Letter

Adaptive strategy optimization in game-theoretic paradigm using reinforcement learning

Kang Hao Cheong *

Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, S637371 Singapore and College of Computing and Data Science, Nanyang Technological University, S639798, Singapore

Jie Zhao

Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, S637371 Singapore and Science, Mathematics and Technology, Singapore University of Technology and Design, S487372 Singapore, Singapore

(Received 12 March 2024; accepted 26 April 2024; published 10 July 2024)

Parrondo's paradox refers to the counterintuitive phenomenon whereby two losing strategies, when alternated in a certain manner, can result in a winning outcome. Understanding the optimal sequence in Parrondo's games is of significant importance for maximizing profits in various contexts. However, the current predefined sequences may not adapt well to changing environments, limiting their potential for achieving the best performance. We posit that the optimal strategy that determines which game to play should be learnable through experience. In this Letter, we propose an efficient and robust approach that leverages Q learning to adaptively learn the optimal sequence in Parrondo's games. Through extensive simulations of coin-tossing games, we demonstrate that the learned switching strategy in Parrondo's games outperforms other predefined sequences in terms of profit. Furthermore, the experimental results show that our proposed method can be easily adjusted to adapt to different cases of capital-dependent games and history-dependent games.

DOI: 10.1103/PhysRevResearch.6.L032009

Inspired by the dynamic nature of flashing Brownian ratchets, Parrondo's paradox demonstrates that alternating between two strategies, each of which will individually result in losses, can surprisingly lead to a winning outcome. This intriguing paradox has piqued interest and found relevance in varied fields such as quantum systems [1,2], biology [3,4], and encryption [5], demonstrating its wide-ranging applications [6]. A player selects between two games, A and B, which is individually a losing game. When played in a certain sequence, it can surprisingly result in a winning outcome in the long run. Determining the optimal sequence for switching between games is crucial for maximizing outcomes, as highlighted by Dinis [7]. There is the capital-dependent Parrondo's paradox which is reliant on the current capital of the player, while the history-dependent variant relies on the past wins/losses to decide on the game to be played. The sequential arrangement of playing the games can determine whether one wins or loses the game in the long term. For example, the periodic "ABABB" sequence has been devised as the optimal sequence for the capital-dependent Parrondo's paradox, drawing support from both theoretical analysis and empirical observations [8]. However, this is optimal only when the sequence is specified as periodic. Therefore, several attempts have further been made to discover the optimal sequence in an arbitrary order, with researchers exploring different methodologies. One notable approach is the utilization of genetic algorithms [9,10], which are renowned for their powerful search abilities in combinatorial problems [11,12]. The convergence of these multifaceted approaches empowers researchers to study the intricate interplay between game ordering and winning outcomes within Parrondo's paradox, gaining new insights into the underlying dynamics and optimizing strategies for maximum advantage. Nevertheless, these methods typically search for a fixed sequence, which inherently restricts the scope of their potential applications for the following reasons.

Parrondo's games exhibit complex dynamics and nonlinear relationships, making it difficult to anticipate optimal sequences solely based on static analysis. The performance of Parrondo's games is sensitive to specific conditions, such as the current capital. Predefined sequences are typically designed based on theoretical or empirical insights, which fail to adapt to these changing conditions, as they remain fixed regardless of the environment. Consequently, the suboptimal sequencing may fail to fully capitalize on advantageous conditions or adequately mitigate the deleterious effects of unfavorable circumstances, thus diminishing the game's overall performance. This suboptimality can arise due to a lack of adaptability to dynamic environments, or inadequate incorporation of feedback mechanisms and control strategies. The consequence is an attenuation of the game's ability to exploit beneficial opportunities and navigate challenging scenarios, highlighting the critical importance of discerning and implementing optimal sequencing schemes to attain favorable outcomes.

*kanghao.cheong@ntu.edu.sg

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Some works such as Refs. [13–15] have already noticed that policy based on the current state can achieve the maximum profit and proposed to use adaptive strategy. However, there still lacks a general framework that can accommodate complex situations, such as different M in the coin-tossing game within the capital-dependent Parrondo's paradox or other complex tasks. To address this issue, we use reinforcement learning that emerges as a compelling solution to overcome the limitations of predefined sequences in Parrondo's games. Our contributions can be summarized as follows.

(i) The dynamic and adaptable nature of reinforcement learning, driven by learning from experience, enables informed decision-making in response to changing conditions. Therefore, we explore employing reinforcement-learning algorithms to find the optimal sequence in Parrondo's games. In this way, we can adaptively update the sequence, allowing for the discovery of optimal sequences that outperform static approaches.

(ii) We have studied the extensibility of reinforcement learning in finding the optimal sequence in Parrondo's games. The experimental results show that our proposed method can be easily tailored to adapt to different problems by simply modifying the size of state space. Furthermore, in addition to the well-studied capital-dependent games, we have also explored the applicability of our proposed method in history-dependent games.

We first review some basic definitions of Parrondo's games [8] and elucidate the methodology for modeling Parrondo's games [6,16] within the framework of reinforcement learning. In particular, we will explore the application of Q-learning techniques to identify the optimal sequence of moves.

Game A. In game A, a player tosses a coin and receives a win if it lands on heads, and a loss if it lands on tails. In particular, the probability p_1 of the coin landing on heads is given by $0.5-\epsilon$, where ϵ is a small positive constant, and the probability of the coin landing on tails is given by $0.5+\epsilon$. In other words, the coin is biased toward tails. The player starts with an initial capital of 0, and each win adds 1 unit to their capital, while each loss subtracts 1 unit from their capital. Game A is a Markov process and can be modeled as follows:

$$\Pi_A = \begin{bmatrix} 0 & 1-p_1 & p_1 \\ p_1 & 0 & 1-p_1 \\ 1-p_1 & p_1 & 0 \end{bmatrix}. \tag{1}$$

Game B. In game B, the player has two coins to choose from, coin 2 and coin 3. Both coins have a different probability of landing on heads or tails. Coin 2 has a probability of p_2 of landing on heads, and a probability of $1-p_2$ for landing on tails. Coin 3 has a probability of p_3 for landing on heads, and a probability of $1-p_3$ for landing on tails. Again, the player starts with an initial capital of 0, and each win adds 1 unit to their capital, while each loss subtracts 1 unit from their capital. The player's coin selection is based on the current capital and a parameter M. If the current capital is a multiplier of M, coin 2 is chosen; otherwise, coin 3 is selected. Game B is a Markov process and can be modeled as follows:

$$\Pi_B = \begin{bmatrix} 0 & 1-p_3 & p_3 \\ p_2 & 0 & 1-p_3 \\ 1-p_2 & p_3 & 0 \end{bmatrix}. \tag{2}$$

For game A, let t be the round of game and $X(t)$ be the average capital at round t, then we have

$$X(t+1) = X(t) + 2p_1 - 1. \tag{3}$$

Similarly, the average capital at round t can be obtained as follows if game B is played at round t,

$$X(t+1) = X(t) + 2p_{\text{win}_\text{B}}(t) - 1, \tag{4}$$

where p_{win_B} refers to the probability of winning of game B.

Let $\pi_0(t)$ be the probability that the capital at round t is a multiple of 3, then we can have $p_{\text{winB}}(t)$ as follows:

$$p_{\text{winB}}(t) = \pi_0(t)p_2 + [\pi_1(t) + \pi_2(t)]p_3. \tag{5}$$

Similarly, $\pi_1(t)$ and $\pi_2(t)$ refer to the likelihood that the capital is a multiple of 3 with the remainder of 1 or 2, respectively, and $\pi(t) \equiv [\pi_0(t), \pi_1(t), \pi_2(t)]^T$.

We can define the expected gain at round t as follows:

$$g(t) \equiv X(t+1) - X(t). \tag{6}$$

The above equation is satisfied for whatever game is played. For each case, we can have

$$g(t) = \begin{cases} g^A & \text{if A is played at } t, \\ g^B & \text{if B is played at } t. \end{cases} \tag{7}$$

The expected gain of game A is as follows:

$$g^A \equiv 2p_1 - 1. \tag{8}$$

Similarly, game B is as follows:

$$g^B \equiv 2[\pi_0(t)p_2 + [\pi_1(t) + \pi_2(t)]p_3] - 1. \tag{9}$$

Therefore, the total gain of playing game for T rounds is

$$G_T = \sum_{t=1}^{T} g(t). \tag{10}$$

Let α_t denote the game to play at round t which can only have values of A or B. The problem now becomes that of finding the sequence $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ to maximize G_T, with

$$\begin{aligned} \boldsymbol{\pi}(t+1) &= \Pi_A \boldsymbol{\pi}(t), \quad \text{if } \alpha_t = \text{A}, \\ \boldsymbol{\pi}(t+1) &= \Pi_B \boldsymbol{\pi}(t), \quad \text{if } \alpha_t = \text{B}. \end{aligned} \tag{11}$$

Let $\hat{G}_n(\pi)$ be the maximum expected gain at round n. Then we can have the expected gain if we play game A at round $n-1$ as follows:

$$g^A + \hat{G}_{n-1}(\Pi_A \pi). \tag{12}$$

Similarly, we can have the following if we play game B:

$$g^B + \hat{G}_{n-1}(\Pi_B \pi). \tag{13}$$

Then, we can choose the game to play based on

$$\hat{G}_n(\boldsymbol{\pi}) = \max\left\{g^A + \hat{G}_{n-1}(\Pi_A \boldsymbol{\pi}), g^B + \hat{G}_{n-1}(\Pi_B \boldsymbol{\pi})\right\}. \tag{14}$$

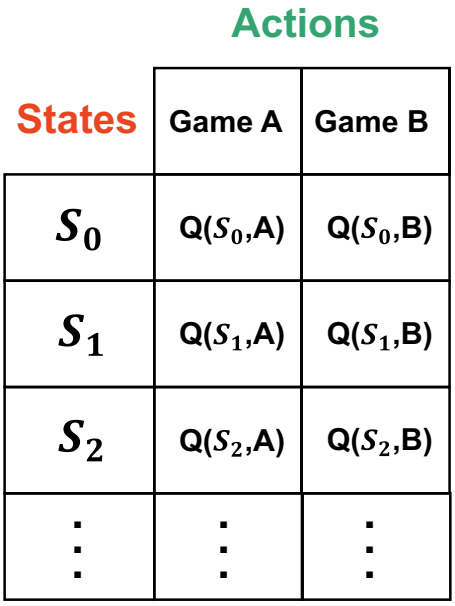

FIG. 1. The diagram of Q-table in Parrondo's paradox.

It is obvious that the optimal game to play is dependent on the latest result only. Due to the Markov nature of Parrondo's games [15], we are motivated to approach the problem with a technique that can adapt to the changing environment. In this Letter, our focus is on employing the Q-learning algorithm [17,18], a reinforcement-learning method, to discover the optimal strategy. Unlike the typical switching strategy (periodic or random) [19,20], this method will follow a sequential decision-making framework, where an agent interacts with the environment, learns from its actions, and updates its policy based on observed rewards. In this work, the reward is set to the gain of the played game. Our proposed method begins by defining the state space, which is based on the type of game. For the capital-dependent game, the state space consists of three possible remainders of the current capital divided by 3, i.e., 0, 1, or 2.

In the context of the history-dependent Parrondo's paradox, game A is the same as the capital-dependent one while game B is different. Determining the game scenario to be played hinges upon the sequential results of the preceding two games instead of the current capital. As each game can culminate in either a win or a loss, a total of four distinct combinations emerge from the possible outcomes of the previous two games: {lose, lose}, {lose, win}, {win, lose}, and {win, win}. These four combinations, in turn, correspond to four distinct coin options that can be chosen. In this case, the state space is set to the aforementioned four possible scenarios. The action space is determined by the available choices, specifically selecting either game A or game B at each decision.

The core implementation revolves around the Q-learning algorithm loop. The agent iteratively selects actions based on an ϵ-greedy strategy. Upon executing an action, the agent receives a reward and subsequently updates the Q value for the corresponding state-action pair using the Q-learning update rule as follows,

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha[r + \gamma \max_a Q(s',a)], \quad (15)$$

where $Q(s,a)$ is a state-action pair, referring to the expected cumulative reward. α denotes the learning rate. r is the immediate reward obtained after taking action a in state s. γ refers to the discount factor and determines the importance of future rewards in the learning process. The diagram of the Q-table defined in our problem is shown in Fig. 1 and the pseudocode is shown in Algorithm 1.

ALGORITHM 1. Q learning for optimal sequence in Parrondo's paradox.

```
Initialize Q-table with random or zero values
Define state space and action space
Set hyperparameters: learning rate α, discount factor γ,
    exploration rate ϵ
Initialize episode counter e ← 1
While not convergence criteria met do
    Initialize state s
    while not end of episode do
        Select action a based on ϵ-greedy strategy
        Execute action a in the environment
        Observe reward r and new state s′
        Update Q-value
        Update current state: s ← s′
    end while
    Increment episode counter: e ← e + 1
end while
Output: Optimal strategy
```

Our proposed method will be validated using a genetic algorithm and theoretical strategy to search for optimal game sequences. This approach is conducted under various settings, with a relaxed structural framework. The benchmark methods are as follows:

Random sequence. A random sequence is introduced as a comparison with the performance of our proposed method. The random sequence involves randomly selecting game A or game B, without any bias or intelligent decision-making process.

Fixed sequence. The ABABB sequence, identified through theoretical analysis when $M = 3$ [8], is also taken as one of the benchmarks. As for $M = 2$ and $M = 4$, we empirically use AB and AABB as a form of comparison.

Genetic algorithm. Inspired by Ref. [9], we adopt the genetic algorithm as one of the methods in searching for the optimal sequence, comprising the following components:

Population. We start by creating an initial population of random sequences of games. In this case, we can represent a sequence of games as a list of 0 and 1, where 0 represents playing game A, and 1 represents playing game B.

Fitness. For each individual in the population, we calculate its fitness (i.e., capital) by averaging 10^3 simulations of the sequence.

Selection. We use tournament selection to select individuals for reproduction. This means we randomly choose a subset of individuals with a high fitness from the population.

Crossover. We use the single-point crossover to combine two individuals. This involves selecting a random point in the sequence and swapping the subsequences before and after that

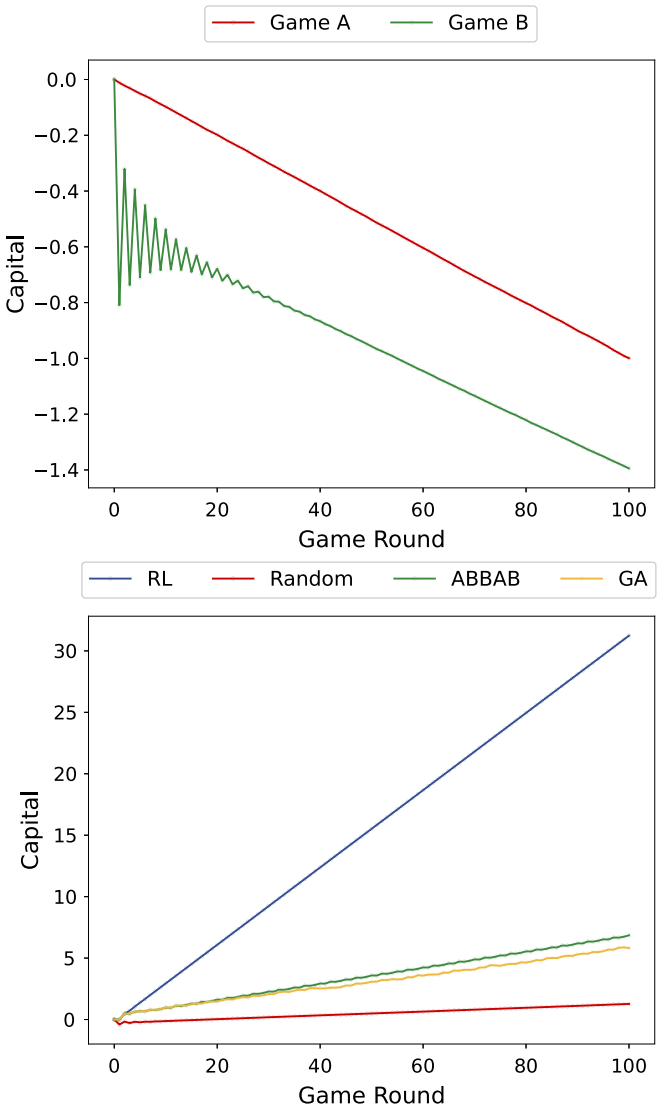

FIG. 2. Simulation results for the capital-dependent game ($M = 3$).

point between the two parents. The crossover probability is set to 0.5.

Mutation. We adopt random mutations by flipping a random bit in the sequence with a small probability. The mutation probability is set to 0.1.

We start with a standard scenario of $M = 3$, conducting 100 game rounds across 10^6 simulation repetitions. In keeping with the consistency of other works [9,13], we adopt the following parameter settings: p_1 is set to $1/2 - \epsilon$, and p_2 and p_3 are set to $1/10 - \epsilon$ and $3/4 - \epsilon$ when playing the coin-tossing game, respectively. The initial capital is set to 0 and ϵ is set to 0.005. As part of our experiments, we will also include the case of $M = 2$ and $M = 4$ as exploratory exposition. To illustrate the adaptability of our method, the history-dependent Parrondo's paradox is also investigated. Through iterative updates to the Q values based on observed rewards and actions, the agent incrementally learns the optimal policy. Each episode is completed when the capital reaches 20 and the number of episodes is set to 1000. The

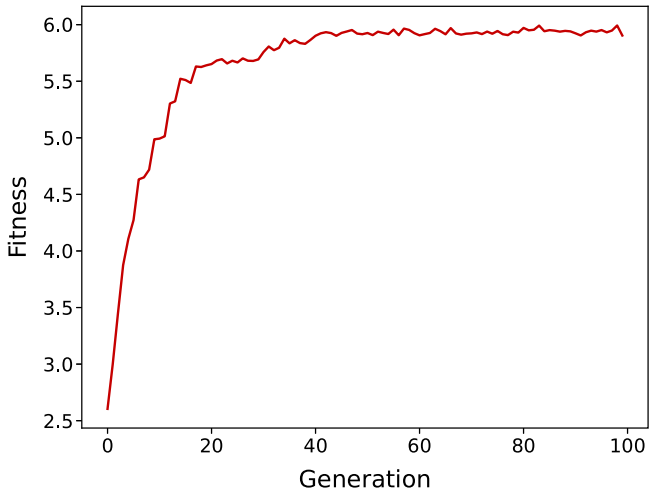

FIG. 3. Simulation results for the sequence searched by the genetic algorithm.

learning rate in updating the Q-table is set to 0.001, and the discount factor is set to 0.9.

The results from playing game A and game B individually can be found in Fig. 2. As observed, both games will yield a losing result in the long run. The comparison of the different methods such as reinforcement learning, fixed sequence, and random sequence is given in Fig. 2. These experimental results underscore the superior performance of the proposed adaptive reinforcement-learning (RL) algorithm in capital accrual, outstripping both the random sequence and the optimal sequence predicated on theoretical calculations. This indicates the efficacy of the adaptive RL-based approach, particularly when the switching strategy is not constrained to adhere strictly to periodic or stochastic sequences. As another method deviating from the periodic or random sequence, the genetic algorithm is also used to search the optimal sequence, as shown in Fig. 3. We employ elitism to make sure the individual with the highest fitness can be retained. Nevertheless, it is found that the optimal fitness values in the population do not always rise as expected but fluctuate slightly. We then take the best individual obtained in the last round as the optimal sequence, which achieves the capital of about 5, higher than the random sequence. However, the sequence searched by the genetic algorithm is still predefined. Although this metaheuristic method breaks the constraint of random and periodic manner, its performance is still not comparable to the status-aware method, as shown in Fig. 2. This result reveals a promising potential for utilizing adaptive decision-making models, in contrast to predefined or random approaches, thereby advancing the prospects of greater economic yield.

The result of $M = 2$ is shown in Fig. 4. As observed, game B will lose more money than in the case of $M = 3$ in the long run, reaching a much lower capital of -15. Unlike the traditional Parrondo's game, the random interplay of the two strategies cannot turn the result into a positive one, as observed in Fig. 4. On the other hand, the gain of profit is heavily subject to the selection of sequence when adopting the periodic manner. The sequence AABB can only achieve the same profit as random sequence, both of which are negative, while AB can lead to a capital gain of up to 25, as observed in Fig. 4.

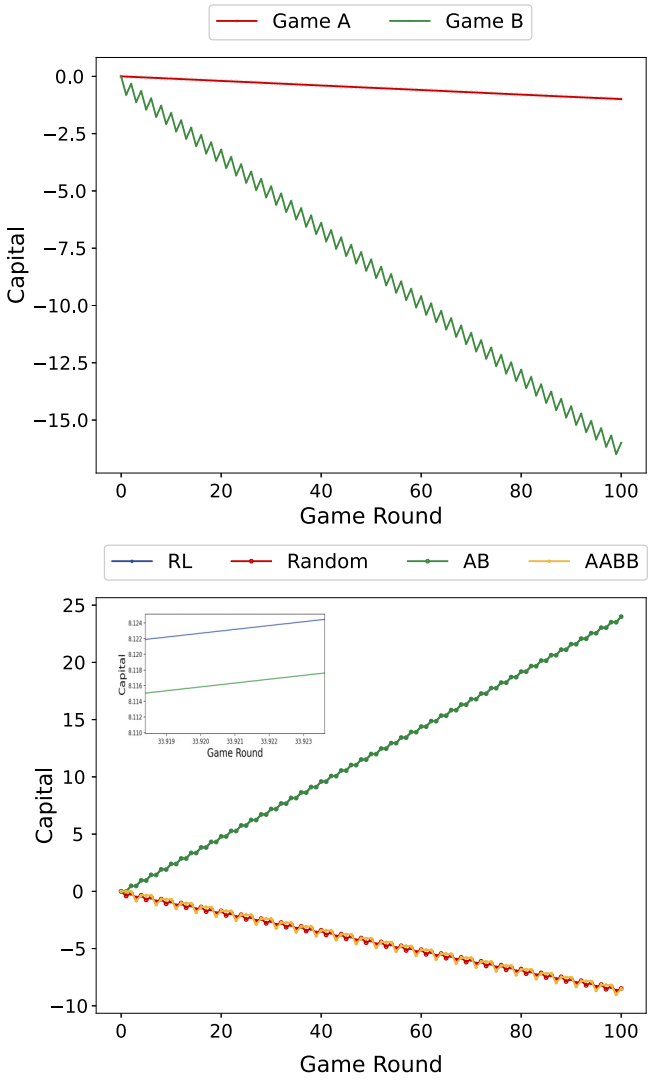

FIG. 4. Simulation results for the capital-dependent game ($M = 2$).

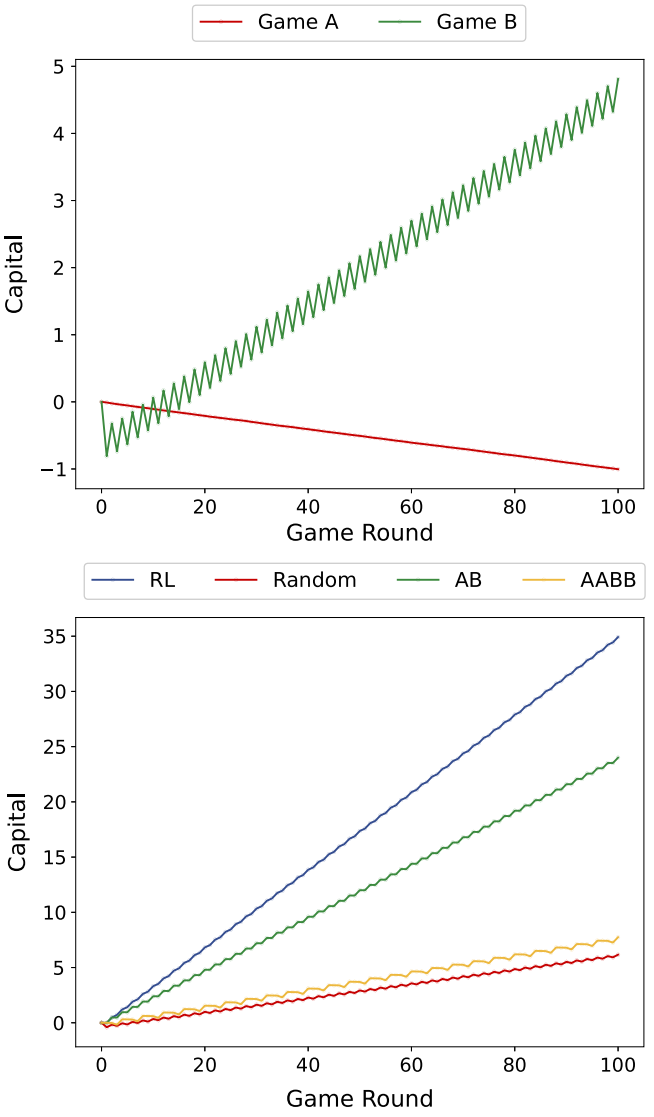

FIG. 5. Simulation results for the capital-dependent game ($M = 4$).

The extensibility of reinforcement learning empowers us to fine tune the algorithm to suit specific game parameters, facilitating its adaptability to different values of M and enabling the exploration of more complex variations of Parrondo's games. In practice, we only need to change the size of the Q-table and make the number of states equal to M, then the optimal sequence under different settings can be easily found. As observed in Fig. 4, reinforcement learning can still find a much better sequence than random sequence such as in the previous case of $M = 3$, revealing the effectiveness of our proposed method. Other evidence of the adaptability of our proposed method is shown in Fig. 5 where the case of $M = 4$ is tested (and no longer exhibiting the Parrondo effect). In this setting, game B is no longer a losing game. Then, the main focus should be shifted to maximizing the profit instead of turning lose to win, as in traditional Parrondo's games. As seen, the reinforcement-learning method can still achieve the dominant strength over the random sequence and periodic sequence of AB and AABB (included as a form of comparison).

In the history-dependent Parrondo's paradox, paradoxical effects are also exhibited, as shown in Fig. 6. Here, game B involves four coins, each with probabilities of winning of $0.9 - \epsilon$, $0.25 - \epsilon$, $0.25 - \epsilon$, and $0.7 - \epsilon$, respectively. By modifying the state space, the optimal strategy can also be learned and our proposed method can gain a capital of up to 25, much higher than the random sequence. As observed from the update process of the Q-table in Fig. 7, the best action in each state is well determined after a certain episodew of simulations, which suggests our method can easily converge to the optimal strategy. By leveraging the insights from Parrondo's paradox, a reinforcement-learning-based game can help optimize decision making and achieve better results in certain real-world scenarios where switching between different strategies based on certain outcomes is required. For example, in the context of the COVID-19 pandemic shown in Ref. [21], reinforcement learning can help determine the optimal switching strategy between lockdown and open community based on the infection numbers per day. The RL agent can learn from past data and adjust the switching decisions to minimize the loss caused by the pandemic. In an agricultural optimization example [22], Parrondo's paradox is applied

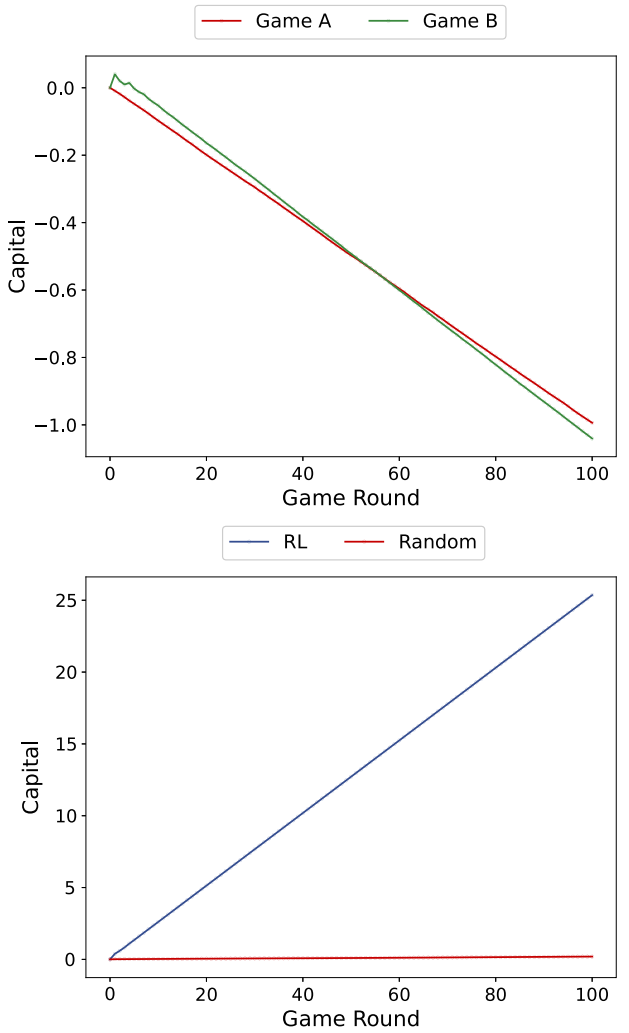

FIG. 6. Simulation results for the history-dependent game averaged over four possible initial states.

to crop rotations, alternating between cover and cash crops to mitigate intensive farming damage. The adaptable nature of reinforcement learning could enhance the yield further. In this context, the agent evaluates the farm's current state, considering factors such as soil quality, weather, and pest presence, then selects the next crop to plant, influencing future conditions and yield. The experimental results of different settings of the capital-dependent game and the history-dependent game also highlight the applicability and extensibility of our method. In studies concerning the modification of actions contingent upon the system state, the underlying model that governs the switch timing often remains intuitive and fails to reach an optimal level. These models, while demonstrating some efficacy, do not fully capitalize on the potential for decision-making enhancement, thus revealing an area of improvement by our proposed method discussed here. As demonstrated by the coin-tossing game, the switching strategy is learnable through reinforcement learning, suggesting its effectiveness.

In conclusion, our investigation was primarily concentrated on optimizing play sequences within two inherently losing games via a departure from the conventional methodologies of periodic or random sequences. This approach is well adapted to dynamically changing environments, a crucial aspect for applying Parrondo's paradox in real-world scenarios. We have implemented an algorithmic strategy aimed at learning and determining the most advantageous switching strategy. This strategy was pivotal in identifying the optimal sequence of play under varying conditions, a task that posed significant analytical challenges. Our numerical experiments have demonstrated the efficacy of our proposed model. Not only was our model capable of learning and applying the optimal switching strategy, but it also significantly outperformed the traditional approaches in these complex, dynamic scenarios. This advancement underscores the potential of adaptive strategies in game-theoretic paradigms and decision-making processes, particularly in environments characterized by uncertainty and fluctuation. Transitioning the proposed framework from simulation to real-world applications offers a valuable opportunity to tackle challenges such as computational requirements, scalability, and system integration, which will inspire future work. Our research will also be extended to the broader field of engineering, as well as economics,

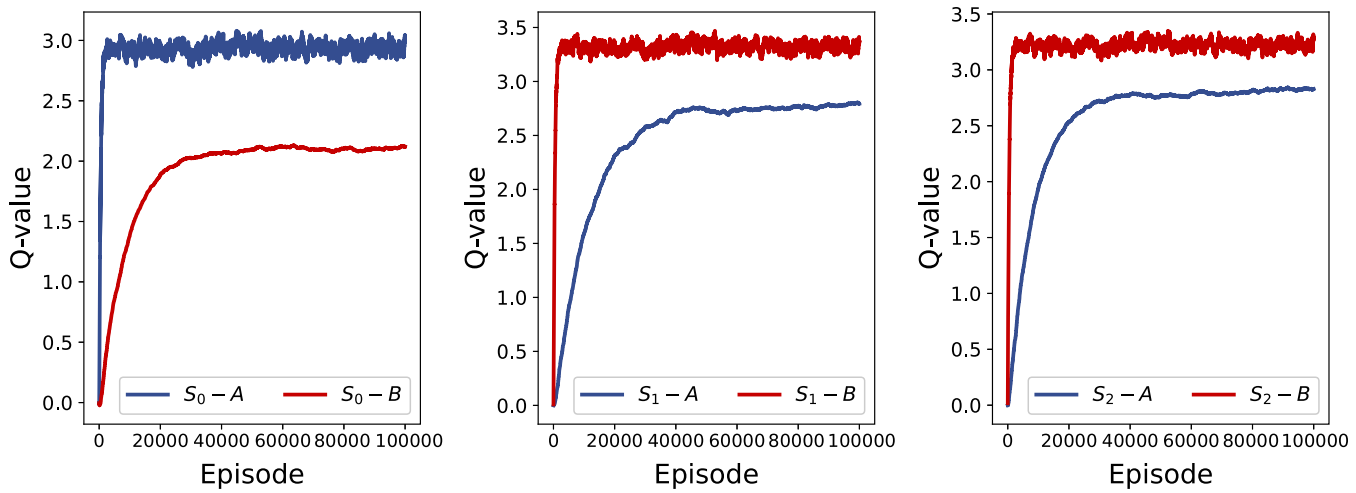

FIG. 7. The convergence test of capital-dependent game when $M = 3$.

statistical physics, and complex systems, by moving beyond the standard paradigms and incorporating elements of dynamic adaptability and environmental responsiveness.

This work was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 2 Grant No. MOET2EP50120-0021.

[1] J. Rajendran and C. Benjamin, Europhys. Lett. **122**, 40004 (2018).
[2] J. Rajendran and C. Benjamin, R. Soc. Open Sci. **5**, 171599 (2018).
[3] K. H. Cheong, J. M. Koh, and M. C. Jones, BioEssays **41**, 1900027 (2019).
[4] T. Wen, K. H. Cheong, J. W. Lai, J. M. Koh, and E. V. Koonin, Phys. Rev. Lett. **128**, 218101 (2022).
[5] J. W. Lai and K. H. Cheong, Phys. Rev. Res. **3**, L022019 (2021).
[6] G. P. Harmer and D. Abbott, Nature (London) **402**, 864 (1999).
[7] L. Dinis and J. M. Parrondo, Europhys. Lett. **63**, 319 (2003).
[8] L. Dinis, Phys. Rev. E **77**, 021124 (2008).
[9] D. Wu and K. Y. Szeto, Applications of genetic algorithm on optimal sequence for Parrondo games, in *International Conference on Evolutionary Computation Theory and Applications* (SCITEPRESS, 2014), Vol. 2.
[10] D. Wu and K. Y. Szeto, Sequence analysis with motif-preserving genetic algorithm for iterated parrondo games, in *Computational Intelligence: International Joint Conference* (Springer, Berlin, 2016), pp. 33–48.
[11] J. Zhao and K. H. Cheong, IEEE Trans. Evol. Comput. **27**, 1926 (2023).
[12] J. Zhao, Z. Wang, J. Cao, and K. H. Cheong, IEEE Trans. Syst., Man, Cybern. Syst. **53**, 4954 (2023).
[13] K. W. Cheung, H. F. Ma, D. Wu, G. C. Lui, and K. Y. Szeto, J. Stat. Mech.: Theory Exp. (2016) 054042.
[14] S. Rahmann, Optimal adaptive strategies for games of the Parrondo type, Science Direct Working Paper No. S1574-0358 (2002) 04.
[15] X. Molinero and C. Mégnien, arXiv:2304.05876.
[16] D. Wu and K. Y. Szeto, Phys. Rev. E **89**, 022142 (2014).
[17] C. J. C. H. Watkins and P. Dayan, Mach. Learn. **8**, 279 (1992).
[18] F. S. Melo, Tech. Rep. (2001).
[19] J. W. Lai and K. H. Cheong, Chaos **32**, 103107 (2022).
[20] K. H. Cheong, Z. X. Tan, N.-g. Xie, and M. C. Jones, Sci. Rep. **6**, 34889 (2016).
[21] K. H. Cheong, T. Wen, and J. W. Lai, Adv. Sci. **7**, 2002324 (2020).
[22] C. S. Gokhale and N. Sharma, R. Soc. Open Sci. **10**, 221401 (2023).