DTHBR: An Asymmetric NE-solving Paradigm with Last-iterate Convergence for Two-player Zero-sum Games

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

1. School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

2. State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

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

Abstract: Most traditional learning algorithms, such as no-regret algorithms, fail to converge to Nash equilibrium (NE) and may exhibit unpredictable behavior even in simple games like Matching Pennies. Consequently, researchers have focused on developing learning algorithms with last-iterate convergence. Most of this work has centered on optimistic variants of existing algorithms, including Hedge, Gradient Descent Ascent, and Follow the Regularized Leader. A critical requirement for achieving last-iterate convergence and a linear convergence rate is that all players must adopt optimistic variants. However, this assumption often fails in practical scenarios, as opponents may not cooperate. In this paper, we propose a novel asymmetric NE-solving paradigm with last-iterate convergence for two-player zero-sum games. Specifically, we allow one player to update their strategy using the Hedge algorithm while the other player adopts a corresponding best response. By incorporating a doubling trick into the learning rate of the Hedge algorithm, we prove that this paradigm achieves last-iterate convergence and provide a theoretical convergence rate. This approach offers a novel framework for leveraging learning algorithms to solve for Nash equilibrium with guaranteed last-iterate convergence.

Key Words: Last-iterate convergence, Asymmetric paradigm, Doubling trick, Hedge algorithm

1 Introduction

Game theory provides a powerful framework for modeling interactions among self-interested and rational agents in real-world scenarios [22]. At its core lies the concept of Nash equilibrium (NE) [21], which describes a stable state where no player can benefit from unilaterally changing their strategy, given that others adhere to the NE strategies. In two-player zero-sum games, linear programming offers an efficient method to compute NE in polynomial time [16]. However, practical challenges such as scalability and imperfect information, make the study of two-player zero-sum games an ongoing area of research, attracting attention from diverse fields [18, 25].

To address the challenges of solving NE in complex scenarios, researchers have turned to learning-based approaches. Over the years, a rich body of literature has emerged, leading to significant achievements such as AlphaGo [30] and Deepstack [20]. For instance, the Fictitious Play (FP) algorithm [13] ensures that the empirical distribution of actions converges to NE in zero-sum games [19]. When each player employs a no-regret algorithm [7] to determine their stage strategy in repeated games, their time-averaged strategy profile converges to the coarse correlated equilibrium in general-sum games and to NE in two-player zero-sum games [15]. In imperfect-information extensive-form games, the Counterfactual Regret Minimization (CFR) algorithm has been shown to converge to NE in two-player zero-sum settings [34]. A common feature of these convergence results is that they can only guarantee time-averaged convergence.

However, time-averaged convergence does not imply that the stage strategies themselves converge, a property known as last-iterate convergence. This distinction is crucial because, in many practical applications such as training Generative Adversarial Networks (GANs), time-averaged convergence is difficult to implement, making last-iterate convergence more desirable [10]. Further investigation into learning dynamics reveals that even in simple game models, basic algorithms can exhibit highly complex behavior and may fail to converge [24, 26, 27, 29]. For example, Palaiopanos et al. [23] demonstrated that the Multiplicative Weights Update (MWU) algorithm can exhibit bifurcation in specific instances of 2×2 potential games. Bailey and Piliouras [4] showed that in two-player zero-sum games, MWU dynamics can deviate from equilibrium and converge toward the boundary. Mertikopoulos et al. [18] further proved the Poincaré recurrence of regularized learning algorithms in two-player zero-sum games, implying the impossibility of convergence to NE from any initial strategy profile. These findings highlight the complexity of learning dynamics in games and demonstrate that most learning algorithms fail to converge to NE.

Recent research has focused on developing learning algorithms with last-iterate convergence [3, 5]. Many of these works modify existing algorithms, which often exhibit chaotic behavior, by incorporating predictive methods to create optimistic variants. Examples include Optimistic Follow the Regularized Leader [31], Optimistic Multiplicative Weights Update [6], Optimistic Online Mirror Descent [28], and Optimistic Gradient Descent/Ascent [33]. To further enhance convergence rates [9], researchers have explored techniques such as entropy regularization [6] and the use of self-concordant barriers as the regularizer [31]. These advancements have been extended to various game settings, including multi-player general-sum matrix games [1], two-player zero-sum stochastic games [6], and multiplayer general-sum

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDA27030201 and the Natural Science Foundation of China under Grant T2293770.

stochastic games [11]. Despite these developments, most existing algorithms assume symmetry among players, where all participants employ the same learning dynamics.

Our work focuses on an asymmetric setting of learning dynamics in two-player zero-sum games, proposing a novel paradigm that achieves late-iterate convergence to NE. Specifically, we allow one player to update their strategy using the Hedge algorithm while the other player adopts a corresponding best response. We establish a theoretical convergence rate for this paradigm and demonstrate, through simulations, the effectiveness of integrating a doubling trick into the Hedge algorithm to attain last-iterate convergence. This paper offers a novel framework, which is different from most of existing literature, for modifying learning algorithms to achieve last-iterate convergence.

Paper Organization: Section 2 provides the necessary preliminaries, including the zero-sum game model and the Hedge algorithm. Section 3 introduces an NE-solving paradigm and proves its last-iterate convergence to the NE strategy with a sublinear convergence rate. Section 4 demonstrates the effectiveness of the proposed paradigm through simulations. Finally, Section 5 concludes the paper, discusses limitations of the proposed paradigm, and suggests interesting research problems for future work.

2 Preliminary

2.1 Zero-sum game

Consider a $n \times n$ two-player zero-sum game. The row player is called Player X and the column player is called Player Y. Their action sets are denoted by $\mathcal{I} = \{1, 2, \cdots, n\}$ and $\mathcal{J} = \{1, 2, \cdots, n\}$, respectively. Given any action profile $(i, j) \in \mathcal{I} \times \mathcal{J}$, the payoff obtained by Player X is a_{ij}. Thus, the payoff matrix of this game is $A = (a_{ij})_{i\in\mathcal{I}, j\in\mathcal{J}}$, which is the payoff matrix for Player X and the loss matrix for Player Y.

A mixed strategy of Player X, denoted by $\mathbf{x} \in \Delta(\mathcal{I})$, is a probability distribution over the action set of Player X, i.e., the strategy $\mathbf{x}$ satisfies that $x_i \geq 0, \forall i \in \mathcal{I}$ and $\sum_{i\in\mathcal{I}} x_i = 1$. A mixed strategy of Player Y, denoted by $\mathbf{y} \in \Delta(\mathcal{J})$, is a probability distribution over the action set of Player Y, i.e., the strategy $\mathbf{y}$ satisfies that $y_j \geq 0, \forall j \in \mathcal{J}$ and $\sum_{j\in\mathcal{J}} y_j = 1$. A strategy profile $(\mathbf{x}, \mathbf{y})$ is called a Nash Equilibrium (NE) if

$$\mathbf{x}^T A\mathbf{y} \geq (\mathbf{x}')^T A\mathbf{y}, \ \forall \mathbf{x}' \in \Delta(\mathcal{I});$$
$$\mathbf{x}^T A\mathbf{y} \leq \mathbf{x}^T A\mathbf{y}', \ \forall \mathbf{y}' \in \Delta(\mathcal{J}).$$

We denote such an NE by $(\mathbf{x}^*, \mathbf{y}^*)$. The value of the game is denoted by v^* and equals to $(\mathbf{x}^*)^T A\mathbf{y}^*$.

An NE is called an interior NE or a fully-mixed NE if the support sets of strategies $\mathbf{x}^*$ and $\mathbf{y}^*$ are the respective action set of Player X and Player Y, i.e., $\{i|x_i^* > 0\} = \mathcal{I}$ and $\{j|y_j^* > 0\} = \mathcal{J}$.

Let the zero-sum game repeated for an infinite number of stages. The stage strategy of Player X at time t is denoted by $\mathbf{x}_t$, and the stage strategy of Player Y at time t is denoted by $\mathbf{y}_t$.

2.2 Online learning and Hedge algorithm

The Hedge algorithm, also known as weighted majority [17] or exponential weighted average prediction [7], or Multiplicative Weights Update [2], is a popular no-regret learning algorithm proposed by Freund and Schapire, based on the context of boost learning [12].

Consider the online learning framework known as learning with expert advice [7]. In this framework, the decision maker is a forecaster whose goal is to predict an unknown sequence $d_1, d_2, \cdots$, where d_t belongs to an outcome space $\mathcal{D}$. The prediction of the forecaster at time t, denoted by $\hat{p}_t$, is assumed to belong to a convex subset $\mathcal{P}$ of $\mathcal{D}$. At each time t, the forecaster receives a finite set of expert advice $\{f_{i,t} \in \mathcal{P} : i = 1, 2, \cdots, N\}$, then the forecaster computes his own guess $\hat{p}_t$ based on the given set of expert advice. Subsequently the true outcome d_t is revealed. Predictions of the forecaster and experts are scored using a non-negative loss function $\ell : \mathcal{P} \times \mathcal{D} \to \mathbb{R}$ and the *cumulative regret* is introduced to measure how much better the forecaster could have done compared to how he did in hindsight, which is defined to be

$$R_t = \max_{i=1,2,\cdots,N} \left\{ \sum_{\tau=1}^{t} (\ell(\hat{p}_\tau, d_\tau) - \ell(f_{i,\tau}, d_\tau)) \right\}.$$

The prediction $\hat{p}_t$ at time t given by the Hedge algorithm is the weighted average of the predictions from the experts, i.e., $\hat{p}_t = \sum_{i=1}^{N} w_{i,t-1} f_{i,t}$, where

$$w_{i,t-1} = \frac{\exp(-\eta \sum_{\tau=1}^{t-1} \ell(f_{i,\tau}, d_\tau))}{\sum_{j=1}^{N} \exp(-\eta \sum_{\tau=1}^{t-1} \ell(f_{j,\tau}, d_\tau))}, \ t \geq 1,$$

where η is a positive parameter and is called *the learning rate*.

3 An Asymmetric NE-solving Paradigm with Last-iterate Convergence

In this section, we propose an asymmetric NE-solving paradigm for two-player zero-sum games and prove that this paradigm has late-iterate convergence to the NE strategy by applying doubling trick to the learning rate of the Hedge algorithm. More than this, we also give the convergence rate of this paradigm.

We state some assumptions below.

Assumption 1: The zero-sum game has a unique interior NE.

Assumption 2: The matrices $\{A_i \in \mathbb{R}^{n\times n}, i = 1, 2, \cdots, n\}$ are all non-singular, where A_i is defined as

$$\begin{pmatrix} a_{1,1} & a_{2,1} & \cdots & a_{n,1} \\ \vdots & & & \vdots \\ a_{1,i-1} & a_{2,i-1} & \cdots & a_{n,i-1} \\ a_{1,i+1} & a_{2,i+1} & \cdots & a_{n,i+1} \\ \vdots & & & \vdots \\ a_{1,n} & a_{2,n} & \cdots & a_{n,n} \\ 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{1}$$

The matrix A_i is obtained by suppressing the i-th row from A^T and adding a line of 1 at its bottom.

Assumption 1 is a wild and common condition on the game when studying the convergence of the Hedge-kind algorithm, like Optimistic Multiplicative Weights Update [8] and Hedge algorithm itself [4]. Assumptions 1 and 2 are not

unrelated for zero-sum $n \times n$ games with a non-zero value. Specifically, if (1) the value of the game is non-zero, and (2) there is an interior Nash equilibrium, and (3) Assumption 2 holds, then the interior NE is unique, as assumed in Assumption 1. To account for all cases, we require the NE to be both unique and interior in Assumption 1.

Now, we stipulate the action rules for both players.

Let Player X employ the Hedge algorithm to update their stage strategy, i.e.,

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

where $e_i = (0, \cdots, 1, \cdots, 0)$ with the i-th element being 1 and η is a sufficiently small constant parameter.

Let Player Y take a best response to the stage strategy of Player X at each stage, i.e.,

$$\mathbf{y}_t = \arg\min_{\mathbf{y} \in \Delta(\mathcal{J})} \mathbf{x}_t^T A \mathbf{y}. \tag{3}$$

The best response of Player Y may not be unique for some strategy $\mathbf{x}$. When it happens, Player Y can choose any best response, and this does not affect the last-iterate convergence to the NE strategy. This game system is called *the Hedge-myopic system*.

For this game system, we have the following theoretical result, which gives the convergence rate to NE strategy of Player X. In this theorem, we use the the Kullback-Leibler divergence (KL divergence) to measures the distance of the strategy $\mathbf{x}$ to the NE strategy $\mathbf{x}^*$, which is defined to be

$$D_{KL}(\mathbf{x}^* || \mathbf{x}) = -\sum_{i=1}^{n} x_i^* \ln x_i + \sum_{i=1}^{n} x_i^* \ln x_i^*. \tag{4}$$

The smaller the KL divergence $D_{KL}(\mathbf{x}^* || \mathbf{x})$ is, the closer the strategy $\mathbf{x}$ is to the NE strategy $\mathbf{x}^*$.

Theorem 1. *Consider a two-player zero-sum game satisfying Assumptions 1 and 2. Let the game infinitely repeated. Let Player X employ the Hedge algorithm to determine their stage strategy and let Player Y take a best response to the stage strategy of Player X. If the parameter η of the Hedge algorithm is sufficiently small, then after $\frac{8 \ln n}{\delta^2 \eta^2}$ stages, we have*

$$D_{KL}(\boldsymbol{x}^* || \boldsymbol{x}_t) \leq C\eta, \tag{5}$$

where $C = 2\lambda_m n^{3/2} \delta^2 + \delta$, $\lambda_m = \max_{k=1,2,\cdots,n} ||A_k^{-1}||_2$ and $\delta = \max_{i \in \mathcal{I}, j \in \mathcal{J}} a_{ij} - \min_{i \in \mathcal{I}, j \in \mathcal{J}} a_{ij}$.

Remark. This theorem demonstrates that the convergence rate of the strategy sequence of Player X to the NE strategy in the Hedge-myopic system is sublinear rather than linear. This is because the provided number of stages serves as an upper bound for the required stages, and simulations show that the actual convergence rate is faster.

3.1 DTHBR paradigm

By Theorem 1, we know that the KL divergence of the strategy $\mathbf{x}^*$ from the strategy $\mathbf{x}_t$ is bounded by the parameter η of the Hedge algorithm. Therefore, it is easy to understand that the last-iterate convergence can be attained by using "doubling trick" to gradually shrink the size of the parameter η. Based on this natural idea, we propose the following paradigm with last-iterate convergence for games satisfying the required assumptions, as presented by the pseudocode 1. We call this paradigm **DTHBR paradigm**.

The parameter η can be treated as a hyper-parameter that depends on the payoff matrix. Typically, it can be set to the order of $O(\sqrt{1/T})$, where T is the time horizon. However, if the structure of the Nash equilibrium (NE) is not well-behaved, i.e., if it is close to the boundary of the simplex $\Delta(I)$, the parameter η must be set to a smaller value.

Algorithm 1: The DTHBR paradigm

Input : Time horizon T, payoff matrix $A = (a_{ij})_{i \in \mathcal{I}, j \in \mathcal{J}}$, n, $\delta = \max_{i \in \mathcal{I}, j \in \mathcal{J}} a_{ij} - \min_{i \in \mathcal{I}, j \in \mathcal{J}} a_{ij}$;
Initialization: $\eta, \mathbf{x}_1 = (\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}), t = 1$;
repeat
  repeat
    $\mathbf{y}_t \leftarrow \arg\min_{\mathbf{y} \in \Delta(\mathcal{J})} \mathbf{x}_t^T A \mathbf{y}$;
    $x_{i,t} \leftarrow \frac{\exp(\eta \sum_{\tau=1}^{t-1} e_i^T A y_\tau)}{\sum_{j=1}^{n} \exp(\eta \sum_{\tau=1}^{t-1} e_j^T A y_\tau)}$, $\forall i$;
    $t \leftarrow t + 1$;
  until $t > \frac{8 \ln n}{\eta^2 \delta^2}$;
  $\eta \leftarrow \eta / 2$
until $t \geq T$;
Output: $\mathbf{x}_T$

4 Simulation Results

In this section, we conduct two simulations to verify the effectiveness of the DTHBR paradigm.

First, we consider a variant of the game Matching Pennies [32], in which the reward obtained by matching obverse is doubled. The payoff matrix of this game is

$$\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$$

and the NE is $((2/5, 3/5), (2/5, 3/5))$. The Kullback-Leibler (KL) divergence is used to measure the distance between the strategy sequence of Player X and the NE strategy of Player X.

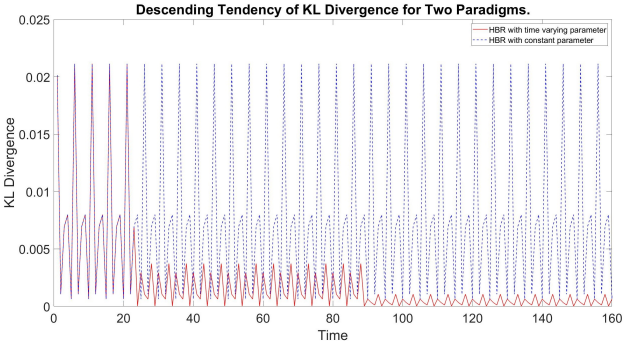

Fig. 1: The KL divergence of the HBR paradigm and the DTHBR paradigm for the Matching Pennies game.

We employ two paradigms with an identical time horizon T to solve the NE of this game. The first paradigm is referred to as the HBR paradigm, where the Hedge algorithm utilizes a constant parameter $\eta = \sqrt{8 \ln 2 / T}$. The second paradigm is the DTHBR paradigm, which incorporates the doubling trick into the parameter of the Hedge algorithm. The initial

parameter η is also set to $\eta = \sqrt{8 \ln 2/T}$ and is decreased according to the paradigm described in the pseudocode 1.

From Figure 2, it can be seen that the strategy sequence of Player X maintained by the HBR paradigm enters a cycle after a finite number of stages. This observation aligns with the theoretical result established in [14]. Since the NE of this game is close to the average distribution, the dynamics quickly enter a cycle, rendering further computations seemingly redundant, as they do not improve the quality of the obtained result.

In contrast, the strategy sequence of Player X maintained by the DTHBR paradigm breaks free from this cyclical behavior and indeed gets closer to the NE strategy of Player X! This simulation result demonstrates that by reducing the size of the parameter η through the doubling trick, last-iterate convergence is indeed achieved.

Next, we consider a 3×3 game with payoff matrix being

$$\begin{pmatrix} 2 & -1 & -3 \\ -1 & -2 & 2 \\ -2 & 0 & 1 \end{pmatrix}$$

This game admits a unique interior NE, which is $((3/8, 1/24, 7/12), (3/8, 1/3, 7/24))$.

We also employ two paradigms with an identical time horizon T to solve the NE of this game. The parameter of the Hedge algorithm is set to $\eta = \sqrt{8 \ln 3/T}$. The trend of the KL divergence is illustrated by Figure 2. The periodicity of the HBR paradigm and the efficiency of shrinking the parameter by using doubling trick are also verified in this simulation.

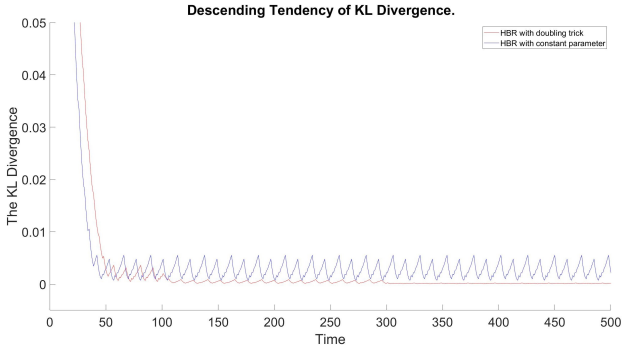

Fig. 2: The KL divergence of the HBR paradigm and the DTHBR paradigm for the second game.

5 Conclusion and Future Work

In this paper, we propose a novel asymmetric NE-solving paradigm with last-iterate convergence for two-player zero-sum games. Specifically, we allow one player to update their strategy using the Hedge algorithm while the other player adopts a corresponding best response. This approach offers a novel framework for leveraging learning algorithms to solve Nash equilibrium with guaranteed last-iterate convergence.

However, this paradigm still has some limitations, such as the requirement that the Nash equilibrium (NE) strategy must be unique. Future research could explore scenarios where the game admits multiple NEs or where the NE is not interior. Moreover, the convergence rate derived here may not be tight, and whether a faster rate can be achieved remains an open question. Future work could explore analyzing this system using different techniques to determine if a better convergence rate is attainable.

References

[1] Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Tuomas Sandholm. Uncoupled learning dynamics with $o(\log t)$ swap regret in multiplayer games. *Advances in Neural Information Processing Systems*, 35:3292–3304, 2022.

[2] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

[3] Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities. In *Conference on Learning Theory*, pages 326–358. PMLR, 2021.

[4] James P Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 321–338, 2018.

[5] Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Uncoupled and convergent learning in two-player zero-sum markov games with bandit feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[6] Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34:27952–27964, 2021.

[7] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[8] Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *arXiv preprint arXiv:1807.04252*, 2018.

[9] Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 235–254. SIAM, 2011.

[10] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

[11] Liad Erez, Tal Lancewicki, Uri Sherman, Tomer Koren, and Yishay Mansour. Regret minimization and convergence to equilibria in general-sum markov games. In *International Conference on Machine Learning*, pages 9343–9373. PMLR, 2023.

[12] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[13] Drew Fudenberg and David K Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

[14] Xinxiang Guo, Yifen Mu, and Xiaoguang Yang. Periodicity in hedge-myopic system and an asymmetric ne-solving paradigm for two-player zero-sum games. *arXiv preprint arXiv:2403.04336*, 2024.

[15] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68 (5):1127–1150, 2000.

[16] Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311, 1984.

[17] Nick Littlestone and Manfred K Warmuth. The weighted ma-

jority algorithm. *Information and computation*, 108(2):212–261, 1994.

[18] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the twenty-ninth annual ACM-SIAM symposium on discrete algorithms*, pages 2703–2717. SIAM, 2018.

[19] Koichi Miyasawa. *On the convergence of the learning process in a 2 x 2 non-zero-sum two-person game*. Princeton University Princeton, 1961.

[20] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

[21] John F. Nash. Non-cooperative games. *Annals of mathematics*, 54(2):286–295, 1951.

[22] Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

[23] Gerasimos Palaiopanos, Ioannis Panageas, and Georgios Piliouras. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. *Advances in Neural Information Processing Systems*, 30, 2017.

[24] Christos Papadimitriou and Georgios Piliouras. From nash equilibria to chain recurrent sets: Solution concepts and topology. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 227–235, 2016.

[25] Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, et al. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *International Conference on Machine Learning*, pages 8525–8535. PMLR, 2021.

[26] Georgios Piliouras and Jeff S Shamma. Optimization despite chaos: Convex relaxations to complex limit sets via poincaré recurrence. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 861–873. SIAM, 2014.

[27] Georgios Piliouras, Carlos Nieto-Granda, Henrik I Christensen, and Jeff S Shamma. Persistent patterns: multi-agent learning beyond equilibrium and utility. In *AAMAS*, pages 181–188, 2014.

[28] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013.

[29] Yuzuru Sato, Eizo Akiyama, and J Doyne Farmer. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences*, 99(7):4748–4751, 2002.

[30] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[31] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems*, 28, 2015.

[32] John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior (60th Anniversary Commemorative Edition)*. Princeton university press, 2007.

[33] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Conference on learning theory*, pages 4259–4299. PMLR, 2021.

[34] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20, 2007.