2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 1-4, 2023, Oahu, Hawaii, USA

Truncated Quantile Critics Algorithm for Cryptocurrency Portfolio Optimization

Leibing Xiao
Northwestern Polytechnical University
Shanxi, China
x13033968676@mail.nwpu.edu.cn

Xinchao Wei
Northwestern Polytechnical University
Shanxi, China
xcwei@mail.nwpu.edu.cn

Yuelei Xu
Northwestern Polytechnical University
Shanxi, China
xuyuelei@nwpu.edu.cn

Xin Xu
Northwestern Polytechnical University
Shanxi, China
xinxu1999@163.com

Kun Gong
Northwestern Polytechnical University
Shanxi, China
gongkun@mail.nwpu.endu.cn

Huafeng Li
Northwestern Polytechnical University
Shanxi, China
2021264560@mail.nwpu.edu.cn

Fan Zhang
Northwestern Polytechnical University
Shanxi, China
fanzhang9876@163.com

2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC) | 979-8-3503-3702-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/SMC53992.2023.10393867

Abstract—This paper investigates portfolio management algorithm for the cryptocurrency market by using the TQC (Truncated Quantile Critics) algorithm. The study is based on the daily prices of cryptocurrencies. TQC is a deep reinforcement learning algorithm with the Actor-Critic architecture. It alleviates the overestimation problem of traditional value learning algorithm. In this paper, the data of cryptocurrencies are first processed as input to the networks. The inputs to the networks include not only the closing prices of cryptocurrencies, but also the relative strength index, moving average line, and moving average convergence divergence. Various metrics measuring algorithm returns and algorithm stability are used as evaluation criteria in this paper. In this paper, common deep reinforcement learning algorithms are compared. The experimental results show that the TQC algorithm has a highest return of 33.9 % during the test period, which is 3 %, 3 % and 15.6 % higher than A2C, PPO and DDPG respectively. And, the TQC algorithm has the highest stability of return, which is an important evaluation metric for portfolio management algorithms. Despite the high volatility of the cryptocurrency market, the performance of the TQC algorithm has remained relatively stable. This illustrates the positive effects of the TQC algorithm.

Index Terms—Portfolio management, Modern portfolio theory, Deep reinforcement learning, Technical analysis

I. INTRODUCTION

The problem of portfolio management is a hot issue in finance research. Through the concerted efforts of many researchers, many portfolio management algorithms have emerged and have yielded good results. Currently, traditional trading policies are mainly based on pre-defined subjective indicators to determine trading moves. For example, the use of the relative strength index [1] or the opening range breakout [2] indicator can reduce the risk of an investment and increase the return on investment. But this approach ignores the non-linear nature of markets and the uncertainty in investment decision making. In the cryptocurrency trading market, uncertainty can be even greater due to the dramatic volatility of cryptocurrency prices.

Currently, to solve the portfolio management problem, researchers use classic machine learning and deep learning algorithms. By using machine learning algorithms, researchers can analyse and forecast market data to improve portfolio management result [3]. There are some machine learning algorithms which is used include support vector machines [4], decision tree [5], etc. However, the use of classical machine learning algorithms in portfolio management problems faces the disadvantages of high data dependency and vulnerability to interference. Deep learning algorithms are used to analyse and forecast market data to improve the effectiveness of portfolio management [6]. The algorithm has strong non-linear modelling and generalisation capabilities. However, deep learning algorithms also require a large amount of annotated data for training, and annotated data is relatively expensive to acquire and process.

To address these issues, some academics have chosen to train portfolio management agents using deep reinforcement learning algorithms. Giorgio Lucarelli [7] has designed a portfolio management framework based on deep Q-learning algorithms. They demonstrate experimentally that deep reinforcement learning algorithms have excellent performance in portfolio management problems. But he is using a value learning algorithm. More data features considered as states in the algorithm design process is an effective way to improve the performance of the algorithm. Policy learning algorithms are more advantageous when faced with large scale state spaces. Beyond this, the space for action on portfolio management issues is continuous. Policy learning is more applicable to problems with continuous action spaces because it is easier to search for optimal policy in a continuous action space.

Therefore, in this paper, we decided to extend Lucarelli's research and design the cryptocurrency market portfolio management framework based on the TQC (Truncated Quantile Critics) algorithm [8]. The main advantage of the TQC algorithm is its ability to reduce bias in the estimator and to better cope with reward sparsity and high-dimensional state spaces when dealing with continuous control task. The agent trained by the algorithm can give the best set of actions at each transaction. The agent trained by the algorithm can give the best set of actions at each transaction. The ultimate goal of this paper is to maximize portfolio returns for the agent.

More precisely, the agent gives the ratio corresponding to the current number of asset types. More precisely, the agent gives the ratio corresponding to the current number of asset types. The agent receives a reward from the environment after performing the action. The state of the agent is the daily price characteristic of the investment asset. The proposed TQC portfolio framework is tested on a cryptocurrency portfolio consisting of six cryptocurrency assets: Bitcoin (BTC), Ethereum (ETH), Litecoin (LTC), Riple (XRP), Tether (USDT), and Polkadot (DOT).

II. LITERATURE REVIE

A. Related works

The portfolio management approach has been receiving a lot of attention from academia and industry [9] because of its strong realistic significance. But deep reinforcement learning algorithms show great potential, such as AlphaGo [10] , ChatGPT, etc. Using deep reinforcement learning to solve portfolio management problems is in the spotlight.

For now, many scholars have improved the algorithm performance by changing the reward function. Because reinforcement learning is goal-oriented, setting the reward function is one of the key aspects of the reinforcement learning agent training process. A recurrent reinforcement learning objective function was proposed by Saud Almahdi et al [11] . This method has a Kalmar ratio for risk adjustment. It is experimentally demonstrated that the objective function based on the maximum risk of the expected goal yields superior return performance compared to the classical reward function. Also, they propose an adaptive portfolio rebalancing decision system, which due to hedge fund criteria. Amine et al. [12] constructed a real-time optimal portfolio framework by recursive reinforcement learning algorithms. The Sharpe rate, which is a risk measure, is used as a performance indicator during the algorithm training process.

In addition to improving the reward function, many scholars improve deep reinforcement learning algorithms applied to portfolio management problem. This approach allows the full application of research results in the field of deep reinforcement learning. For now, reinforcement learning is a hot research topic. Every year, new deep reinforcement learning algorithms appear. It is highly relevant to explore algorithms that are more suitable for application to portfolio management problems. Yunan Ye et al. [13] proposed a new state-augmented reinforcement learning framework to address the data heterogeneity and environmental uncertainty that arise in financial asset portfolio management. Farzan Soleymani et al. [14] used the Deep Breath algorithm which combines a restricted stacked autoencoder with a convolutional neural network. They used the Deep Breath algorithm to construct an asset portfolio management framework. The framework is divided into two parts: offline learning, which is used to train the convolutional neural network, and online learning, which deals with conceptual drift.

B. Reinforcement learning

Reinforcement learning [15] is an emerging machine learning technique that mimics the human learning process. Through constant trial, the agent can learn the correct action, the one that meets the requirements of the goal, through the reward returned from the environment. The reinforcement learning agent selects an action based on the state it is currently in.

Common reinforcement learning algorithm can be divided into two types: value learning algorithm and policy learning algorithm. The core of the value learning algorithm is the construction of a value function that estimates the next state [16]. After learning a more accurate value function, the agent chooses the action that achieves the next state of maximum value. The core of the value learning algorithm is the construction of a value function that estimates the next state. After learning a more accurate value function, the agent chooses the action that achieves the next state of maximum value. But the disadvantages of value learning algorithm are also obvious. The goal of value learning is to learn the optimal value function, but in practical application the problem of over- or under-estimation may occur. The treatment of continuous action spaces is more difficult: in continuous action spaces it is necessary to use function approximation method to learn the value function, but this can lead to over-fitting problem and requires special treatment.

The core of a policy learning algorithm is the construction of a policy function, usually a neural network [17] . The inputs to the policy network are the agent's states and the outputs are the agent's actions. Based on the reward obtained from the environment after the agent performs the action, the policy network can update the parameters by stochastic gradient descent. In contrast to value learning, policy learning can deal directly with the continuous action space without the need for discretization.

C. Truncated Quantile Critics,TQC

The TQC algorithm is a policy learning algorithm with an Actor-Critic [18] architecture. The Actor-Critic architecture is a reinforcement learning algorithm architecture that combines the advantages of policy learning and value learning, while overcoming their respective disadvantages. In the Actor-Critic architecture, the agent is divided into two parts: the Actor and the Critic.

The Actor is the policy learning component, which selects an action based on the current state and passes it to the environment for execution. The goal of the Actor is to maximise the expected return, and it updates its parameters so as to improve its policy. The output of the Actor is a probability distribution describing the probability that the agent will select each action in the current state. Critic is the value learning component which evaluates Actor's actions and estimates the value of the current state. the goal of Critic is to minimise the mean squared error in the current state.

Since the Cirtic network is used, the algorithm inevitably faces the problem of overestimation of value. In the continuous domain, the addition of a method to mitigate overestimation bias can significantly improve the performance of continuous policies [19] . Based on this idea, the TQC algorithm was proposed [8] . The TQC uses three ideas: distributional representation of a critic, truncation of approximated distribution, and ensembling.

III. METHODOLOGY

This section describes how to use the TQC algorithm to build a system for optimising portfolio management in the cryptocurrency market. First, the paper begins by describing the data used for the experiments. Afterwards, a deep reinforcement learning framework for applying cryptocurrency market is presented. Finally, indicators are presented, which are used to evaluate the performance of the algorithms.

A. Data

To validate the performance of the proposed algorithm, we backtracked to a cryptocurrency portfolio management environment that would have been constructed using test data. A total of six cryptocurrencies were used in the experimental environment for this paper.

The dataset is downloaded from the Yahoo Finance website. The dataset contains the opening, closing, high and low prices of cryptocurrency prices, and trading volume. For this article, we have chosen to use the daily closing prices of cryptocurrencies. The dataset has a start date of 1 January 2021 and an end date of 1 January 2023. The dataset is divided into two sets: a training set and a test set. The starting date for the training set is 1 January 2021 and the ending date is 1 October 2022. The rest of the data is the test set.

In addition to the given closing prices of cryptocurrencies, this paper considers a number of technical indicators from modern financial theory. The technical indicators used in the model are described in Section III-B.

B. Technical analysis

Technical indicators indicate the current state of cryptocurrency price and also contain information from the past. It can explain how cryptocurrency prices have changed over time. By using these technical indicators, agents can apply modern financial knowledge and then get a better performance. A total of six technical indicators are used in this paper, namely moving average line(MA) , the relative strength index (RSI), and moving average convergence divergence (MACD).

MACD describes the price movement trend of a cryptocurrency and is defined by two moving averages:

$$MA_w(t) = \frac{\sum_0^{w-1} P_t}{W} \tag{1}$$

$$EMA_w(t) = \alpha * P(t) + (1-\alpha) * EMA_w(t-1), \alpha = \frac{2}{1+w} \tag{2}$$

$$MACD_w(t) = \sum_{i=1}^{w} EMA_k(i) = \sum_{i=1}^{w} EMA_d(i) \tag{3}$$

where P_t represents the closing price of the cryptocurrency on day t and w denotes the length of the time window. MA represents a moving average; EMA is an exponential moving average with attenuation weighting.

The relative strength index (RSI) is a technical indicator used in the analysis of financial markets. It is intended to chart the current and historical strength or weakness of a market based on the closing prices of a recent trading period. The calculation formula is as follows:

$$RSI_{\mathrm{t}}(\mathrm{w}) = 100^* \left(1 - \frac{1}{1 + \frac{EMA_w(\max(P_t - P_{t-1}, 0)}{EMA_w(\min(P_t - P_{t-1}, 0))}} \right) \tag{4}$$

C. Reinforcement learning

This paper uses the TQC algorithm to solve the portfolio allocation problem in the cryptocurrency market. The environment of this paper is constructed through the daily closing prices of cryptocurrencies. Therefore, the time step for each action in this paper is 1 day. The termination date of the data used in the training environment of this paper is October 1, 2022 and the start date is January 1, 2021. Since the price of the 6 cryptocurrencies selected in this

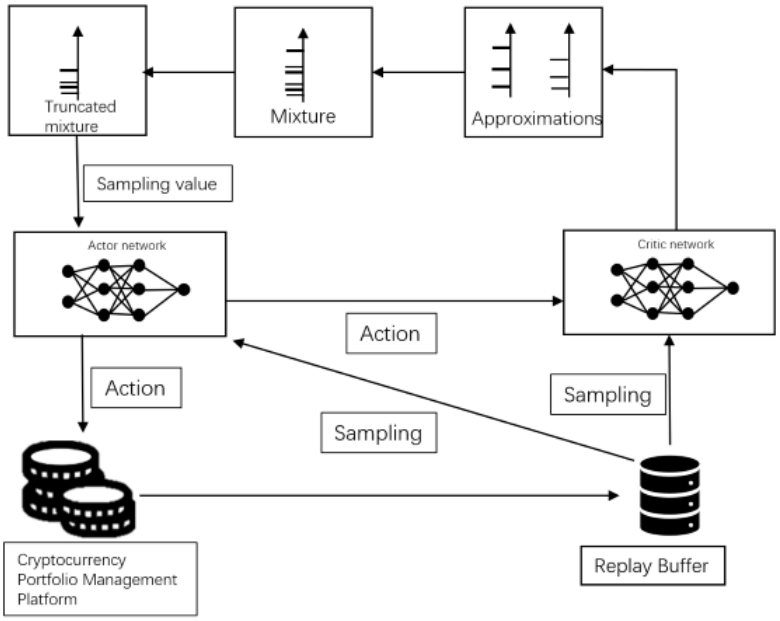

Fig. 1. TQC portfolio management framework.

paper is not 0, it is not necessary to consider the case where the account is 0.

The portfolio that we study in this paper has a total of seven assets, including cash and six cryptocurrencies. Assume that at time step t, the weight of each asset i in the portfolio is $w_{i,t}, (0 \leq w_{i,t} \leq 1)$, and the closing price of the asset is $c_{i,t}, (c_{i,t} > 0)$. Assume that at time t, the cash in the portfolio is m_t and the number of i cryptocurrencies is $n_{i,t}$.

At this point, the formula for calculating the value $y_{i,t}$ of the ith cryptocurrency is shown below:

$$y_{i,t} = n_{i,t} * c_{i,t} \tag{5}$$

The total assets v_t of the portfolio at this time are

$$v_t = m_t + \sum_{i=0}^{6} n_{i,t} * c_{i,t} \tag{6}$$

After performing the allocation w_{it}, the portfolio is assumed to have a value of v_{t+1} at $t+1$.

Assume that the number of ith cryptocurrency is $n_{i,t+1}$ at $t+1$.

$$n_{i,t+1} = v_t * w_{i,t}/c_{i,t} \tag{7}$$

The transaction fee for the ith cryptocurrency is $fee_{i,t}$.

$$fee_{i,t} = \max(2, (n_{i,t+1} - n_{i,t}) * c_{i,t} * 0.02) \tag{8}$$

Then, v_{t+1} can be obtained.

$$v_{t+1} = v_t * w_{7,t} + \sum_{i=0}^{6} n_{i,t+1} * c_{i,t+1} - fee_{i,t} \tag{9}$$

where $w_{7,t}$ denotes the proportion of cash in the portfolio.

At time step t, the environmental reward is r_t. In this paper, we choose to use the change in property after performing the action as a reward for simulating the environment. The calculation formula is as follows:

$$r_t = (v_{t+1} - v_t) * \beta \tag{10}$$

where β is a reward scale factor, $\beta = 10^{-4}$ in this paper. The specific TQC framework is shown in Figure 1.

IV. Experimental setup and results

A. Benchmarks

For comparison purposes, a number of widely used benchmarks have been selected for this paper. A total of three common deep reinforcement learning algorithms are chosen as evaluation criteria in this paper, namely A2C (Advantage Actor-Critic), DDPG (Deep Deterministic Policy Gradient), and PPO (Proximal Policy Optimization).

1) A2C: The A2C algorithm is a reinforcement learning algorithm based on the Actor-Critic architecture, designed to solve the optimization problem of the policy (Actor) and value functions (Critic).
2) DDPG [21] : The DDPG algorithm is also a deep reinforcement learning algorithm based on the Actor-Critic architecture, which solves the problem of continuous action space. The DDPG algorithm is based on an empirical replay mechanism using historical data for training, which is more effective.
3) PPO [22] : The PPO algorithm is a policy iteration algorithm for reinforcement learning that belongs to the category of model-independent algorithm in policy-based reinforcement learning method. The

main advantage of the PPO algorithm is its ability to stabilise model updates to some extent, while avoiding some of the problems of policy gradient algorithms.

B. Evaluation

A variety of indicators are used to judge the performance of a portfolio's selection strategy. In this paper, six common metrics have been selected to test the algorithm.

1) Cumulative return: Cumulative returns are the total return earned from holding an asset from a certain point in time.
2) Volatility: Also known as return volatility, it is a statistical indicator used to describe the range of volatility of an investment within the test period.
3) Sharpe ratio: It is an indicator used to measure the performance of a portfolio in terms of risk and return.
4) Calmar ratio: An indicator that assesses the performance of a portfolio and is usually used to measure the trade-off between overall performance and maximum retracement.
5) Stability: Also known as volatility stability, this is a derivative of volatility. The more stable an investment is, the less volatile its return will be, representing a lower capital risk for the investment.
6) Max drawdown: It is the maximum loss of a portfolio from the top to the bottom of the portfolio. It is a standard indicator for assessing risk, as it takes into account the risk of loss of the entire portfolio.

C. Results

In this paper, we choose to train the algorithm 100,000 times in the training environment. Then, the best performing network models in the last 10 training episodes are recorded. The four algorithms are run separately in the test environment and the evaluation metrics of each algorithm are compared.

TABLE I
Backtest Results.

Algorithm	A2C	PPO	DDPG	TQC
Cumulative return	0.329	0.328	0.293	0.339
Volatility	1.755	1.751	1.870	1.512
Sharpe ratio	1.463	1.460	1.441	1.459
Calmar ratio	5.282	5.266	4.313	6.483
Stability	0.684	0.682	0.641	0.732
Max drawdown	-0.275	-0.274	-0.309	-0.212
Omega ratio	1.340	1.339	1.333	1.333

Table I shows the results of the agents in the test environment, where the best values of the evaluation indicators are bolded. It can be seen that the TQC algorithm has a higher return of 33.9 % during the test period, which is 3 %, 3 % and 15.6 % higher than A2C, PPO and DDPG respectively. It can be seen that the TQC algorithm has a higher return of 33.9 % during the test period, which is 3 %, 3 % and 15.6 % higher than A2C, PPO and DDPG respectively. Also, the TQC algorithm shows higher stability in a highly volatile cryptocurrency trading environment due to the lower volatility of the TQC algorithm.

As can be seen from Table I, the TQC algorithm obtained the highest Calmar ratio, and the lowest Max drawdown values. It demonstrates that the TQC algorithm is more tolerant of losses. Unfortunately, however, the Sharpe ratio of the TQC algorithm is 1.459, which is lower than the Sharpe ratios of A2C and PPO. It demonstrates that the TQC algorithm may select riskier assets.

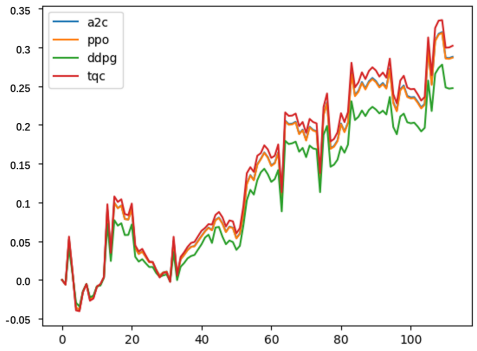

Fig. 2. Cumulative return during testing.

The cumulative return for each algorithm during testing is shown in Figure 2. As can be seen from Figure 2, the trend in cumulative returns over the testing process is similar for all four algorithms. However, the TQC algorithm has the highest cumulative return.

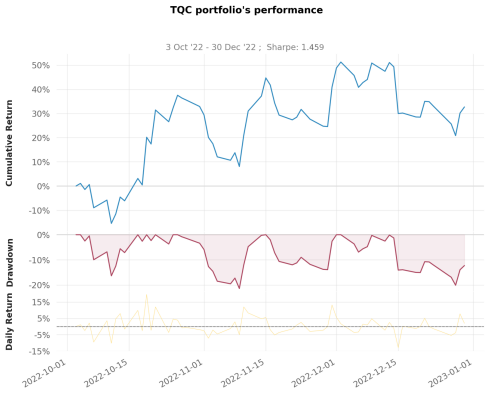

Fig. 3. TQC portfolio's performance

Figure 3 shows the performance during the training period during the test. It includes the cumulative return curve, the Drawdown curve, and the Daily return curve.

V. CONCLUSION AND FUTURE WORK

The issue of portfolio management in the cryptocurrency market will always be a focus of research due

to the widespread interest it has attracted. Due to the high volatility of cryptocurrency price, obtaining a high-performing policy can be difficult. In this paper, the TQC algorithm is applied to the problem of cryptocurrency portfolio management. It is experimentally demonstrated that the TQC algorithm has a higher return of 33.9 % during the test period, which is 3 %, 3 % and 15.6 % higher than A2C, PPO and DDPG respectively. And, the TQC algorithm has the highest stability of return, which is an important evaluation metric for portfolio management algorithms.

This paper chooses to use deep reinforcement learning algorithm to solve the cryptocurrency portfolio management problem. For the input to the neural network, we can in the future test a variety of metrics from modern management science and select the most suitable input metric through the variation of reward.In the meantime, a great deal of research has focused on improvement to the reward function of deep reinforcement learning algorithm. This paper uses the classical reward as the reward. In the future, we hope to combine multiple evaluation criteria as reward. By doing so, the algorithm will be used in a wider range of scenarios.

References

[1] ran-M. Adrian, "The relative strength index revisited," African Journal of Business Management, vol. 5, no. 14, pp. 5855–5862, 2011.

[2] J.-H. Syu, M.-E. Wu, S.-H. Lee, and J.-M. Ho, "Modified orb strategies with threshold adjusting on taiwan futures market," in 2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), 2019, pp. 1–7.

[3] S. M. Bartram, J. Branke, G. De Rossi, and M. Motahari, "Machine learning for active portfolio management," The Journal of Financial Data Science, vol. 3, no. 3, pp. 9–30, 2021.

[4] C. J. Burges, "A tutorial on support vector machines for pattern recognition," Data mining and knowledge discovery, vol. 2, no. 2, pp. 121–167, 1998.

[5] E. B. Hunt, J. Marin, and P. J. Stone, "Experiments in induction.," 1966.

[6] Z. Zhang, S. Zohren, and S. Roberts, "Deep learning for portfolio optimization," The Journal of Financial Data Science, vol. 2, no. 4, pp. 8–20, 2020.

[7] G. Lucarelli and M. Borrotti, "A deep Q-learning portfolio management framework for the cryptocurrency market," Neural Computing and Applications, vol. 32, pp. 17229–17244, 2020.

[8] A. Kuznetsov, P. Shvechikov, A. Grishin, and D. Vetrov, "Controlling overestimation bias with truncated mixture of continuous distributional quantile critics," in International Conference on Machine Learning, 2020, pp. 5556–5566.

[9] D. Kim and J. C. Francis, Modern portfolio theory: Foundations, analysis, and new developments. John Wiley & Sons, 2013.

[10] D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," nature, vol. 529, no. 7587, pp. 484–489, 2016.

[11] S. Almahdi and S. Y. Yang, "An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown," Expert Systems with Applications, vol. 87, pp. 267–279, 2017.

[12] A. M. Aboussalah and C.-G. Lee, "Continuous control with stacked deep dynamic recurrent reinforcement learning for portfolio optimization," Expert Systems with Applications, vol. 140, p. 112891, 2020.

[13] Y. Ye et al., "Reinforcement-learning based portfolio management with augmented asset movement prediction states," in Proceedings of the AAAI conference on artificial intelligence, 2020, vol. 34, no. 01, pp. 1112–1119.

[14] F. Soleymani and E. Paquet, "Financial portfolio optimization with online deep reinforcement learning and restricted stacked autoencoder—DeepBreath," Expert Systems with Applications, vol. 156, p. 113456, 2020.

[15] S. Thrun and M. L. Littman, "Reinforcement learning: an introduction," AI Magazine, vol. 21, no. 1, pp. 103–103, 2000.

[16] V. Mnih et al., "Playing atari with deep reinforcement learning," arXiv preprint arXiv:1312.5602, 2013.

[17] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," arXiv preprint arXiv:1509.02971, 2015.

[18] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," Advances in neural information processing systems, vol. 12, 1999.

[19] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in International conference on machine learning, 2018, pp. 1587–1596.

[20] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in International conference on machine learning, 2017, pp. 449–458.

[21] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," arXiv preprint arXiv:1509.02971, 2015.

[22] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.