ASYNCHRONOUS GRADIENT PLAY IN ZERO-SUM MULTI-AGENT GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Finding equilibria via gradient play in competitive multi-agent games has been attracting a growing amount of attention in recent years, with emphasis on designing efficient strategies where the agents operate in a decentralized and symmetric manner with guaranteed convergence. While significant efforts have been made in understanding zero-sum two-player matrix games, the performance in zero-sum multi-agent games remains inadequately explored, especially in the presence of delayed feedbacks, leaving the scalability and resiliency of gradient play open to questions. In this paper, we make progress by studying *asynchronous gradient plays* in zero-sum polymatrix games under delayed feedbacks. We first establish that the last iterate of entropy-regularized optimistic multiplicative weight updates (OMWU) method converges linearly to the quantal response equilibrium (QRE), the solution concept under bounded rationality, in the absence of delays. The linear convergence continues to hold even when the feedbacks are *randomly* delayed under mild statistical assumptions, albeit at a slower rate. Moving beyond random delays, we further demonstrate entropy-regularized OMWU with two-timescale learning rates enjoys faster last-iterate convergence under fixed delays, and continues to converge provably even when the delays are arbitrarily bounded. Our methods also lead to finite-time guarantees to approximate the Nash equilibrium (NE) by moderating the amount of regularization. To the best of our knowledge, this work is the first that aims to understand asynchronous gradient play in zero-sum polymatrix games under a wide range of delay assumptions.

1 INTRODUCTION

Finding equilibria of multi-player games via gradient play lies at the heart of game theory, which permeates a remarkable breadth of modern applications, including but not limited to competitive reinforcement learning (RL) (Littman, 1994), generative adversarial networks (GANs) (Goodfellow et al., 2014) and adversarial training (Mertikopoulos et al., 2018). While conventional wisdom leans towards the paradigm of centralized learning, retrieving and sharing information across multiple agents raise questions in terms of both privacy and efficiency, leading to a significant amount of interest in designing decentralized learning algorithms that utilize only local payoff feedbacks, with the updates at different agents executed in a symmetric manner.

In reality, there is no shortage of scenarios where the feedback can be obtained only in a delayed manner (He et al., 2014), i.e., the agents only receive the payoff information sent from a previous round instead of the current round, due to communication slowdowns and congestions, for example. Substantial progress has been made towards reliable and efficient online learning with delayed feedbacks in various settings, e.g., stochastic multi-armed bandit (Pike-Burke et al., 2018; Vernade et al., 2017), adversarial multi-armed bandit (Cesa-Bianchi et al., 2016; Li et al., 2019), online convex optimization (Quanrud & Khashabi, 2015; McMahan & Streeter, 2014) and multi-player game (Meng et al., 2022). Typical approaches to combatting delays include subsampling the payoff history (Weinberger & Ordentlich, 2002; Joulani et al., 2013), or adopting adaptive learning rates suggested by delay-aware analysis (Quanrud & Khashabi, 2015; McMahan & Streeter, 2014; Hsieh et al., 2020; Flaspohler et al., 2021). Most of these efforts, however, have been limited to the study of *individual regret*, which characterizes the performance gap between an agent's learning trajectory and the best policy in hindsight. It remains highly inadequate when it comes to guaranteeing

Learning rate	Type of delay	Iteration complexity	
		ϵ-QRE	ϵ-NE
single-timescale	none	$\tau^{-1} d_{\mathsf{max}} \|A\|_\infty \log \epsilon^{-1}$	$d_{\mathsf{max}} \|A\|_\infty \epsilon^{-1}$
	statistical	$\tau^{-2} d_{\mathsf{max}}^2 \|A\|_\infty^2 (\gamma+1)^2 \log \epsilon^{-1}$	$d_{\mathsf{max}}^2 \|A\|_\infty^2 (\gamma+1)^2 \epsilon^{-2}$
two-timescale	constant	$\tau^{-1} d_{\mathsf{max}} \|A\|_\infty (\gamma+1)^2 \log \epsilon^{-1}$	$d_{\mathsf{max}} \|A\|_\infty (\gamma+1)^2 \epsilon^{-1}$
	bounded	$\tau^{-2} n d_{\mathsf{max}}^4 \|A\|_\infty^4 (\gamma+1)^{5/2} \epsilon^{-2}$	$n d_{\mathsf{max}}^4 \|A\|_\infty^4 (\gamma+1)^{5/2} \epsilon^{-4}$

Table 1: Iteration complexities of the proposed OMWU method for finding ϵ-QRE/NE of zero-sum polymatrix games, where logarithmic dependencies are omitted. Here, γ denotes the maximal time delay when the delay is bounded, n denotes the number of agents in the game, d_{max} is the maximal degree of the graph, and $\|A\|_\infty = \max_{i,j} |A_{i,j}|$ is the ℓ_∞ norm of the entire payoff matrix A (over all games in the network). We only present the result of statistical delay when the delays are bounded for direct comparison with other results, while more general bounds are given in Section 3.2.

convergence to the equilibrium in a multi-player environment, especially in the presence of delayed feedbacks, thus leaving the scalability and resiliency of gradient play open to questions.

In this work, we initiate the study of asynchronous learning algorithms for zero-sum polymatrix games (also known as network matrix games (Bergman & Fokin, 1998)), which generalizes two-player zero-sum matrix games to the multiple-player setting and serves as an important stepping stone to more general multi-player general-sum games. In particular, we focus on *finite-time last-iterate* convergence to two prevalent solution concepts in game theory, namely Nash Equilibrium (NE) and Quantal Response Equilibrium (QRE) which considers bounded rationality (McKelvey & Palfrey, 1995). Despite the seemingly simple formulation, few existing works have achieved this goal even in the synchronous setting, i.e., with instantaneous feedback. Leonardos et al. (2021) studied a continuous-time learning dynamics that converges to a QRE at a linear rate. Anagnostides et al. (2022) demonstrated Optimistic Mirror Descent (OMD) (Rakhlin & Sridharan, 2013) enjoys finite-time last-iterate convergence to the NE, yet the analysis therein requires continuous gradient of the regularizer, which incurs computation overhead for solving a subproblem every iteration. In contrast, the typical choice of entropic regularizer leads to closed-form multiplicative updates, which is computationally more desirable but remains poorly understood. In sum, designing efficient learning algorithms that provably converge to the equilibrium has been technically challenging, even in the synchronous setting.

1.1 Our contributions

In this paper, we develop provably convergent algorithms—broadly dubbed as *asynchronous gradient play*—to find the QRE and NE of zero-sum polymatrix games in a decentralized manner with delayed feedbacks. We propose an entropy-regularized Optimistic Multiplicative Weights Update (OMWU) method (Cen et al., 2021), where each player symmetrically updates their strategies without access to the payoff matrices and other players' strategies, and initiate a systematic investigation on the impacts of delays on its convergence. Our main contributions are summarized as follows.

- *Finite-time last-iterate convergence of single-timescale OMWU.* We begin by showing that, in the synchronous setting, the single-timescale OMWU method—when the same learning rate is adopted for extrapolation and update—achieves last-iterate convergence to the QRE at a linear rate, which is independent of the number of agents as well as the size of action spaces (up to logarithmic factors). In addition, this implies a last-iterate convergence to an ϵ-approximate NE in $\widetilde{\mathcal{O}}(\epsilon^{-1})$ iterations by adjusting the regularization parameter, where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic dependencies. Encouragingly, the last-iterate linear convergence to QRE continues to hold in the asynchronous setting, as long as the delay sequence follows certain mild statistical assumptions, albeit at a slower rate, whereas the iteration complexity to find an ϵ-NE degenerates to $\widetilde{\mathcal{O}}(\epsilon^{-2})$.
- *Finite-time convergence of two-timescale OMWU.* Motivated by applications where the delays are constant and known (e.g. from timestamp information), we propose to adopt a two-timescale

OMWU method which separates the learning rates of extrapolation and update in a delay-aware manner, and show it achieves a faster last-iterate linear convergence to QRE in the presence of constant delays, with an improved $\widetilde{\mathcal{O}}(\epsilon^{-1})$ iteration complexity to ϵ-NE. We further generalize the analysis to tackle the more practical yet challenging setting where the feedback sequence is permutated by bounded delays—possibly in an adversarial manner—and demonstrate provable convergence to the equilibria.

We summarize the iteration complexities of the proposed methods for finding ϵ-approximate solutions of QRE and NE in Table 1. To the best of our knowledge, this work presents the first algorithm design and analysis that focus on equilibrium finding in a multi-player game with delayed feedbacks. In contrast, most of existing works concerning individual regret in the synchronous/asynchronous settings typically yield average-iterate convergence guarantees (see e.g., Bailey (2021); Meng et al. (2022)) and fall short of characterizing the actual learning trajectory to the equilibrium.

1.2 Further related works

Learning in two-player zero-sum matrix games. Freund & Schapire (1999) proved that Multiplicative Weights Update (MWU) method achieve an average-iterate convergence rate of $\mathcal{O}(1/\sqrt{T})$ through the lens of regret analysis. Daskalakis et al. (2011) is the first to achieve an optimal convergence rate of $\mathcal{O}(1/T)$ with the excessive gap technique of Nesterov. Rakhlin & Sridharan (2013) achieved the same rate with OMD, which is more commonly referred to as OMWU when entropy regularization is in use for the mirror descent update rule. In terms of last-iterate convergence, Daskalakis & Panageas (2018) established asymptotic last-iterate convergence for OMWU assuming the uniqueness of NE. Wei et al. (2021) improved upon the analysis under the same assumption by showing a problem-dependent linear rate of convergence, which is extended to a class of extensive-form games (Lee et al., 2021). Cen et al. (2021) showed that entropy-regularized OMWU converges linearly to the QRE of two-player zero-sum matrix game, which translates to an iteration complexity of $\widetilde{\mathcal{O}}(1/T)$ for finding an ϵ-NE, without assuming its uniqueness. Sokota et al. (2022) showed that linear convergence to QRE can be achieved without resorting to optimistic update rules, e.g., using entropy-regularized MWU, albeit with a more restrictive learning rate. It is worth pointing out that the idea of learning rate separation has been explored for equilibrium finding in two-player zero-sum games with instant feedback (Fasoulakis et al., 2022) and online learning with delayed feedback (Hsieh et al., 2020), but lacks study in an asynchronous multi-player game setting.

Asynchronous optimization. Asynchronous and decentralized optimization algorithms have been extensively studied since the proposal of Bertsekas & Tsitsiklis (1989), where a number of agents seek to find an approximate global optimizer of a common loss function, by performing iterative gradient-based methods in a collaborative manner. Typical approaches including parallelizing the computation of gradient with regard to data (Tong et al., 2020), or parallelizing the model updates by imposing coordinate update rules (Nesterov, 2012; Liu et al., 2014; Liu & Wright, 2015). Delayed gradient (feedback) is also common in these scenarios due to the existence of other agents updating the model. Moreover, the zero-sum polymatrix setting considered in this work is inherently non-collaborative by requiring every agent to maximize its own utility function and compete with other agents, and leads to substantially difference analysis techniques.

1.3 Notation and paper organization

Denote by $[n]$ the set $\{1, \cdots, n\}$ and by $\Delta(S)$ the probability simplex over the set S. Given two probability distributions $p, p' \in \Delta(S)$, the KL divergence from p' to p is defined by $\mathsf{KL}(p \,\|\, p') := \sum_{k \in S} p(k) \log \frac{p(k)}{p'(k)}$. For any vector $z = [z_i]_{1 \le i \le n} \in \mathbb{R}^n$, we use $\exp(z)$ to represent $[\exp(z_i)]_{1 \le i \le n}$. The rest of this paper is organized as follows. Section 2 provides the preliminary on zero-sum polymatrix games and solution concepts. Performance guarantees of single-timescale OMWU and two-timescale OMWU are presented in Section 3 and Section 4, respectively. Numerical experiments are provided in Section 5 to corroborate the theoretical findings, and finally, we conclude in Section 6. The proofs are deferred to the appendix.

2 PRELIMINARIES

In this section, we introduce the formulation of zero-sum polymatrix games as well as the solution concept of NE and QRE.

Definition 1 (Polymatrix game). *Let $\mathcal{G} := \{(V, E), \{S_i\}_{i\in V}, \{A_{ij}\}_{(i,j)\in E}\}$ be an n-player polymatrix game, where each element in the tuple is defined as follows.*

- *An undirected graph (V, E), with $V = [n]$ denoting the set of players and E the set of edges;*
- *For each player $i \in V$, S_i represents its action set, which is assumed to be finite;*
- *For each edge $(i, j) \in E$, $A_{ij} \in \mathbb{R}^{|S_i|\times|S_j|}$ and $A_{ji} \in \mathbb{R}^{|S_j|\times|S_i|}$ represent the payoff matrices associated with player i and j, i.e., when player i and player j choose $s_i \in S_i$ and $s_j \in S_j$, the received payoffs are given by $A_{ij}(s_i, s_j)$, $A_{ji}(s_j, s_i)$, respectively.*

Given the strategy profile $s = (s_1, \cdots, s_n) \in S = \prod_{i\in V} S_i$ taken by all players, the utility function $u_i : S \to \mathbb{R}$ of player i is given by

$$u_i(s) = \sum\nolimits_{j:(i,j)\in E} A_{ij}(s_i, s_j).$$

Suppose that player i adopts a mixed/stochastic strategy or policy, $\pi_i \in \Delta(S_i)$, where the probability of selecting $s_i \in S_i$ is specified by $\pi_i(s_i)$. With slight abuse of notation, we denote the expected utility of player i with a mixed strategy profile $\pi = (\pi_1, \cdots, \pi_n) \in \Delta(S)$ as

$$u_i(\pi) = \mathop{\mathbb{E}}_{s_i\sim\pi_i,\forall i\in V}[u_i(s)] = \sum_{j:(i,j)\in E} \pi_i^\top A_{ij}\pi_j. \tag{1}$$

It turns out to be convenient to treat π_i and π as vectors in $\mathbb{R}^{|S_i|}$ and $\mathbb{R}^{\sum_{i\in V}|S_i|}$ without ambiguity, and concatenate all payoff matrices associated with player i into

$$A_i = (A_{i1}, \cdots, A_{in}) \in \mathbb{R}^{|S_i|\times\sum_{j\in V}|S_j|}, \tag{2}$$

where A_{ij} is set to 0 whenever $(i, j) \notin E$. In particular, it follows that $A_{ii} = 0$ for all $i \in V$. With these notation in place, we can rewrite the expected utility function (1) as

$$u_i(\pi) = \pi_i^\top A_i\pi, \tag{3}$$

where $A_i\pi \in \mathbb{R}^{|S_i|}$ can be interpreted as the expected utility of the actions in S_i for player i. In addition, we denote the maximum entrywise absolute value of payoff by $\|A\|_\infty = \max_{i,j}\|A_{ij}\|_\infty = \max_i \|A_i\|_\infty$, and the maximum degree of the graph by $d_{\mathsf{max}} = \max_{i\in V} \deg_i$.

Zero-sum polymatrix games. The game $\mathcal{G}$ is a zero-sum polymatrix game if it holds that

$$\sum\nolimits_{i\in V} u_i(s) = 0, \qquad \forall\, s \in S. \tag{4}$$

This also immediately implies that for any strategy profile $\pi \in \Delta(S)$, it follows that $\sum_{i\in V} u_i(\pi) = 0$. Zero-sum polymatrix games are commonly used to describe situations in which agents' interactions are captured by an interaction graph and the entire system of games are closed so that the total payoffs keep invariant in the system, which find applications in security games (Cai et al., 2016), graph transduction (Bernardi, 2021), etc.

Nash equilibrium (NE). A mixed strategy profile $\pi^\star = (\pi_1^\star, \cdots, \pi_n^\star)$ is a Nash equilibrium (NE) when each player i cannot further increase its own utility function u_i by unilateral deviation, i.e., $u_i(\pi_i', \pi_{-i}^\star) \le u_i(\pi_i^\star, \pi_{-i}^\star)$, for all $i \in V,\ \pi_i' \in \Delta(S_i)$, where the existence is guaranteed by the work (Cai et al., 2016). Here we denote the mixed strategies of all players other than i by π_{-i} and write $u_i(\pi_i, \pi_{-i}) = u_i(\pi)$. To measure how close a strategy $\pi \in \Delta(S)$ is to an NE, we introduce

$$\texttt{NE-Gap}(\pi) = \max_{i\in V}\left[\max_{\pi_i'\in\Delta(S_i)} u_i(\pi_i', \pi_{-i}) - u_i(\pi)\right],$$

which measures the largest possible gain in the expected utility when players deviate from its strategy unilaterally. A mixed strategy profile π is called an ϵ-approximate Nash equilibrium (ϵ-NE) when $\texttt{NE-Gap}(\pi) \le \epsilon$, which ensures that $u_i(\pi_i', \pi_{-i}) \le u_i(\pi_i, \pi_{-i}) + \epsilon$, for all $i \in V,\ \pi_i' \in \Delta(S_i)$.

Quantal response equilibrium (QRE). The quantal response equilibrium (QRE), proposed by McKelvey & Palfrey (1995), generalizes the classical notion of NE under uncertain payoffs or bounded rationality, while balancing exploration and exploitation. A mixed strategy profile $\pi_\tau^\star = (\pi_{1,\tau}^\star, \cdots, \pi_{n,\tau}^\star)$ is a QRE when each player assigns its probability of action according to the expected utility of every action in a Boltzmann fashion, i.e., for all $i \in V$,

$$\pi_{i,\tau}^\star(k) = \frac{\exp([A_i \pi_\tau^\star]_k/\tau)}{\sum_{k \in S_i} \exp([A_i \pi_\tau^\star]_k/\tau)}, \qquad k \in S_i, \tag{5}$$

where $\tau > 0$ is the regularization parameter or temperature. Equivalently, this amounts to maximizing an entropy-regularized utility of each player (Mertikopoulos & Sandholm, 2016), i.e., $u_{i,\tau}(\pi_i', \pi_{-i,\tau}^\star) \leq u_{i,\tau}(\pi_{i,\tau}^\star, \pi_{-i,\tau}^\star)$ for all $i \in V$, $\pi_i' \in \Delta(S_i)$. Here, the entropy-regularized utility function $u_i : S \to \mathbb{R}$ of player i is given by

$$u_{i,\tau}(\pi) = u_i(\tau) + \tau \mathcal{H}(\pi_i), \tag{6}$$

where $\mathcal{H}(\pi_i) = -\pi_i^\top \log \pi_i$ denotes the Shannon entropy of π_i. In Leonardos et al. (2021), it is shown that a unique QRE exists in a zero-sum polymatrix game. Similarly, we can measure the proximity of a strategy π to a QRE by

$$\texttt{QRE-Gap}_\tau(\pi) = \max_{i \in V} \left[\max_{\pi_i' \in \Delta(S_i)} u_{i,\tau}(\pi_i', \pi_{-i}) - u_{i,\tau}(\pi) \right]. \tag{7}$$

A mixed strategy profile π is called an ϵ-QRE when $\texttt{QRE-Gap}_\tau(\pi) \leq \epsilon$. It follows immediately that we can link an $\epsilon/2$-QRE to ϵ-NE according to the following relationship

$$|\texttt{QRE-Gap}_\tau(\pi) - \texttt{NE-Gap}(\pi)| \leq \tau \max_{i \in V} \log |S_i|, \tag{8}$$

by setting $\tau = \frac{\epsilon}{2 \max_{i \in V} \log |S_i|}$.

3 Performance guarantees of single-timescale OMWU

In this section, we present and study the entropy-regularized OMWU method (Cen et al., 2021) for finding the QRE of zero-sum polymatrix games. Whilst the method is originally proposed for finding QRE in a two-player zero-sum game, the update rule naturally generalizes to the multi-player setting as

$$\pi_i^{(t+1)}(k) \propto \pi_i^{(t)}(k)^{1-\eta\tau} \exp(\eta [A_i \overline{\pi}^{(t+1)}]_k), \qquad \forall k \in S_i, \tag{9}$$

where $\eta > 0$ is the step size and $\overline{\pi}^{(t+1)}$ serves as a prediction for $\pi^{(t+1)}$ via an extrapolation step

$$\overline{\pi}_i^{(t+1)}(k) \propto \pi_i^{(t)}(k)^{1-\eta\tau} \exp(\eta [A_i \overline{\pi}^{(t)}]_k), \qquad \forall k \in S_i. \tag{10}$$

In the asynchronous setting, however, each agent i receives a delayed payoff vector $A_i \overline{\pi}^{(\kappa_i^{(t)})}$ instead of $A_i \overline{\pi}^{(t)}$ in the t-th iteration, where $\kappa_i^{(t)} = t - \gamma_i^{(t)}$ with $\gamma_i^{(t)} \geq 0$ representing the length of delay. The detailed procedure is outlined in Algorithm 1 using the single-timescale rule for extrapolation.

3.1 Performance guarantees without delays

We first present our theorem concerning the last-iterate convergence of single-timescale OMWU for finding the QRE in the synchronous setting, i.e. $\gamma_i^{(t)} = 0$ for all $i \in V$ and $t \geq 0$. For any $\pi, \pi' \in V$, let $\mathsf{KL}(\pi \,\|\, \pi') = \sum_{i \in V} \mathsf{KL}(\pi_i \,\|\, \pi_i')$.

Theorem 1 (Last-iterate convergence without delays)**.** *Suppose that the learning rate η of single-timescale OMWU in Algorithm 1 obeys*

$$0 < \eta \leq \min\left\{ \frac{1}{2\tau}, \frac{1}{4 d_{\mathsf{max}} \|A\|_\infty} \right\}, \tag{13}$$

then the last iterates of single-timescale OMWU satisfy that

$$\max\left\{ \mathsf{KL}(\pi_\tau^\star \,\|\, \pi^{(t)}), \mathsf{KL}(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}), \texttt{QRE-Gap}_\tau(\overline{\pi}^{(t)}) \right\} \leq \mathcal{O}((1-\eta\tau)^t).$$

Algorithm 1 Entropy-regularized OMWU, agent i

1: Initialize $\pi_i^{(0)} = \overline{\pi}_i^{(0)}$ as uniform distribution. Learning rates η, and $\overline{\eta}$ (optional).
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: Receive payoff vector $A_i \overline{\pi}^{(\kappa_i^{(t)})}$.
4: When $t \geq 1$, update π_i according to

$$\pi_i^{(t)}(k) \propto \pi_i^{(t-1)}(k)^{1-\eta\tau} \exp(\eta[A_i \overline{\pi}^{(\kappa_i^{(t)})}]_k), \qquad \forall k \in S_i.$$

5: Update $\overline{\pi}_i$ according to the single-timescale rule

$$\overline{\pi}_i^{(t+1)}(k) \propto \pi_i^{(t)}(k)^{1-\eta\tau} \exp(\eta[A_i \overline{\pi}^{(\kappa_i^{(t)})}]_k), \qquad \forall k \in S_i. \tag{11}$$

or the two-timescale rule

$$\overline{\pi}_i^{(t+1)}(k) \propto \pi_i^{(t)}(k)^{1-\overline{\eta}\tau} \exp(\overline{\eta}[A_i \overline{\pi}^{(\kappa_i^{(t)})}]_k), \qquad \forall k \in S_i. \tag{12}$$

6: **end for**

Theorem 1 demonstrates that as long as the learning rate η is sufficiently small, the last iterate of single-timescale OMWU converges to the QRE at a linear rate. Compared with prior works for finding approximate equilibrium for zero-sum polymatrix games, our approach features a closed-form multiplicative update and a fast linear last-iterate convergence. Some remarks are in order.

- *Linear convergence to the QRE.* Theorem 1 implies an iteration complexity of $\widetilde{\mathcal{O}}\left(\frac{1}{\eta\tau}\log\frac{1}{\epsilon}\right)$ for finding an ϵ-QRE in a last-iterate manner, which translates to $\widetilde{\mathcal{O}}\left(\left(\frac{d_{\max}\|A\|_\infty}{\tau}+1\right)\log\frac{1}{\epsilon}\right)$ by optimizing the learning rate in (13). The result is especially appealing as it avoids direct dependency on the number of agents n as well as the size of action spaces (up to logarithmic factors), suggesting that learning in competitive multi-agent games can be made quite scalable as long as the interactions among the agents are sparse.
- *Last-iterate convergence to ϵ-NE.* By setting τ appropriately, we end up with an iteration complexity of $\widetilde{\mathcal{O}}\left(d_{\max}\|A\|_\infty/\epsilon\right)$ for achieving last-iterate convergence to an ϵ-NE (cf. (8)), which outperforms the best existing last-iterate rate of $\widetilde{\mathcal{O}}\left(n\|A\|_\infty/\epsilon^2\right)$ from Leonardos et al. (2021) by at least a factor of $n/(d_{\max}\epsilon)$.

Remark 1. *Our results trivially extend to the setting of weighted zero-sum polymatrix games (Leonardos et al., 2021), which amounts to adopting different learning rates $\{\eta_i\}_{i\in V}$ at each player. In this case, the iteration complexity becomes $\widetilde{O}\left(\max_{i\in V}\frac{1}{\eta_i\tau}\log\frac{1}{\epsilon}\right)$. In addition, our convergence result readily translates to a bound on individual regret as detailed in Appendix C.*

3.2 Performance guarantees under random delays

We continue to examine single-timescale OMWU in the more challenging asynchronous setting. In particularly, we show that the last iterate of single-timescale OMWU continues to converge linearly to the QRE, as long as the delays satisfy some mild statistical assumptions given below.

Assumption 1 (Random delays)**.** *Assume that for all $i \in V$, $t \geq 0$, the delay $\gamma_i^{(t)}$ is independently generated and satisfies*

$$\mathbb{E}_{\gamma_i^{(t)}\geq\ell}\left[\gamma_i^{(t)}\right] := \mathbb{E}\left[\gamma_i^{(t)} \mid \gamma_i^{(t)} \geq \ell\right] \leq E(\ell), \qquad \forall \ell = 0, 1, \ldots. \tag{14}$$

Additionally, there exists some constant $\zeta > 1$, such that $L \triangleq \sum_{\ell=0}^{\infty}\zeta^\ell E(\ell) < \infty$.

We remark that Assumption 1 is a rather mild condition that applies to typical delay distributions, such as the Poisson distribution with any $\zeta > 1$ (Zhang et al., 2020), as well as distributions with bounded support (Recht et al., 2011; Liu et al., 2014; Assran et al., 2020). Roughly speaking, Assumption 1 implies that the probability of the delay decays exponentially with its length given

that $\sum_{\ell=0}^{\infty}\zeta^{\ell}\mathbb{E}_{\gamma_i^{(t)}\geq\ell}\left[\gamma_i^{(t)}\right] = \mathbb{E}\left[\frac{\zeta^{\gamma_i^{(t)}+1}-1}{\zeta-1}\gamma_i^{(t)}\right]$, where ζ^{-1} approximately indicates the decay rate. We have the following theorem.

Theorem 2 (Last-iterate convergence with random delays)**.** *Under Assumption 1, suppose that the regularization parameter $\tau < 1$ and the learning rate η of single-timescale OMWU in Algorithm 1 obeys*

$$\frac{1}{\eta} \geq \max\left\{\tau + 2d_{\mathsf{max}}\|A\|_\infty\left[1+\frac{2d_{\mathsf{max}}\|A\|_\infty}{\tau}+2\Big(\frac{d_{\mathsf{max}}\|A\|_\infty}{\tau}+1\Big)L\right], \frac{\zeta}{\tau(\zeta-1)}\right\}, \quad (15)$$

then for any $T \geq 1$, the last iterate $\pi^{(T)}$ converges to $\pi_\tau^\star$ linearly. Optimizing the learning rate, the iteration complexity to ϵ-QRE is no more than

$$\widetilde{\mathcal{O}}\left(\max\left\{d_{\mathsf{max}}^2\|A\|_\infty^2(L+1), \frac{\zeta}{\zeta-1}\right\}\frac{1}{\tau^2}\log\frac{1}{\epsilon}\right).$$

To facilitate understanding of Theorem 2, we showcase the iteration complexity for finding ϵ-QRE/NE under two typical scenarios: bounded delay and Poisson delay.

- *Bounded random delay.* When the delays are bounded above by some maximum delay γ, Assumption 1 is met with $\zeta = 1+\gamma^{-1}$ and $L = e\gamma(\gamma+1)$. Plugging into Theorem 2 yields an iteration complexity of
$$\widetilde{\mathcal{O}}\Big(\frac{d_{\mathsf{max}}^2\|A\|_\infty^2(\gamma+1)^2}{\tau^2}\log\frac{1}{\epsilon}\Big)$$
for finding an ϵ-QRE, or $\widetilde{\mathcal{O}}\big(d_{\mathsf{max}}^2\|A\|_\infty^2\gamma^2/\epsilon^2\big)$ for finding an ϵ-NE, which increases quadratically as the maximum delay increases. Note that these rates are slightly worse than those without delays (cf. Theorem 1).
- *Poisson delay.* When the delays follow the Poisson distribution with parameter $1/\overline{T}$, it suffices to set $\zeta = 1+\overline{T}^{-1}$ and $L = e\overline{T}(1+\overline{T})$ Assumption 1. This leads to an iteration complexity of
$$\widetilde{\mathcal{O}}\Big(\frac{d_{\mathsf{max}}^2\|A\|_\infty^2\overline{T}^2}{\tau^2}\log\frac{1}{\epsilon}\Big)$$
for finding an ϵ-QRE, or $\widetilde{\mathcal{O}}\big(d_{\mathsf{max}}^2\|A\|_\infty^2\overline{T}^2/\epsilon^2\big)$ for finding an ϵ-NE, which is similar to the bounded random delay case.

4 Performance guarantees of two-timescale OMWU

While Theorem 2 demonstrates provable convergence of single-timescale OMWU with random delays, it remains unclear whether the update rule can be better motivated in more general asynchronous settings, and whether the convergence can be further ensured under adversarial delays. Indeed, theoretical insights from previous literature (Mokhtari et al., 2020; Cen et al., 2021) suggest the critical role of $\overline{\pi}^{(t)}$ as a predictive surrogate for $\pi^{(t)}$ in enabling fast convergence, which no longer holds when $\overline{\pi}^{(t)}$ is replaced by a delayed feedback from $\overline{\pi}^{(\kappa_i^{(t)})}$. To this end, we propose to replace the extrapolation update (11) with one equipped with a different learning rate:

$$\overline{\pi}_i^{(t+1)}(k) \propto \pi_i^{(t)}(k)^{1-\overline{\eta}\tau}\exp(\overline{\eta}[A_i\overline{\pi}^{(\kappa_i^{(t)})}]_k), \qquad \forall k\in S_i, \quad (16)$$

which adopts a larger learning rate $\overline{\eta} > \eta$ to counteract the delay. Intuitively, a choice of $\overline{\eta} \approx (\gamma_i^{(t)}+1)\eta$ would allow $\overline{\pi}^{(\kappa_i^{(t)})}$ to approximate $\pi^{(t)}$ by taking the intermediate updates $\{\pi^{(l)} : \kappa_i^{(t)} \leq l < t\}$ into consideration. We refer to this update rule as the two-timescale entropy-regularized OMWU, whose detailed procedure is again outlined in Algorithm 1.

4.1 Performance guarantees under constant and known delays

To highlight the potential benefit of learning rate separation, we start by studying the convergence of two-timescale OMWU in the asynchronous setting with constant and known delays, which has been studied in (Weinberger & Ordentlich, 2002; Flaspohler et al., 2021; Meng et al., 2022). We have the following theorem, which reveals a a faster linear convergence to the QRE by using a delay-aware two-timescale learning rate design.

Theorem 3 (Last-iterate convergence with fixed delays)**.** *Suppose that the delays $\gamma_i^{(t)} = \gamma$ are fixed and known, and for notational simplicity, we denote $A_i\overline{\pi}^{(t-\gamma)} = A_i\pi^{(0)}$ for the first $0 \le t < \gamma$ iterations. Suppose that the learning rate η of two-timescale OMWU in Algorithm 1 satisfies $\eta^{-1} > \mathcal{O}(\gamma\tau + 4d_{\max}\|A\|_\infty(\gamma+1)^2)$, and $\overline{\eta}$ is determined by $1-\overline{\eta}\tau = (1-\eta\tau)^{(\gamma+1)}$, then the last iterate $\pi^{(t)}$ converges to the QRE at a linear rate:*

$$\max\left\{\mathsf{KL}\big(\pi_\tau^\star \,\|\, \pi^{(t)}\big), \mathtt{QRE\text{-}Gap}_\tau(\pi^{(t)})\right\} \le \mathcal{O}((1-\eta\tau/2)^t).$$

By optimizing the learning rate η, Theorem 3 implies that two-timescale OMWU takes at most

$$\widetilde{\mathcal{O}}\left(\frac{d_{\max}\|A\|_\infty(\gamma+1)^2}{\tau}\log\frac{1}{\epsilon}\right)$$

iterations to find an ϵ-QRE in a last-iterate manner, which translates to an iteration complexity of $\widetilde{\mathcal{O}}\left(d_{\max}\|A\|_\infty(\gamma+1)^2/\epsilon\right)$ for finding an ϵ-NE. This significantly improves over the iteration complexity of $\widetilde{\mathcal{O}}\big(d_{\max}^2\|A\|_\infty^2(\gamma+1)^2/\epsilon^2\big)$ for single-timescale OMWU, verifying the positive role of adopting two-timescale learning rate in enabling faster convergence.

4.2 Permuted feedback with bounded delay

The above result requires the exact information of the delay, which may not always be available. Motivated by the need to address arbitrary or even adversarial delays, we consider a more realistic scenario, where the payoff sequence arrives in a permuted order (Agarwal & Duchi, 2011) constrained by a maximum bounded delay (McMahan & Streeter, 2014; Wan et al., 2022).

Assumption 2 (Bounded delay)**.** *For any $i \in V$ and $t > 0$, it holds that $\gamma_i^{(t)} \le \gamma$.*

Assumption 3 (Permuted feedback)**.** *For any $t > 0$, the payoff vector at the t-th iteration is received by agent i only once. The payoff at the 0-th iteration can be used multiple times.*

The following theorem unveils the convergence of two-timescale OMWU to the QRE in an average sense under permutated bounded delays.

Theorem 4 (Average-iterate convergence under permutated delays)**.** *Under Assumption 2 and 3, suppose that the learning rate η of two-timescale OMWU in Algorithm 1 satisfies $\eta^{-1} > \mathcal{O}(d_{\max}\|A\|_\infty(\gamma+1)^{5/2})$, and $\overline{\eta}$ is determined by $1-\overline{\eta}\tau = (1-\eta\tau)^{(\gamma+1)}$, it holds for $T > 2\gamma$ that*

$$\begin{cases} \frac{1}{T-2\gamma}\sum_{t=2\gamma+1}^{T}\mathsf{KL}\big(\pi_\tau^\star \,\|\, \pi^{(t)}\big) & \le \widetilde{\mathcal{O}}\left(\frac{nd_{\max}\|A\|_\infty}{\eta\tau^2(T-2\gamma)}\right) \\ \frac{1}{T-2\gamma}\sum_{t=2\gamma+1}^{T}\mathtt{QRE\text{-}Gap}_\tau(\pi^{(t)}) & \le \widetilde{\mathcal{O}}\left(\sqrt{\frac{nd_{\max}^3\|A\|_\infty^3}{\eta\tau^2(T-2\gamma)}}\right) \end{cases}. \tag{17}$$

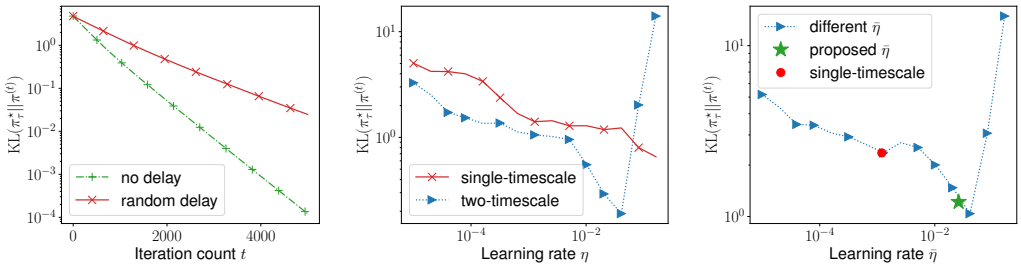
Theorem 4 guarantees that the best iterate among $\{\pi^{(t)}\}_{2\gamma<t\le T}$ is an ϵ-QRE as long as $T \ge \widetilde{\mathcal{O}}\left(\frac{nd_{\max}^4\|A\|_\infty^4(\gamma+1)^{\frac{5}{2}}}{\tau^2\epsilon^2}\right)$, which translates to an iteration complexity of $\widetilde{\mathcal{O}}\big(\frac{nd_{\max}^4\|A\|_\infty^4(\gamma+1)^{\frac{5}{2}}}{\epsilon^4}\big)$ for finding an ϵ-NE. While the rate seems slower than the previous theorems, Theorem 4 holds under arguably the weakest delay assumptions, where it can be even adversarially bounded. We remark that the result in (17) also guarantees the convergence of the last iterate $\pi^{(t)}$ to the QRE asymptotically, although without a finite-time rate. This is in sharp contrast to typical average-iterate analysis that only applies to $\frac{1}{T}\sum_{t=1}^{T}\pi^{(t)}$ without implications on the convergence of the last iterate $\pi^{(t)}$.

5 Numerical Experiments

In this section, we verify our theoretical findings by investigating the performance of both single-timescale and two-timescale OMWU on randomly generated zero-sum entropy-regularized polymatrix games with $n = 10$, $|S_i| = 10, i \in V$ and $\tau = 0.1$. For each $(i,j) \in E$, we set $A_{ij} = -A_{ji}^\top$ with entries of A_{ij} independently sampled from the uniform distribution over $[-1, 1]$. All the results are averaged over five independent runs.

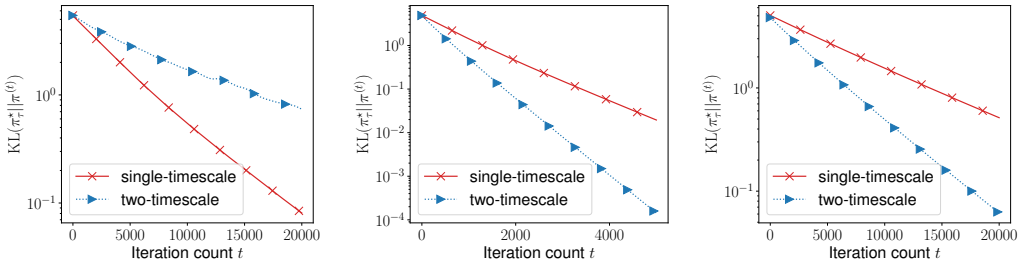
In Fig. 1 (a), we compare the performance of single-timescale OMWU in both synchronous and asynchronous settings, with delay uniformly sampled from $\{0, 1, \ldots, 10\}$. We adopt the optimal learning rate η from $\{0.1, 0.05, 0.02, 0.01, \ldots\}$ that yields the highest accuracy. The method achieves linear convergence in both cases, yet the convergence rate is slowed down by delayed feedbacks in the asynchronous setting. Fig. 1 (b) and (c) compare the effect of different choices of learning rates $\eta, \overline{\eta}$ on the performance of the proposed methods, where the feedback is permutated with bounded delay $\gamma = 25$ (cf. Assumptions 2 and 3). In general, two-timescale OMWU outperforms single-timescale OMWU given appropriate choices of learning rate η. On the other hand, (c) demonstrates that the choice of $\bar{\eta} = \tau^{-1}(1 - (1 - \eta\tau)^{\gamma+1})$ suggested by the theory (marked with star) indeed leads to near-optimal performance of two-timescale OMWU.

Figure 2 shows $\mathsf{KL}\big(\pi_\tau^\star \,\|\, \pi^{(t)}\big)$ with respect to the number of iterations of single-timescale and two-timescale OMWU under different asynchronous scenarios, with optimal choices of η and $\tau = 0.1$. In particular, two-timescale OMWU adopts the extrapolation learning rate suggested by theory $\bar{\eta} = \tau^{-1}(1 - (1 - \eta\tau)^{\gamma+1})$. While both methods yield linear convergence to the QRE, two-timescale method outperforms its single-timescale counterpart in the case with constant and known delay and the case where the feedback is permutated with bounded delay, which verifies our theory.



(a) (b) (c)

Figure 1: $\mathsf{KL}\big(\pi_\tau^\star \,\|\, \pi^{(t)}\big)$ of single-timescale and two-timescale OMWU with respect to different values of learning rate and delay. (a): performance of single-timescale OMWU in the synchronous setting and asynchronous settings. (b) & (c): performance of the two methods after 5000 iterations under various choices of η and $\overline{\eta}$, with $\bar{\eta}$ fixed to $\bar{\eta} = \tau^{-1}(1 - (1 - \eta\tau)^{\gamma+1})$ in (b) and η fixed to 0.001 in (c).



(a) random delay (b) fixed delay (c) permuted feedback

Figure 2: $\mathsf{KL}\big(\pi_\tau^\star \,\|\, \pi^{(t)}\big)$ with respect to iteration count t of single-timescale and two-timescale OMWU under various asynchronous settings. (a): random delays bounded by $\gamma = 25$. (b): constant delays $\gamma = 50$. (c): permuted feedback with delay bounded by $\gamma = 25$.

6 Conclusions

This paper studies asynchronous gradient play in zero-sum polymatrix games, by investigating the convergence behaviors of entropy-regularized OMWU with delayed feedbacks under two different schedules of the learning rates. We demonstrate that the single-timescale OMWU enjoys a linear last-iterate convergence to the QRE even under mild statistical delays, thus enables robust learning of gradient play in a delay-oblivious manner. We further show that the method benefits from adopting a two-timescale learning rate, by achieving a faster convergence when the delay is fixed and known and provable convergence in a more general setting with permuted feedback and bounded delay.

REFERENCES

Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. *Advances in neural information processing systems*, 24, 2011.

Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On last-iterate convergence beyond zero-sum games. *arXiv preprint arXiv:2203.12056*, 2022.

Mahmoud Assran, Arda Aytekin, Hamid Reza Feyzmahdavian, Mikael Johansson, and Michael G Rabbat. Advances in asynchronous parallel and distributed optimization. *Proceedings of the IEEE*, 108(11):2013–2031, 2020.

James P Bailey. $O(1/T)$ time-average convergence in a generalization of multiagent zero-sum games. *arXiv preprint arXiv:2110.02482*, 2021.

LM Bergman and IN Fokin. On separable non-cooperative zero-sum games. *Optimization*, 44(1): 69–84, 1998.

Riccardo Bernardi. Interactive image segmentation using graph transduction games. 2021.

Dimitri P Bertsekas and John N Tsitsiklis. Parallel and distributed computation: numerical methods, 1989.

Yang Cai, Ozan Candogan, Constantinos Daskalakis, and Christos Papadimitriou. Zero-sum polymatrix games: A generalization of minmax. *Mathematics of Operations Research*, 41(2):648–655, 2016.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. *arXiv preprint arXiv:2105.15186*, 2021.

Nicol'o Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pp. 605–622. PMLR, 2016.

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

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*, pp. 235–254. SIAM, 2011.

Michail Fasoulakis, Evangelos Markakis, Yannis Pantazis, and Constantinos Varsos. Forward looking best-response multiplicative weights update methods for bilinear zero-sum games. In *International Conference on Artificial Intelligence and Statistics*, pp. 11096–11117. PMLR, 2022.

Genevieve E Flaspohler, Francesco Orabona, Judah Cohen, Soukayna Mouatadid, Miruna Oprescu, Paulo Orenstein, and Lester Mackey. Online learning with optimism and delay. In *International Conference on Machine Learning*, pp. 3363–3373. PMLR, 2021.

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, et al. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the eighth international workshop on data mining for online advertising*, pp. 1–9, 2014.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Multi-agent online optimization with delays: Asynchronicity, adaptivity, and optimism. *arXiv preprint arXiv:2012.11579*, 2020.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pp. 1453–1461. PMLR, 2013.

Chung-Wei Lee, Christian Kroer, and Haipeng Luo. Last-iterate convergence in extensive-form games. *Advances in Neural Information Processing Systems*, 34, 2021.

Stefanos Leonardos, Georgios Piliouras, and Kelly Spendlove. Exploration-exploitation in multi-agent competition: convergence with bounded rationality. *Advances in Neural Information Processing Systems*, 34:26318–26331, 2021.

Bingcong Li, Tianyi Chen, and Georgios B Giannakis. Bandit online learning with unknown delays. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 993–1002. PMLR, 2019.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings*, pp. 157–163. Elsevier, 1994.

Ji Liu and Stephen J Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, 25(1):351–376, 2015.

Ji Liu, Steve Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. In *International Conference on Machine Learning*, pp. 469–477. PMLR, 2014.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

Brendan McMahan and Matthew Streeter. Delay-tolerant algorithms for asynchronous distributed online learning. *Advances in Neural Information Processing Systems*, 27, 2014.

Min Meng, Xiuxian Li, and Jie Chen. Decentralized nash equilibria learning for online game with bandit feedback. *arXiv preprint arXiv:2204.09467*, 2022.

Panayotis Mertikopoulos and William H Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.

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*, pp. 2703–2717. SIAM, 2018.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 1497–1507. PMLR, 2020.

Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pp. 4105–4113. PMLR, 2018.

Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. *Advances in neural information processing systems*, 28, 2015.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *arXiv preprint arXiv:1311.1869*, 2013.

Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in neural information processing systems*, 24, 2011.

Samuel Sokota, Ryan D'Orazio, J Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. *arXiv preprint arXiv:2206.05825*, 2022.

Qianqian Tong, Guannan Liang, and Jinbo Bi. Effective federated adaptive gradient methods with non-iid decentralized data. *arXiv preprint arXiv:2009.06557*, 2020.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. *arXiv preprint arXiv:1706.09186*, 2017.

Yuanyu Wan, Wei-Wei Tu, and Lijun Zhang. Online strongly convex optimization with unknown delays. *Machine Learning*, 111(3):871–893, 2022.

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. *arXiv preprint arXiv:2102.04540*, 2021.

Marcelo J Weinberger and Erik Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, 2002.

Xin Zhang, Jia Liu, and Zhengyuan Zhu. Taming convergence for asynchronous stochastic gradient descent with unbounded delay in non-convex learning. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 3580–3585. IEEE, 2020.