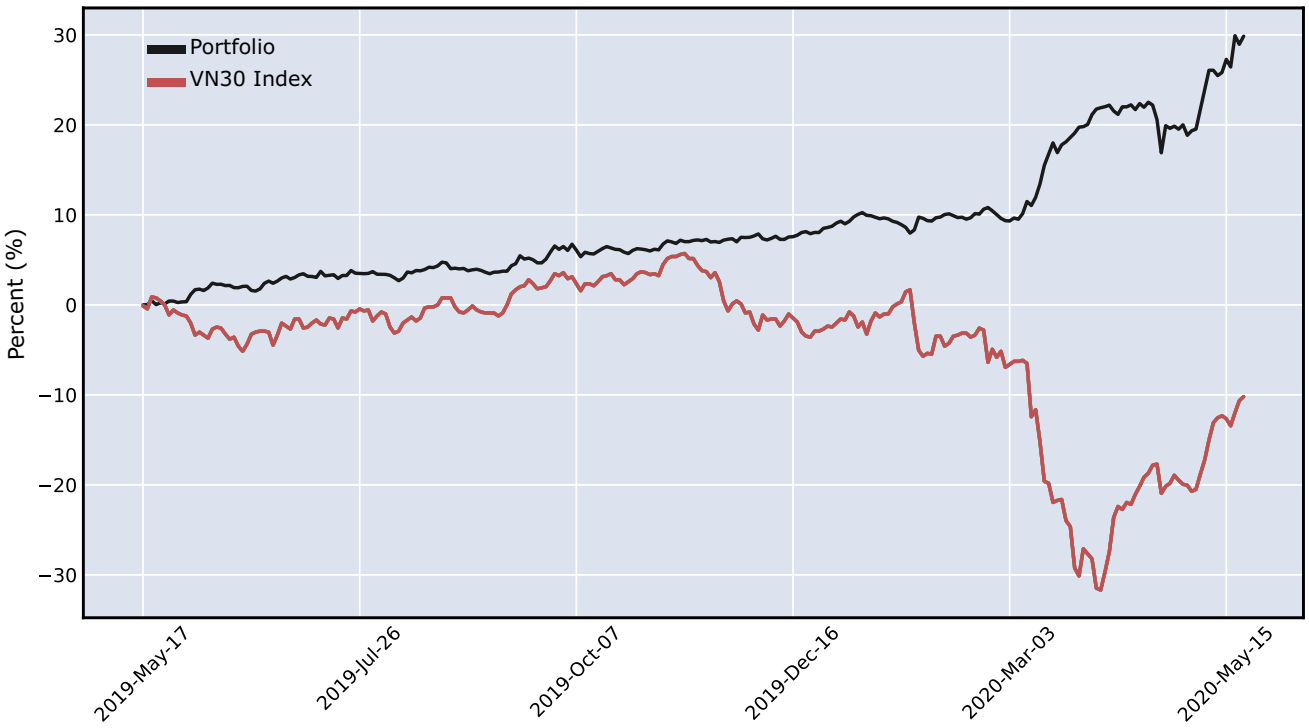

Fig. 4 Cumulative return of portfolio and VN30 Index in percentage from May 17, 2019, to May 21, 2020

Input : Selected valid action $\pi(a)$ of learned agent π, Transaction fee F, Price of asset m, Asset p in portfolio P, Total time step T
Output: Portfolio value PV
for *t=1 to t=T* **do**
  for *p in P* **do**
    if $\pi(a_t) \leftarrow 0$ **then**
      Hold position;
    else
      if $\pi(a_t) \leftarrow 1$ **then**
        Trade return $r_p \leftarrow$ Long buy at m_t;
      else
        Trade return $r_p \leftarrow$ Short sell at m_t;
      end
      r_p -= F;
      PV += r_p;
    end
  end
end

Algorithm 1: Simulation of stock market environment

as it is T+0 settlement market. We include transaction fees for every trade return (see Algorithm. 1). For every agent, reward can be 1 if overall profit of an episode is positive and -1 in case of negative profit. In addition, we discount reward for every long position of agent in future market to encourage hedging.

Finally, we use RLLib Liang et al. (2017) with 32 workers to train the agent. RMSProp algorithm is used as optimizer for training.

4.3.1 Buy and Hold Strategy Baseline

We compare our proposed trading strategy result with performance of buy and hold strategy to determine effectiveness of the approach. During the evaluation periods, the stock market is highly volatile due to impact of COVID-19 pandemic. Buy and hold strategy may lead to negative return (see Table. 2).

4.3.2 Multi-Agent Reinforcement Learning

Learned deep RL agent was deployed to trade out-of-sample market data from May 17, 2019, to May 21, 2020. The result shows that the learned agent can protect portfolio by short selling in futures market. In addition, in some cases, our agent can cut loss and achieve higher performance than market return in equity market even market plunged as traders had panic-sold out of COVID-19 pandemic fear (see Table. 3).

Specifically, VN30 Index was lost about 300 points (33%) during the first three months (Period 9, Period 10, Period 11) of 2020 (see Fig. 1). In terms of equity trading, every stock in portfolio had negative market return in the periods. In equity market, after transaction and commission fees, the RL agent cannot maintain positive return. In contrast, our agent executed many orders for opening and closing position to hedge equity assets dynamically in futures market. It leads to positive return of the portfolio (see Fig. 4). The portfolio profit did not decrease when the market rebounded due to

dynamic hedge strategy. However, in some cases, the agent cannot achieve higher performance than return of market.

As a result, we show that our method can reduce losses and achieve positive profit in trading. Furthermore, the trading data also suggest that dynamic hedging strategy for equity in portfolio is feasible in cross-hedging case. The futures trading agent generated far profit than losses. Overall, during evaluation periods, our deep RL agent earned about 30% profit of portfolio value and maintains positive return in case market collapsed systematically.

5 Conclusion

This study proposed a feasible approach for cross-hedging in trading without domain knowledge by applying deep reinforcement learning. Our result also suggests that the approach can cut loss efficiently when market is in selling panic as happening in COVID-19 event. Overall, the proposed method can generate positive profit with dynamic hedge strategy. The result is desirable as our approach earns higher performance than the risk-free rate Hancock and Weise (1994).

It is important to develop a deterministic behavior of agent to maintain reliable outcome. In future work, we should further study stability and safety in reinforcement learning for trading.

Acknowledgements Hien Duy Tran was supported by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) grant 101.02-2019.319.

Declarations

Conflict of Interest Uyen Pham declares that she has no conflict of interest. Quoc Luu declares that he has no conflict of interest. Hien Tran declares that he has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Acerbi C, Tasche D (2002) Expected shortfall: a natural coherent alternative to value at risk. Econ Notes 31(2):379–388

Alexander S, Coleman TF, Li Y (2003) Derivative portfolio hedging based on cvar. New Risk Measures in Investment and Regulation: Wiley

Artzner P, Delbaen F, Eber JM, Heath D (1997) Thinking coherently. risk, 10. November, 68, 71

Artzner P, Delbaen F, Eber JM, Heath D (1999) Coherent measures of risk. Math Finance 9(3):203–228

Berner C, Brockman G, Chan B, Cheung V, Debiak P, Dennison C, Farhi D, Fischer Q, Hashme S, Hesse C, Others (2019) Dota 2 with Large Scale Deep Reinforcement Learning. arXiv preprint. arXiv:1912.06680

Brown SL (1985) A reformulation of the portfolio model of hedging. Am J Agric Econ 67(3):508–512

Buehler H, Gonon L, Teichmann J, Wood B (2019) Deep hedging. Quant Finance 19(8):1271–1291

Buehler Hans, Gonon Lukas, Wood Ben, Teichmann Josef, Mohan Baranidharan, Kochems Jonathan (2019) Deep hedging: Hedging derivatives under generic market frictions using reinforcement learning-machine learning version. *Available at SSRN*

Butterworth D, Holmes P (2001) The hedging effectiveness of stock index futures: evidence for the FTSE-100 and FTSE-Mid250 indexes traded in the UK. Appl Financ Econ 11(1):57–68

Cao Z, Harris RDF, Shen J (2010) Hedging and value at risk: a semi-parametric approach. J Futures Markets Futures Opt Other Derivat Prod 30(8):780–794

Ederington LH (1979) The hedging performance of the new futures markets. J Finance 34(1):157–170

Espeholt L, Soyer H, Munos R, Simonyan K, Mnih V, Ward T, Doron Y, Firoiu V, Harley T, Dunning I et al (2018) Importance weighted actor-learner architectures: Scalable distributed deep-rl with importance weighted actor-learner architectures. arXiv preprint. arXiv:1802.01561

Evans R, Gao J (2016) Deepmind AI reduces google data centre cooling bill By 40%. DeepMind Blog 20:158

Figlewski S (1984) Hedging performance and basis risk in stock index futures. J Finance 39(3):657–669

Floros C, Vougas DV (2004) Hedge ratios in greek stock index futures market. Appl Financial Econ 14(15):1125–1136

Ganesh S, Vadori N, Xu M, Zheng H, Reddy P, Veloso M (2019) Reinforcement learning for market making in a multi-agent dealer market. *arXiv preprint* arXiv:1911.05892

Ghosh A (1993) Hedging with stock index futures: estimation and forecasting with error correction model. J Futures Markets 13(7):743–752

Halperin I (2017) Qlbs: Q-learner in the black-scholes (-merton) worlds. *Available at SSRN 3087076*

Hancock GD, Weise PD (1994) Competing derivative equity instruments: Empirical evidence on hedged portfolio performance. J Futures Markets (1986–1988) 14(4):421

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Huggenberger M, Albrecht P, Pekelis A (2016) Tail risk hedging and regime switching. In: Asian Finance Association (AsianFA) 2015 Conference Paper

Konda VR, Tsitsiklis JN (2000) Actor-critic algorithms. In: Advances in Neural Information Processing Systems, pp 1008–1014

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105

Liang E, Liaw R, Moritz P, Nishihara R, Fox R, Goldberg K, Gonzalez JE, Jordan MI, Stoica I (2017) Rllib: Abstractions for distributed reinforcement learning. arXiv preprint. arXiv:1712.09381

McDonald Robert L (2006) Derivatives markets, 2nd edn. Addison Wesley, Boston

Miller MB (2018) Quantitative financial risk management. Wiley, Hoboken

Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M (2013) Playing atari with deep reinforcement learning. arXiv preprint. arXiv:1312.5602

Moody J, Saffell M, Liao Y, Wu L (1998) Reinforcement learning for trading systems and portfolios: Immediate vs future rewards. In: Decision technologies for computational finance, pp 129–140. Springer

Moody J, Lizhong W, Liao Y, Saffell M (1998) Performance functions and reinforcement learning for trading systems and portfolios. J Forecast 17(5–6):441–470

O'Kelly M, Sinha A, Namkoong H, Tedrake R, Duchi JC (2018) Scalable end-to-end autonomous vehicle testing via rare-event simulation. In: Advances in neural information processing systems, pp 9827–9838

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A (2017) Mastering the game of go without human knowledge. Nature 550(7676):354–359

Sutton RS, Barto AG (2018) Reinforcement learning: an introduction, 2nd edn. MIT press, Cambridge

Topaloglou N, Vladimirou H, Zenios SA (2002) Cvar models with selective hedging for international asset allocation. J Bank Finance 26(7):1535–1561

Watkins C, John CH (1989) Learning from delayed rewards

Watts S (2015) Hedging basis risk using reinforcement learning. Technical report, Working Paper, University of Oxford

Zhang K, Yang Z, Başar T (2019) Multi-agent reinforcement learning: a selective overview of theories and algorithms. arXiv preprint. arXiv:1911.10635

Zhang Z, Zohren S, Stephen R (2020) Deep reinforcement learning for trading. J Financial Data Sci