Taking Myopic Best Response Against The Hedge Algorithm

Xinxiang Guo[1,2], Yifen Mu[3]

1. School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

2. The State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

3. The Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

Abstract: With the rapid development of artificial intelligence (AI), the game between the human and machine/AI becomes more and more common. The related theoretical analysis becomes significant and necessary, which however is still rare. This problem involves evolution analysis and optimal control of dynamic game systems driven by learning algorithms. In this paper, we consider a finitely repeated two-player zero-sum game. We study the dynamic evolution process of the game system, in which one player employs the Hedge algorithm and the other player takes the myopic best response. By the updating formula of the Hedge algorithm, we define a quantity called the state and construct the State Transition Triangle Graph (STTG). Then, we prove that the game system is periodic after $o(T)$ stages.

Key Words: Dynamic games, Opponent exploitation, No-regret learning, Artificial intelligence

1 Introduction

In recent years, artificial intelligence (AI) has made rapid developments, especially considering AI algorithms to play games. Many AI algorithms have been proposed and achieved human-level performance in different kinds of games, some of which even can defeat human experts. This makes the basic question arise naturally: how should we play games against an opponent who adopts an AI algorithm? The answer to this problem is necessary to understand the system involving AI and humans. However, theoretical analysis of this problem is rare in the literature.

The AI algorithms to play games have been studied for a long time and compelling achievements have been made in the past decade, which is shown by their performance against humans. AlphaGo defeated the world champion in the game of Go, which was regarded as impossible due to the high complexity of the game [17]. For the imperfect information games, DeepStack [16] and Libratus [2], based on some variants of counterfactual regret minimization [21], beat professional poker players in heads-up no-limit Texas hold 'em poker with statistical significance. For six-player no-limit Texas hold 'em poker, Pluribus [4] is capable of defeating elite human professionals. For more complicated imperfect-information games, AlphaStar was rated at Grandmaster level for all three StarCraft races and above 99.8% of officially ranked human players [20]. Although these algorithms perform well, it is still open whether there exist certain strategies for humans to exploit these algorithms and win the games.

To play against these algorithms, one way is to take Nash equilibrium strategy. However, on the one hand, finding a Nash equilibrium in a two-player game is PPAD-complete [6], that is to say, it is usually computationally costly to obtain a Nash equilibrium in a complex game. On the other hand, taking Nash equilibrium strategy, which may lead the potential benefit of finding and exploiting the weakness of opponent to be missed [10], is conservative. For example, in the game paper-rock-scissors, if the strategy of the opponent is to play paper, rock, scissor in order and in turn, it is obvious that taking Nash equilibrium strategy forgoes the underlying benefit of exploiting such sub-optimal opponent. Therefore, considering how to (maximally) exploit opponents is desirable and profitable.

In fact, the problems of opponent exploitation could be viewed as games between a human and a machine/AI. Concerning the optimal play against a machine/AI. Mu and Guo [14] proved and gave the optimal strategy against an opponent with a simple finite-memory strategy which leads the system to cycles, Tang et al. [19] utilized recurrent neural networks (RNNs) to approximate the opponent and utilized the look-ahead strategy (called Rolling Horizon Evolution Algorithm) to optimize his utility, which won the 2020 Fighting Game AI Competition. Concerning the revenue against a machine/AI, Braverman et al. [3] studied the auction, in which the buyer runs a no-regret learning algorithm, and constructed the bounds for the seller's revenue. Deng et al. [7] also built the bounds for the optimizer against the no-regret learner for different cases. Besides, Ganzfried et al. [9] considered the safe exploitation under the guarantee of the game value in finitely repeated games and developed efficient algorithms for exploiting sub-optimal opponents safely. However, few of these research studied the dynamic of the game system.

The dynamic of the game system is complex to analyze. To take the first step, we study the evolution of the game ststem when one player employs the Hedge algorithm and the other player takes the myopic best response. The Hedge [13] algorithm is a no-regret learning algorithm which is a well-known class of adaptive learning algorithms and has many applications in repeated interactions (e.g. social choice [12]). Hedge possesses many better properties than general no-regret learning algorithms such as regret matching[11]. For example, for the problem of prediction with expert advice, Hedge only depends on the past performance of the experts, whereas the predictions made using other no-regret

This work was supported by the Major Project on the New Generation of Artificial Intelligence from the Ministry of Science and Technology of China [Grant No.2018AAA0101002] and the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDA27000000..

learning algorithms depend also on the past predictions [5]. Moreover, there is a parameter η in the setting of Hedge algorithm, which makes Hedge usually perform better than other no-regret learning algorithms [1]. Recently, Mourtada and Gaïffas [15] studied Hedge algorithm in the online stochastic setting and proved its optimality in some sense. Song et al. [18] proposed a method for online visual-object tracking problem based on Hedge algorithm. Due to the superior properties of Hedge, we will study the optimal play against Hedge, which can be regarded as an appropriate start for exploiting more complicated algorithms.

In this paper, we consider a 2×2 zero-sum game repeated for T stages and assume that one player (player X) employs the Hedge algorithm to update his stage strategy, and the other player (player Y) knows this and takes the myopic best response to the stage strategy of player X.

From the updating formula of Hedge, we find that the stage strategy of player X is totally determined by a quantity called *state* and deduce the updating formula with respect to this quantity. Based on this, we construct a graph, called *the State Transition Triangle Graph* (STTG). Then, we investigate the evolution of the system when player Y adopts the myopic best response to update his action at each stage. We prove that the state sequence of the system and the action sequence of player Y will enter a cycle after $o(T)$ stages. For certain special games, the system is periodic over the whole time.

This paper is organized as below: Section 2 gives the problem formulation; Section 3 introduces the deduction of State Transition Triangle Graph; Section 4 studies the game system when player Y takes myopic best response; Section 5 concludes the paper and talks about future work.

2 Problem Formulation

Consider a 2×2 zero-sum game. To be specific, there are two players in the game, called player X and player Y. Player X has two actions: 1 and 2, while player Y also has two actions: L and R. Given any action profile, player X and Y obtain their individual losses, which are represented by the following matrix,

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \tag{1}$$

In the matrix, a_{11} is the loss of player X given the action profile $(1, L)$. Since the game is zero-sum, a_{11} is the payoff of player Y given the action profile $(1, L)$. For the other profiles $(1, R), (2, L), (2, R)$, the loss of player X is a_{12}, a_{21}, a_{22} respectively.

In this paper, we assume that the values of all the losses are taken to be integers, i.e., $a_{ij} \in \mathbb{Z},\ \forall\, i, j = 1, 2$. Let

$$\delta_1 = a_{11} - a_{12},\ \delta_2 = a_{21} - a_{22},$$
$$\Delta_1 = a_{11} - a_{21},\ \Delta_2 = a_{12} - a_{22}.$$

Sometimes, player X and Y would take mixed strategies, denoted by $x = (x_1, x_2)^T$ and $y = (y_1, y_2)^T$, where $x_1 + x_2 = 1, x_1 \geq 0, x_2 \geq 0$ and $y_1 + y_2 = 1, y_1 \geq 0, y_2 \geq 0$. Then, the expected loss of player X (i.e. the expected payoff of player Y) is $x^T A y$.

Now, let the game be repeated for T times. The stage strategy of player X and player Y at time t is denoted by $x_t = (x_{1,t}, x_{2,t})^T$ and $y_t = (y_{1,t}, y_{2,t})^T$ respectively. At each time t, player Y would obtain *the instantaneous expected payoff* (IEP) $x_t^T A y_t$. After T stages, *the cumulative expected payoff* (CEP) of player Y is $\sum_{t=1}^{T} x_t^T A y_t$.

To play the game, player X could observe the action taken by player Y and follow the Hedge algorithm to update its stage strategy x_t. To be specific, $x_1 = (\frac{1}{2}, \frac{1}{2})$ and at each time $t \geq 2$, based on his previous observations of actions taken by player Y, player X will compute *the regret* $R_{i,t}$ with respect to action $i = 1, 2$,

$$R_{i,t} = \sum_{\tau=1}^{t-1} (x_\tau^T A y_\tau - e_i^T A y_\tau).$$

Then, player X's mixed strategy at time $t \geq 2$ is given by

$$x_{i,t} = \frac{\exp(\eta R_{i,t})}{\sum_{j=1}^{2} \exp(\eta R_{j,t})}, \quad i = 1, 2,$$

which equals to

$$x_{i,t} = \frac{\exp(-\eta \sum_{\tau=1}^{t-1} e_i^T A y_\tau)}{\sum_{j=1}^{2} \exp(-\eta \sum_{\tau=1}^{t-1} e_j^T A y_\tau)},\ i = 1, 2. \tag{2}$$

In the above formula, η is a positive parameter satisfying that $\eta = O(\sqrt{1/T})$, which is regarded as constant. Usually, it is taken to be $\sqrt{8 \ln 2 / T}$, which is optimal in the sense of regret bound [5].

We assume that player Y knows that player X employs the Hedge algorithm, that is to say, player Y could accurately predict the stage strategy of player X at each stage. It is natural for player Y takes the myopic best response to the stage strategy of player X. Then, the basic question arises: how will the game system behave when player X employs the Hedge algorithm and player Y takes the myopic best response.

In this paper, we assume that there does not exist a dominant strategy for players X and Y, i.e., $\Delta_1 \Delta_2 < 0, \delta_1 \delta_2 < 0$. If player Y has a dominant strategy, it is easy to see that his myopic best response to the stage strategy of player X would always be the dominant strategy. Therefore, the dynamic of the game system in this case is obvious. Without loss of generality, we assume that $\delta_1 > 0, \delta_2 < 0, \Delta_1 > 0, \Delta_2 < 0$ and $|\Delta_1| \leq |\Delta_2|$.

3 Method: State Transition Triangle Graph

In this section, we construct a graph, on which the state transition process and IEP of player Y is vividly shown.

In the updating rule of Hedge (2), we notice that $x_{1,t}$ and $x_{2,t}$ have the same denominator. So, for time $t \geq 2$, we have

$$\begin{aligned} \frac{x_{2,t}}{x_{1,t}} &= \frac{\exp(-\eta \sum_{\tau=1}^{t-1} (a_{21}, a_{22}) y_\tau)}{\exp(-\eta \sum_{\tau=1}^{t-1} (a_{11}, a_{12}) y_\tau)} \\ &= \exp(-\eta(-\Delta_1, -\Delta_2) \sum_{\tau=1}^{t-1} y_\tau), \end{aligned} \tag{3}$$

where $\Delta_1 = a_{11} - a_{21}, \Delta_2 = a_{12} - a_{22}$.

Define

$$s_t = (-\Delta_1, -\Delta_2) \sum_{\tau=1}^{t-1} y_\tau, \tag{4}$$

for $t \geq 2$. Then we have

$$\frac{x_{2,t}}{x_{1,t}} = \exp(-\eta s_t). \tag{5}$$

Since $x_{1,t} + x_{2,t} = 1$, we have

$$x_t = \left(\frac{1}{e^{-\eta s_t} + 1}, \frac{e^{-\eta s_t}}{e^{-\eta s_t} + 1} \right), \tag{6}$$

for all time $t \geq 2$.

Moreover, when $t = 1$, we have taken x_1 to be $(\frac{1}{2}, \frac{1}{2})$. Hence, by (5), we have $s_1 = 0$, which replenishes the definition (4) of s_t. Then, stage strategy x_t of player X has a one-to-one correspondence with s_t for all time $t \geq 1$ and s_t is called *the state of the system* or simply *state* at time t.

By (4), we have the iteration formula

$$s_{t+1} = s_t + (-\Delta_1, -\Delta_2) y_t. \tag{7}$$

For convenience, we define a state transition function

$$h(s, y) = s + (-\Delta_1, -\Delta_2) y, \tag{8}$$

then $s_{t+1} = h(s_t, y_t)$. Since either $y_t = e_L$ or $y_t = e_R$, we can rewrite (7) as

$$s_{t+1} = \begin{cases} s_t - \Delta_1, & \text{if } y_t = e_L; \\ s_t - \Delta_2, & \text{if } y_t = e_R. \end{cases} \tag{9}$$

From the definition (4) of s_t, we know that there are just t different possible values for state s_t at time t, denoted by $s_{i,t}, 1 \leq i \leq t$, one by one from small to large. Furthermore, by (9), $s_{i,t} = -(t-i)\Delta_1 - (i-1)\Delta_2, 1 \leq i \leq t$.

Given that $s_{t-1} = s_{i,t-1}$, formula (7) can further be written of form

$$s_t = \begin{cases} s_{i,t}, & \text{if } y_{t-1} = e_L; \\ s_{i+1,t}, & \text{if } y_{t-1} = e_R. \end{cases} \tag{10}$$

At each time t, after player X and Y take their strategies, player Y would obtain IEP $x_t^T A y_t$. Since strategy x_t is corresponding to s_t one to one, IEP of player Y at time t can be given by

$$r(s_t, y_t) \triangleq \left(\frac{1}{e^{-\eta s_t} + 1}, \frac{e^{-\eta s_t}}{e^{-\eta s_t} + 1} \right) A y_t. \tag{11}$$

By (7), state s_{t+1} depends only on state s_t of the system and action y_t of player Y. By (11), the payoff obtained by player Y at time t also depends only on state s_t of the system and action y_t of player Y. We can illustrate the state transition process by a graph in Figure 1, which is called *the State Transition Triangle Graph* (STTG).

4 Periodicity of the Hedge-Myopic System

In this section, we study the evolution the game system when player X employs the Hedge algorithm and player Y takes the myopic best response.

Since player Y knows that player X follows Hedge to update his stage strategy x_t, then player Y can know x_t exactly based on the updating rule (2) of Hedge before he chooses

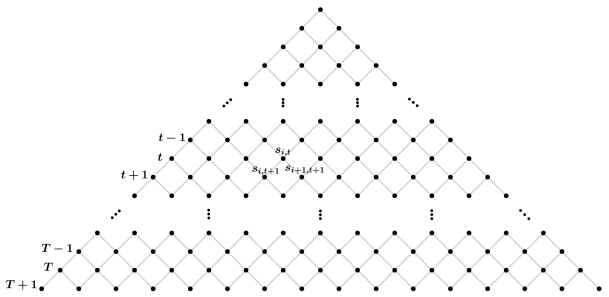

Fig. 1: The State Transition Triangle Graph. In STTG, each black point (also called node) represents a state and nodes at the t-th row represent all possible states at time t: $s_{1,t}, s_{2,t}, \cdots, s_{t,t}$. Each light gray line segment connecting two nodes represents an action of player Y. From state $s_{i,t}$, player Y can choose action e_L(the left branch), leading to state $s_{i,t+1}$, or chooses action e_R(the right branch), leading to state $s_{i+1,t+1}$. Although the game is played for T times, states of the system at time $T+1$ are also considered.

y_t. In this section, we assume that player Y takes *the myopic best response* (MBR) according to x_t, i.e.,

$$y_t = \arg\max_{y \in \mathcal{Y}} x_t^T A y,$$

which equals to

$$y_t = \arg\max_{y \in \mathcal{Y}} r(s_t, y) \triangleq \mathrm{MBR}(s_t). \tag{12}$$

By (11), we have

$$r(s, e_R) - r(s, e_L) = -\frac{\delta_1 + \delta_2 e^{-\eta s}}{1 + e^{-\eta s}}.$$

Let $r(s, e_R) - r(s, e_L) = 0$, we have $s = -\frac{1}{\eta} \ln\left(-\frac{\delta_1}{\delta_2}\right)$.

Define

$$s^*(T) = -\frac{1}{\eta} \ln\left(-\frac{\delta_1}{\delta_2}\right). \tag{13}$$

We can see that: (i) $s^*(T) = O(\sqrt{T}) = o(T)$; (ii) since $\delta_1 > 0, \delta_2 < 0$, if $s \geq s^*(T)$, then $r(s, e_R) \leq r(s, e_L)$ while if $s < s^*(T)$, then $r(s, e_R) > r(s, e_L)$. Therefore, we can write the formula (12) as

$$\mathrm{MBR}(s) = \begin{cases} e_R, & \text{if } s < s^*(T); \\ e_L, & \text{if } s \geq s^*(T). \end{cases} \tag{14}$$

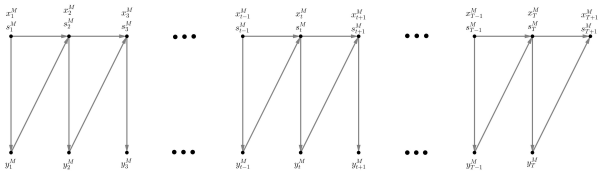

Fig. 2: The evolution of the system: Hedge VS. MBR. At each time t, state s_t^M of the system has one-to-one correspondence to stage strategy x_t^M of player X and action y_t^M of player Y is totally determined by x_t^M. Then, by (7), state s_{t+1}^M can be obtained by s_t^M and y_t^M. Further, stage strategy x_{t+1}^M can then be obtained by (6).

Now, in the repeated game, player X uses Hedge while player Y takes MBR. The evolution process of the system can be illustrated by Figure 2. In this system, we denote the action sequence of player Y by $y_1^M, y_2^M \cdots, y_T^M$, the strategy sequence of player X by $x_1^M, x_2^M, \cdots, x_{T+1}^M$ and the state sequence of the system by $s_1^M, s_2^M, \cdots, s_{T+1}^M$. Moreover, sequence $s_1^M, s_2^M, \cdots, s_{T+1}^M$ is called *the myopic sequence*, and sequence $s_1^M, y_1^M, s_2^M, y_2^M, \cdots, s_T^M, y_T^M, s_{T+1}^M$ is called *the myopic path*.

Let $m(p,q)$ to be the least common multiple of two positive integers p and q. Then, we have

Theorem 1. *In this system, after T_s stages where $T_s = o(T)$, the action sequence $y_{T_s}^M, \cdots, y_T^M$ of player Y and the state sequence $s_{T_s}^M, \cdots, s_{T+1}^M$ are both periodic and they have the same least positive period, which is $T^* = \frac{m(|\Delta_1|,|\Delta_2|)}{|\Delta_1|} + \frac{m(|\Delta_1|,|\Delta_2|)}{|\Delta_2|}$.*

Next, we prove Theorem 1 in several steps.

Firstly, suppose $p, q \in \mathbb{N}$ and $0 < p < q$. Given any $\lambda \in \mathbb{R}$, it is easy to see that for any $\gamma \in \mathbb{R}$, there always exists a unique integer k such that $\gamma + kp \in [\lambda - p, \lambda)$.

Now, consider a sequence $\{\gamma_i\}_{i=1}^{\infty}$, which is iteratively generated by

$$\gamma_{i+1} = \gamma_i + q + k_i p, \tag{15}$$

where k_i is the unique integer such that $\gamma_{i+1} \in [\lambda - p, \lambda)$. Take $\gamma_1 \in [\lambda - p, \lambda)$, then $\gamma_i \in [\lambda - p, \lambda)$ for all $i \geq 1$ and we have

Lemma 2. *The sequence $\{\gamma_i\}_{i=1}^{\infty}$ defined above is periodic and its least positive period is $\frac{m(p,q)}{q}$.*

Proof. By the generating rule of γ_i, we have

$$\gamma_{i+i_0} = \gamma_i + i_0 q + \sum_{\tau=i}^{i+i_0-1} k_\tau p.$$

Since $\gamma_i \in [\lambda - p, \lambda)$ for all $i \geq 1$, $\gamma_{i+i_0} = \gamma_i$ if and only if

$$i_0 q + \sum_{\tau=i}^{i+i_0-1} k_\tau p \equiv 0 \mod p.$$

This is further equivalent to $i_0 q \equiv 0 \mod p$. Obviously, $i_0 = \frac{m(p,q)}{q}$ is the least positive integer satisfying this equation. That proves the lemma.

□

Below, we will state some lemmas without proofs, which can be seen in our full version paper [22].

Lemma 3. *In the state sequence $s_1^M, s_2^M, \cdots, s_{T+1}^M$, there exists time $T_c = o(T)$ such that $s_{T_c}^M < s^*(T), s_{T_c+1}^M \geq s^*(T)$.*

From Lemma 3, we denote the least T_c to be t_0, then $t_0 = o(T)$. Define a time sequence

$$t_{i+1} = \min\{t : t_i < t \leq T, s_t^M < s^*(T),\ s_{t+1}^M > s^*(T)\} \tag{16}$$

for $i \geq 0$. If the set $\{t | s_t^M < s^*(T), s_{t+1}^M > s^*(T), t_i < t \leq T\}$ is empty for some i, we denote such i by $i_{\max}$.

Now, we consider the obtained finite state sequence $s_{t_0}^M, s_{t_1}^M, \cdots, s_{t_{i_{\max}}}^M$.

Lemma 4. *For sequence $s_{t_1}^M, s_{t_2}^M, \cdots, s_{t_{i_{\max}}}^M$, we have*

$$s_{t_i+T^*}^M = s_{t_i}^M$$

for $1 \leq i \leq i_{\max}$, $t_i + T^* \leq T + 1$ where $T^* = \frac{m(|\Delta_1|,|\Delta_2|)}{|\Delta_1|} + \frac{m(|\Delta_1|,|\Delta_2|)}{|\Delta_2|}$ as defined in Theorem 1.

Then, we can prove Theorem 1.

Proof. Consider the state sequence $s_{t_1}^M, s_{t_1+1}^M, \cdots, s_{T+1}^M$.

By Lemma 4, $s_{t_1+T^*} = s_{t_1}$. By the myopic best response formula (14), we have $\mathrm{MBR}(s_{t_1+T^*}) = \mathrm{MBR}(s_{t_1})$ and then by the state updating formula (7), we have $s_{t_1+1+T^*} = s_{t_1+1}$. Repeatedly use (14) and (7), we can prove that sequence $s_{t_1}^M, s_{t_1+1}^M, \cdots, s_{T+1}^M$ is periodic. Since $t_0 = o(T)$, then $t_1 = o(T)$. Take $T_s = t_1$ and we prove Theorem 1. □

In the above results, Lemma 4 is the key tool to prove Theorem 1 while Lemma 2 is the key step to prove Lemma 4. In the proof of Lemma 4, we obtain a sequence $s_{t_0}^M, s_{t_1}^M, \cdots, s_{t_{i^{\max}}}^M$ where $s_{t_j}^M < s^*(T)$ for all j, $0 \leq j \leq i^{\max}$ and find that its subsequence $s_{t_1}^M, s_{t_2}^M, \cdots, s_{t_{i^{\max}}}^M$ is periodic, which leads to the periodicity of the almost whole sequence generated by the Hedge-Myopic system in Theorem 1.

Remark. From the proof of Lemma 4, we can see that there are $\frac{m(|\Delta_1|,|\Delta_2|)}{|\Delta_1|}$ action L and $\frac{m(|\Delta_1|,|\Delta_2|)}{|\Delta_2|}$ action R of player Y in one period of Hedge-Myopic system. Therefore, time-averaged strategy of player Y is $\left(\frac{|\Delta_2|}{|\Delta_2|+|\Delta_1|}, \frac{|\Delta_1|}{|\Delta_2|+|\Delta_1|}\right)$ which is Nash equilibrium strategy. However, player Y obtain higher average payoff than game value. This is because player Y takes MBR to the strategy of player X and his IEP is at least the game value.

We provide an example to illustrate the above results.

Example 4.1. Given that $T = 700$ and

$$A = \begin{pmatrix} 1 & 0 \\ -1 & 3 \end{pmatrix}.$$

By simple calculation, $\Delta_1 = 2, \Delta_2 = -3, \delta_1 = 1, \delta_2 = -4, s^*(T) \approx 15.58$. In this case, sequence $s_{t_0}^M, s_{t_1}^M, \cdots, s_{t_{i_{\max}}}^M$ is $s_6^M, s_9^M, s_{11}^M, s_{14}^M$ and from time $t = 9$, sequence $s_9^M, s_{10}^M, \cdots, s_{16}^M$ (i.e. 14, 17, 15, 18, 16, 14, 17, 15) is periodic and its least positive period is 5. In each period, the action sequence of player Y is R, L, L, R, L and thus time-averaged strategy of player Y over one period is $(0.6, 0.4)$. On the other hand, in one period, payoff sequence obtained by player Y is 0.6646, 0.6119, 0.6702, 0.6390, 0.6250 and thus time-averaged payoff of player Y is 0.6421, which is higher than game value 0.6 .

The State Transition Triangle and the myopic path for this example are illustrated by Figure 3.

Now, consider a special kind of game where the Nash equilibrium strategy of player X is $(\frac{1}{2}, \frac{1}{2})$. In this case, we have $\delta_1 = -\delta_2$ and $s^*(T) = -\frac{1}{\eta}\ln(-\frac{\delta_1}{\delta_2}) = 0$.

For this game, the state sequence is called *the zero sequence* and denoted by $s_1^0, s_2^0, \cdots, s_{T+1}^0$. The according action sequence of player Y is denoted by $y_1^0, y_2^0, \cdots, y_T^0$ and sequence $s_1^0, y_1^0, s_2^0, y_2^0, \cdots, s_T^0, y_T^0, s_{T+1}^0$ is called *the zero path*. Once Δ_1 and Δ_2 is given, the zero sequence and the zero path can be obtained.

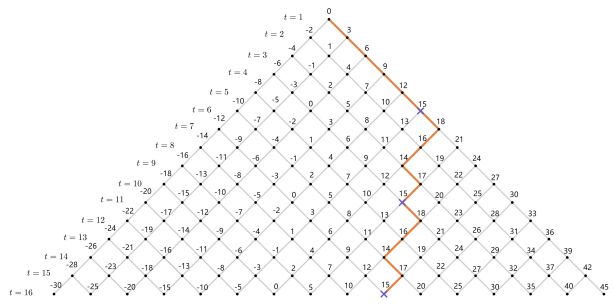

Fig. 3: STTG with the myopic path when $\Delta_1 = 2, \Delta_2 = -3, s^*(T) = 15.58$. Because of the limited space, we only draw the first 16 stages. In the graph, the number over a node represents the value of state s_t to which the node corresponds. And the orange polyline represents the myopic path.

Lemma 5. *For the zero sequence, we have*

(1) $s_1^0, s_2^0, \cdots, s_{T+1}^0$ *is periodic and its least positive period is* $T^* = \frac{m(|\Delta_1|,|\Delta_2|)}{|\Delta_2|} + \frac{m(|\Delta_1|,|\Delta_2|)}{|\Delta_1|}$.

(2) $-|\Delta_1| \leq s_t^0 < |\Delta_2|$.

(3) For possible value $s_{i,t}$ *of state at time t, if* $s_{i,t} < 0$ *and* $s_{i+1,t} \geq 0$*, then* $s_{t+1}^0 = s_{i+1,t+1}$.

5 Conclusion and Future Work

In this paper, we study the dynamic evolution of the game system when player X employs the Hedge algorithm and player Y takes the myopic best response in a two-player finitely-repeated 2×2 zero-sum game. We define the state and construct the State Transition Triangle Graph. Then, we prove that the myopic play will enter a cycle after few stages. Specifically, the zero path shows periodic behavior throughout the whole time. Based on the results in this paper, we can get the optimal play of player Y. The complete solution can be seen in our full paper [22].

For future work, one can study the game system when player X employs the other algorithms, such as regret matching. Besides, one can also consider other game setting, such as the non-zero-sum game and infinitely repeated game.

References

[1] Burch N, *Time and Space: Why Imperfect Information Games Are Hard*, PhD thesis, Alberta University, Canada, 2018.

[2] Brown N, Sandholm T, Superhuman AI for heads-up no-limit poker: Libratus beats top professionals, *Science*, 2018, 359(6374): 418-424.

[3] Braverman M, Mao J, Schneider J, et al, Selling to a no-regret buyer, *Proceedings of the 2018 ACM Conference on Economics and Computation*, 2018: 523-538.

[4] Brown N, Sandholm T, Superhuman AI for multiplayer poker, *Science*, 2019, 365(6456): 885-890.

[5] Cesa-Bianchi N, Lugosi G, *Prediction, Learning, and Games*, Cambridge University Press, 2006.

[6] Chen X, Deng X. Settling the complexity of two-player Nash equilibrium, *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*., 2006: 261-272.

[7] Deng Y, Schneider J, Sivan B, Strategizing against no-regret learners, arXiv:1909.13861, 2019.

[8] Fudenberg D, Levine D K, *The Theory of Learning in Games*, MIT Press, 1998.

[9] Ganzfried S, Sandholm T, Safe opponent exploitation, *ACM Transactions on Economics and Computation (TEAC)*, 2015, 3(2): 1-28.

[10] Ganzfried S, Sun Q. Bayesian opponent exploitation in imperfect-information games, *IEEE Conference on Computational Intelligence and Games (CIG)*. 2018: 1-8.

[11] Hart S, Mas-Colell A, A simple adaptive procedure leading to correlated equilibrium, *Econometrica*, 2000, 68(5): 1127-1150.

[12] Haghtalab N, Noothigattu R, Procaccia A D, Weighted voting via no-regret learning, *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[13] Littlestone N, Warmuth M K, The weighted majority algorithm, *Information and computation*, 1994, 108(2): 212-261.

[14] Mu Y, Guo L, Towards a theory of game-based non-equilibrium control systems, *Journal of Systems Science and Complexity*, 2012, 25(2): 209-226.

[15] Mourtada J, Gaïffas S, On the optimality of the Hedge algorithm in the stochastic regime, *Journal of Machine Learning Research*, 2019, 20: 1-28.

[16] Moravčík M, Schmid M, Burch N, et al, Deepstack: Expert-level artificial intelligence in heads-up no-limit poker, *Science*, 2017, 356(6337): 508-513.

[17] Silver D, Schrittwieser J, Simonyan K, et al, Mastering the game of go without human knowledge, *Nature*, 2017, 550(7676): 354-359.

[18] Song H, Suehiro D, Uchida S, Adaptive aggregation of arbitrary online trackers with a regret bound, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020: 681-689.

[19] Tang Z, Zhu Y, Zhao D, et al, Enhanced Rolling Horizon Evolution Algorithm with Opponent Model Learning, *IEEE Transactions on Games*, 2020, doi: 10.1109/TG.2020.3022698.

[20] Vinyals O, Babuschkin I, Czarnecki W M, et al, Grandmaster level in StarCraft II using multi-agent reinforcement learning, *Nature*, 2019, 575(7782): 350-354.

[21] Zinkevich M, Johanson M, Bowling M, et al, Regret minimization in games with incomplete information, *Advances in neural information processing systems*, 2007, 20: 1729-1736.

[22] Guo X, Mu Y, The optimal play against Hedge algorithm in finitely repeated two-player zero-sum games, submitted.