Optimal Learning in Games Under Delayed Feedback

Ruotong Zhuang
The University of Tokyo

Taira Tsuchiya
The University of Tokyo & RIKEN

Shinji Ito
The University of Tokyo & RIKEN

Abstract

Learning in games is a central topic in both learning theory and game theory, and learning dynamics based on online learning have made significant theoretical and practical advances in recent years. In particular, in two-player zero-sum games, it has been shown that using optimistic follow-the-regularized-leader (OFTRL) as the online learning algorithm allows us to upper bound the individual regret for each player by a constant independent of the time horizon. However, in realistic game scenarios, players are not always able to observe the outcomes of their interactions immediately. Motivated by this, very recently, the problem of learning from delayed feedback in games has been proposed, and it has been shown that by using a variant of OFTRL, one can achieve a social regret upper bound of $\widetilde{O}(D^2 + 1)$ for a fixed delay time D. This study investigates the optimal dependence on the delay parameter D in the setting of learning from delayed feedback in games. In particular, we show that a simple algorithm that runs $D + 1$ independent copies of the standard OFTRL designed for the non-delayed setting achieves social and individual regret upper bounds of $\widetilde{O}(D + 1)$, thereby improving the existing bounds by a factor of D. Moreover, we provide a matching lower bound: for any learning dynamic, there exists a payoff matrix such that the regret of every player is at least $\Omega(D + 1)$.

1 INTRODUCTION

Learning in games studies the dynamics that arise when multiple players interact in a shared environment, each aiming to maximize cumulative reward. In particular, its connection to game-theoretic equilibria has been extensively investigated, and it has found applications in minimax optimization (Freund and Schapire, 1999), multi-agent reinforcement learning (Bai et al., 2020; Wei et al., 2021), and, more recently, in the development of superhuman-level AI systems (Bowling et al., 2015; Moravčík et al., 2017) as well as the post-training of large language models (Munos et al., 2024).

One of the most representative approaches to learning equilibria is to employ the framework of *online learning*, where players minimize cumulative losses based on past observations (Freund and Schapire, 1999; Hart and Mas-Colell, 2000; Cesa-Bianchi and Lugosi, 2006). In the game-theoretic context, the most standard online learning algorithm is the Hedge algorithm (also known as multiplicative weights update or simply the exponential weight algorithm) (Littlestone and Warmuth, 1994; Freund and Schapire, 1997). The Hedge algorithm selects actions according to weights that are exponentially scaled by the negative cumulative losses, and it is known to achieve $\widetilde{O}(\sqrt{T})$ regret in the worst case for the number of rounds T.

This upper bound holds for worst-case environments, whereas in learning in games, it is known that the regrets of the players can be significantly improved if each of them employs an online learning algorithm with *optimistic predictions*. In particular, a representative algorithm is optimistic Hedge (Rakhlin and Sridharan, 2013b; Syrgkanis et al., 2015), which selects actions according to weights that are exponentially scaled by the negative cumulative losses together with the most recent loss. In two-player zero-sum games, when the learning rate of optimistic Hedge is set to an absolute constant, each player's individual regret can be bounded by $\widetilde{O}(1)$ external regret, independent of the number of rounds T. In multiplayer general-sum games, it is also known that, with an appropriate choice of regularizer in optimistic follow-the-regularized-leader (OFTRL), which is a generalization of optimistic Hedge, both external regret and swap regret can be bounded by $O(\log T)$, ignoring dependencies on variables other than T (Anagnostides et al., 2022a,b).

Despite this significant progress, all of these studies assume that the outcomes of the game interactions can be observed immediately. In real-world scenarios, however, this assumption does not always hold, and feedback may be ob-

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

Table 1: Individual regret upper and lower bounds in two-player zero-sum games under delayed feedback. The honest regime refers to the setting where all players follow a prescribed algorithm, while the corrupted regime refers to situations where the opponent deviates from the prescribed algorithm or where the observations are corrupted by adversarial noise. "# of updates" refers to the number of updates of optimistic Hedge. The variable T denotes the number of rounds, D the length of the observation delay in rounds, and $C \in [0, 6T]$ the total degree of the opponent's deviation.

References	Honest regime	Adversarial scenario	Corrupted regime	# of updates
Fujimoto et al. (2025)	$\widetilde{O}(D^2+1)$	–	–	$\Theta(T)$
This work (Theorem 2)	$\widetilde{O}(D+1)$	$\widetilde{O}(\sqrt{(D+1)T})$	$\widetilde{O}(\sqrt{(D+1)T})$	$\Theta(T)$
This work (Theorem 6)	$\widetilde{O}(D+1)$	$\widetilde{O}(\sqrt{(D+1)T})$	$\widetilde{O}(D+1+\sqrt{(D+1)C})$	$\Theta(T)$
This work (Theorem 7)	$\widetilde{O}(D+1)$	$\Omega(T)$	$\Omega(T)$	$\Theta(T/(D+1))$
Lower bounds	$\Omega(D+1)$ (**This work, Theorem 8**)	$\Omega(\sqrt{(D+1)T})$	–	Not applicable

served only after a certain delay. Such a *delayed feedback* setting has in fact been extensively studied in the contexts of online learning and bandit problems (Weinberger and Ordentlich, 2002; Joulani et al., 2013; Cesa-Bianchi et al., 2016; Joulani et al., 2016; Thune et al., 2019; Ito et al., 2020; Zimmert and Seldin, 2020; van der Hoeven et al., 2023; Shibukawa et al., 2025).

Recently, Fujimoto et al. (2025) introduced a framework for delayed feedback in the context of learning in games. They consider a fixed-delay setting in which each player's observations are received with a delay of exactly D rounds. By carefully designing the optimistic prediction within optimistic Hedge, they show that it is possible to achieve a social regret upper bound of $O(D^2+1)$, which is independent of the number of rounds T. However, this dependence on the delay time D through a quadratic factor appears highly suboptimal. Moreover, their analysis does not provide an explicit upper bound on individual regret: individual regret bounds can be derived for their learning dynamic, but the dependence on T remains $T^{1/4}$, leaving a substantial gap compared with the recent $O(\log T)$ bounds (Anagnostides et al., 2022a). Further related work that could not be included in the main text can be found in Appendix A.

Contributions of This Paper

In this study, to develop a more detailed understanding of learning dynamics in games under delayed feedback, we investigate the optimal dependence on the delay parameter D, and make the following contributions that address the limitations of the existing work mentioned above.

We begin by focusing on the case of two-player zero-sum games in order to highlight the key idea. Our algorithm adopts a well-known approach from online learning with delayed feedback, which runs $(D+1)$ independent online learning algorithms designed for the non-delayed setting (Weinberger and Ordentlich, 2002; Joulani et al., 2013). We employ optimistic Hedge as the algorithm for the non-delayed setting. We then show that this learning dynamic achieves $O(D+1)$ social and individual regret upper bounds, improving the regret bounds of Fujimoto et al. (2025) by a factor of D.

Furthermore, we investigate in detail the scenario where the opponent does not follow the prescribed algorithm, namely optimistic Hedge, which was not investigated by Fujimoto et al. (2025). We first show that by monitoring the second-order variation of the gain vectors induced by the opponent's chosen strategies and switching an algorithm with a worst-case regret of $\widetilde{O}(\sqrt{(D+1)T})$ appropriately, it is possible to simultaneously guarantee performance of $\widetilde{O}(\sqrt{(D+1)T})$ even in adversarial scenarios.

However, as pointed out by Tsuchiya et al. (2025), such an algorithm exhibits undesirable behavior in that even a slight deviation by the opponent can deteriorate the regret from constant to $O(\sqrt{T})$. To address this, we employ optimistic Hedge with an adaptive learning rate as the OFTRL algorithm for the non-delayed setting, and, in the *corrupted regime* (Tsuchiya et al., 2025) (where the opponent's degree of deviation is quantified), we establish a regret upper bound of $O(D+1+\sqrt{(D+1)C})$, where C denotes the cumulative deviation of the opponent. This result recovers that of Tsuchiya et al. (2025) when $D = 0$. We further discuss an approach to improve the time and space complexities of these learning dynamics by a factor of D, at the cost of sacrificing adversarial robustness, which is particularly useful for dynamics using complex regularizers for OFTRL (*e.g.,* Farina et al. 2022).

We then demonstrate that the above regret upper bounds with linear dependence on the delay D are in fact optimal by providing individual regret lower bounds. Specifically, we show that for any learning dynamic, there exist utility functions under which the individual regret of *every player* is at least $\Omega(D+1)$. A summary of these results is provided in Table 1.

Finally, we discuss how the learning dynamic running $(D+1)$ copies of non-delayed optimistic Hedge can also be applied to external and swap regret minimization in multi-player general-sum games. In particular, by combining the above approach with the existing OFTRL framework using a log-barrier regularizer (Anagnostides et al., 2022b; Tsuchiya et al., 2025), we show that both external regret and swap regret can be bounded by $O((D+1)\log T)$, ignoring

dependencies on variables other than D and T. This substantially improves upon the individual regret upper bound with $T^{1/4}$ dependence obtained by the learning dynamic of Fujimoto et al. (2025).

2 PRELIMINARIES

In this section, we introduce the basic concepts used throughout this paper. We begin by describing the problem setting of online linear optimization, which is the class of online learning algorithms employed by each player in games. We then present the setting of learning in two-player zero-sum games under delayed feedback. We then explain the relationship between Nash equilibrium and regret minimization in online learning. Finally, we introduce algorithms based on optimistic prediction that are used in learning in games.

Notation For a natural number $n \in \mathbb{N}$, we write $[n] = \{1, \dots, n\}$ and $[n]_+ = \{0, \dots, n\}$. We denote by $\mathbf{0}$ and $\mathbf{1}$ the all-zero and all-one vectors, respectively. For a vector x, the notation $x(i)$ refers to its i-th coordinate, and $\|x\|_p$ denotes its ℓ_p-norm for any $p \in [1, \infty]$. The $(m-1)$-dimensional probability simplex is defined as $\Delta_m = \{x \in [0,1]^m \colon \|x\|_1 = 1\}$. We use the order notation $\widetilde{O}(\cdot)$ to indicate asymptotic bounds up to logarithmic factors in the number of actions.

2.1 Online Linear Optimization

In online linear optimization, a player is given a convex set $\mathcal{K} \subseteq \mathbb{R}^m$ before the game starts. Then at each round $t = 1, 2, \dots, T$,

1. The player selects a point $w^t \in \mathcal{K}$ based on observations up to round $t-1$;
2. The environment determines a loss vector h^t;
3. The player suffers a loss of $\langle w^t, h^t \rangle \in \mathbb{R}$ and observes the loss vector h^t.

The most common objective for the player is to minimize the cumulative losses, which is equivalent to minimizing the (external) regret Reg^T given by

$$\mathsf{Reg}^T = \max_{u \in \mathcal{K}} \mathsf{Reg}^T(u), \quad \mathsf{Reg}^T(u) = \sum_{t=1}^{T} \langle w^t - u, h^t \rangle .$$

An online linear optimization algorithm is said to be *no-regret* if its regret satisfies $\mathsf{Reg}^T = o(T)$.

2.2 Two-Player Zero-Sum Games under Delayed Feedback

Here we describe the problem setting of two-player zero-sum games under delayed feedback (Fujimoto et al., 2025). The problem is characterized by a payoff matrix $A \in [-1,1]^{m_x \times m_y}$ and a fixed-delay parameter D, where m_x and m_y are the number of actions of the x- and y-players, respectively. The procedure of this game is as follows: at each round $t = 1, 2, \dots, T$,

1. The x-player selects a strategy $x^t \in \Delta_{m_x}$ and the y-player selects $y^t \in \Delta_{m_y}$ based on their own observations up to round $t-1$;
2. The x-player gains a payoff of $\langle x^t, g^t \rangle$ for an expected gain vector $g^t = Ay^t \in [-1,1]^{m_x}$ and the y-player incurs a loss of $\langle y^t, \ell^t \rangle$ for an expected loss vector $\ell^t = A^\top x^t \in [-1,1]^{m_y}$;
3. The x-player observes g^{t-D} and the y-player observes ℓ^{t-D} if $t - D \geq 1$;

In what follows, we assume that $T \geq D+1$ and $D \geq 1$. The goal of the x-player and the y-player is to maximize the cumulative gain and to minimize the cumulative loss, respectively. As in online linear optimization, this is equivalent to minimizing the corresponding (external) regrets Reg_x^T and Reg_y^T, defined as follows: $\mathsf{Reg}_x^T = \max_{x^* \in \Delta_{m_x}} \mathsf{Reg}_T^x(x^*)$ and $\mathsf{Reg}_y^T = \max_{y^* \in \Delta_{m_y}} \mathsf{Reg}_T^y(y^*)$ for $\mathsf{Reg}_x^T(x^*) = \sum_{t=1}^T \langle x^* - x^t, Ay^t \rangle = \sum_{t=1}^T \langle x^* - x^t, g^t \rangle$ and $\mathsf{Reg}_y^T(y^*) = \sum_{t=1}^T \langle y^t - y^*, A^\top x^t \rangle = \sum_{t=1}^T \langle y^t - y^*, \ell^t \rangle$. From this formulation, each player in learning in two-player zero-sum games under delayed feedback can be viewed as performing online linear optimization over the probability simplex, with the additional constraint that observations are received with a delay of D rounds.

2.3 Nash Equilibrium and No-Regret Learning Dynamics

Let (x^*, y^*) be probability distributions over the action sets $[m_x]$ and $[m_y]$. We call (x^*, y^*) an *ε-approximate Nash equilibrium* for $\varepsilon \geq 0$ if, for every choice of $x \in \Delta_{m_x}$ and $y \in \Delta_{m_y}$, the following inequalities are satisfied: $\langle x, Ay^* \rangle - \varepsilon \leq \langle x^*, Ay^* \rangle \leq \langle x^*, Ay \rangle + \varepsilon$. The pair (x^*, y^*) is a *Nash equilibrium* if it is a 0-approximate Nash equilibrium.

It is a classical result that no-regret learning dynamics lead to approximate equilibria when all players employ no-regret algorithms (Freund and Schapire, 1999). In particular, if each player follows a no-regret strategy, then the empirical distribution of play, namely $(\frac{1}{T}\sum_{t=1}^T x^t, \frac{1}{T}\sum_{t=1}^T y^t)$ is a $((\mathsf{Reg}_x^T + \mathsf{Reg}_y^T)/T)$-approximate Nash equilibrium. Note that the sum of the regrets, $\mathsf{Reg}_x^T + \mathsf{Reg}_y^T$, is called the *social regret*.

In the standard online linear optimization setting with a fixed delay D and loss vectors $h^t \in [-1,1]^m$, the regret upper bound of $\mathsf{Reg}^T = O(\sqrt{(D+1)T \log m})$ is known (*e.g.,* Joulani et al. 2013). Therefore, by employing such an online linear optimization algorithm, one can guarantee a convergence rate to a Nash equilibrium of $\widetilde{O}(\sqrt{(D+1)/T})$. In the next section, we show that this convergence rate can be significantly improved to $\widetilde{O}((D+$

$1)/T)$.

2.4 Optimistic FTRL and Optimistic Hedge

In learning in games, it is well known that online learning algorithms based on optimistic prediction are particularly useful (Rakhlin and Sridharan, 2013b; Syrgkanis et al., 2015). Here, we introduce the framework of optimistic follow-the-regularized-leader (OFTRL) and one of its notable instances, the optimistic Hedge algorithm.

Optimistic FTRL At each round $t \in [T]$, OFTRL selects $w^t \in \mathcal{K}$ by

$$w^t \in \arg\min_{w \in \mathcal{K}} \left\{ \left\langle w, m^t + \sum_{s=1}^{t-1} h^s \right\rangle + \psi^t(w) \right\}, \quad (1)$$

where ψ^t is a convex regularizer over $\mathcal{K}$ and $m^t \in \mathbb{R}^m$ is an optimistic prediction of the true loss vector h^t. The optimistic prediction m^t is an estimate of h^t based on the information available up to time $t-1$, and the closer it is to h^t, the smaller the resulting regret can be (Rakhlin and Sridharan, 2013a). In the setting of learning in games without delay, it is common to use $m^t = h^{t-1}$. We use the convention $g^t = \mathbf{0}$ and $\ell^t = \mathbf{0}$ for all $t \le 0$, and the optimistic prediction of each copy is initialized to zero before it receives its first observation.

Optimistic Hedge Optimistic Hedge is an instance of OFTRL that uses the negative Shannon entropy $\psi^t(w) = \frac{1}{\eta^t}\sum_{i=1}^m w(i)\log(w(i))$ as its regularizer. Here, for simplicity, we assume that the feasible set is the probability simplex $\mathcal{K} = \Delta_m$. In this case, the OFTRL update rule in (1) can be written in closed form as follows: for each $i \in [m]$,

$$\begin{aligned} w^t(i) &= \frac{v^t(i)}{\sum_{j\in[m]} v^t(j)}, \\ v^t(i) &= \exp\left(-\eta^t\left(m^t(i) + \sum_{s=1}^{t-1} h^s(i)\right)\right), \end{aligned} \quad (2)$$

It is known that optimistic Hedge achieves the following RVU bound (Regret bounded by Variation in Utilities):

Lemma 1. *Suppose that $(w^t)_{t=1}^T$ is generated by the optimistic Hedge algorithm with nonincreasing learning rate $\eta^1 \ge \cdots \ge \eta^T$. Then,*

$$\mathsf{Reg}^T \le \frac{\log m}{\eta^{T+1}} + \sum_{t=1}^T \eta^t \|h^t - m^t\|_\infty^2 - \sum_{t=2}^T \frac{1}{4\eta^t}\|w^t - w^{t-1}\|_1^2.$$

The analysis of OFTRL with a constant learning rate is standard and can be found in Rakhlin and Sridharan (2013a); Syrgkanis et al. (2015). The proof of Lemma 1, which uses the time-varying (or adaptive) learning rate, can be found in Tsuchiya et al. (2025, Lemma 17).

Optimistic Hedge for delayed feedback In the delayed setting, when determining the strategy at round t, only the information up to round $t-D-1$ is available, and therefore updates of the form given in (1) and (2) cannot be performed. A standard approach in such cases is to mimic the updates of (1) and (2) using only the information that has already been received (*e.g.,* Zimmert and Seldin 2020; van der Hoeven et al. 2023). In the context of games, Fujimoto et al. (2025) consider the following optimistic-Hedge-based update with the following definition of v^t in (3)[1]:

$$v^t(i) = \exp\left(-\eta\left(m^t(i) + \sum_{s=1}^{t-D-1} h^s(i)\right)\right),\ m^t = (D+1)h^{t-D-1}.$$

Note that the summation extends only up to $t-D-1$. In other words, their algorithm can be seen as an OFTRL variant that relies on the most recently observed feedback, where the latest observation h^{t-D-1} is treated as a good optimistic prediction for both the past D rounds and the next round of $t+1$. However, as discussed above, this approach yields only an $\widetilde{O}(D^2+1)$ individual regret upper bound in two-player zero-sum games to our knowledge. The next section shows that this bound can be improved by a factor of D using another algorithm for dealing with delayed feedback.

3 REGRET UPPER BOUNDS

In this section, we first show that a learning dynamic based on running $(D+1)$ independent copies of the optimistic Hedge algorithm for the non-delayed setting achieves $\widetilde{O}(D+1)$ social and individual regret upper bounds. We then discuss how the same learning dynamic can guarantee $\widetilde{O}(\sqrt{DT})$ regret in adversarial scenarios where the opponent may deviate from the output of optimistic Hedge. Finally, we provide a learning dynamic that achieves a regret bound of $O(D+1+\sqrt{(D+1)C})$ with respect to the opponent's deviation level $C \in [0, 6T]$.

3.1 Proposed Learning Dynamic

Our learning dynamic employs a standard algorithm for online learning under delayed feedback. Specifically, we employ a simple yet well-known approach in the online learning literature: running $(D+1)$ independent copies of optimistic Hedge for the non-delayed setting (Weinberger and Ordentlich, 2002; Joulani et al., 2013).

We present our learning dynamic in Algorithm 1. Here we only describe the algorithm for the x-player (the y-player's algorithm is defined analogously). First, the algorithm prepares $(D+1)$ copies of optimistic Hedge with a constant learning rate $\eta^t = \eta$, denoted by $(\mathsf{Alg}_x(d))_{d=0}^D$. Each $\mathsf{Alg}_x(d)$ for $d \in [D]_+ = \{0, 1, \ldots, D\}$ is responsible for

[1] They consider a general regularizer with 1-strong convexity with respect to $\|\cdot\|_1$, but it suffices to focus only on OFTRL with the negative Shannon entropy, that is, optimistic Hedge.

Algorithm 1 Optimal learning dynamic in two-player zero-sum games under delayed feedback

(# Algorithm for the x-player)
Initialize $(D+1)$ independent copies of optimistic Hedge algorithms $(\mathsf{Alg}_x(d))_{d=0}^{D}$.
for $t = 1, 2, \ldots, T$ **do**
  Compute $d = t \bmod (D+1)$.
  Compute a strategy $x^t \in \Delta_{m_x}$ by $\mathsf{Alg}_x(d)$.
  Observe a gain vector $g^{t-D} = Ay^{t-D} \in [-1,1]^{m_x}$.
  Update $\mathsf{Alg}_x((t-D) \bmod (D+1))$ based on past observations.

(# Algorithm for the y-player)
Initialize $(D+1)$ independent copies of optimistic Hedge algorithms $(\mathsf{Alg}_y(d))_{d=0}^{D}$.
for $t = 1, 2, \ldots, T$ **do**
  Compute $d = t \bmod (D+1)$.
  Compute a strategy $y^t \in \Delta_{m_y}$ by $\mathsf{Alg}_y(d)$.
  Observe a loss vector $\ell^{t-D} = A^\top x^{t-D} \in [-1,1]^{m_y}$.
  Update $\mathsf{Alg}_y((t-D) \bmod (D+1))$ based on past observations.

the rounds in $[T] = \{1, \ldots, T\}$ given by

$$\mathcal{T}_d = \{t \in [T] : t \bmod (D+1) = d\}.$$

Then, for each $d \in [D]_+$ and $t \in \mathcal{T}_d$, the x-player determines its strategy as follows: for each $i \in [m_x]$,

$$\begin{aligned} x^t(i) &= \frac{v_x^t(i)}{\sum_{j\in[m_x]} v_x^t(j)}, \\ v_x^t(i) &= \exp\left(\eta_x^t\left(g^{t-(D+1)}(i) + \sum_{s\in\mathcal{T}_d : s<t} g^s(i)\right)\right). \end{aligned} \tag{3}$$

We will use η_y^t to denote the learning rate of the y-player at round t. Note that this x^t (and the corresponding y^t) can be computed in our setting where the observations are delayed by D rounds.

3.2 Individual and Social Regret Upper Bounds

With the learning dynamic in Algorithm 1, the individual regrets can be upper bounded as follows:

Theorem 2. *With the learning dynamic in Algorithm 1 with $\eta_x^t = \eta_y^t = \eta = 1/4$, the regret of each player is upper bounded as*

$$\begin{aligned} \mathsf{Reg}_x^T &\le (D+1)\left(\frac{4}{3}\log(m_x m_y) + 4\log m_x + 1\right), \\ \mathsf{Reg}_y^T &\le (D+1)\left(\frac{4}{3}\log(m_x m_y) + 4\log m_y + 1\right). \end{aligned}$$

As a consequence, this result implies a social regret upper bound of $\widetilde{O}(D)$, improving the social regret bound of Fujimoto et al. (2025) by a factor of D. As we will see in the next section, this regret upper bound is optimal.

3.3 Proof of Theorem 2

Here we provide the proof of Theorem 2. For notational simplicity, let $\tau_t = t - (D+1)$ for each $t \in [T]$. Define the regrets of $\mathsf{Alg}_x(d)$ and $\mathsf{Alg}_y(d)$ by

$$\begin{aligned} \mathcal{R}_x(d) &= \max_{x\in\Delta_{m_x}} \sum_{t\in\mathcal{T}_d} \langle x, g^t\rangle - \sum_{t\in\mathcal{T}_d} \langle x^t, g^t\rangle \\ \mathcal{R}_y(d) &= \sum_{t\in\mathcal{T}_d} \langle y^t, \ell^t\rangle - \min_{y\in\Delta_{m_y}} \sum_{t\in\mathcal{T}_d} \langle y, \ell^t\rangle, \end{aligned} \tag{4}$$

where we recall $\mathcal{T}_d = \{t \in [T] : t \bmod (D+1) = d\}$.

From Lemma 1, we immediately obtain the following lemma:

Lemma 3. *Define $\mathcal{T}_d^- = \mathcal{T}_d \setminus \{\min \mathcal{T}_d\}$. Then, for each $d \in [D]_+$, the regret of each base algorithm is upper bounded as*

$$\begin{aligned} \mathcal{R}_x(d) &\le \frac{\log m_x}{\eta} + \eta \sum_{t\in\mathcal{T}_d} \|g^t - g^{\tau_t}\|_\infty^2 - \frac{1}{4\eta} \sum_{t\in\mathcal{T}_d^-} \|x^t - x^{\tau_t}\|_1^2, \\ \mathcal{R}_y(d) &\le \frac{\log m_y}{\eta} + \eta \sum_{t\in\mathcal{T}_d} \|\ell^t - \ell^{\tau_t}\|_\infty^2 - \frac{1}{4\eta} \sum_{t\in\mathcal{T}_d^-} \|y^t - y^{\tau_t}\|_1^2, \end{aligned}$$

where we recall $\tau_t = t - (D+1)$.

Lemma 4. *For each $d \in [D]_+$, it holds that $\mathcal{R}_x(d) + \mathcal{R}_y(d) \ge 0$.*

Proof. Let (x, y) be a Nash equilibrium of the payoff matrix A. Then, we can lower bound $\mathcal{R}_x(d) + \mathcal{R}_y(d)$ as

$$\begin{aligned} &\mathcal{R}_x(d) + \mathcal{R}_y(d) \\ &\ge \sum_{t\in\mathcal{T}_d} \langle x - x^t, Ay^t\rangle + \sum_{t\in\mathcal{T}_d} \langle y^t - y, A^\top x^t\rangle \\ &= \sum_{t\in\mathcal{T}_d} \langle x, Ay^t\rangle - \sum_{t\in\mathcal{T}_d} \langle y, A^\top x^t\rangle \\ &= \sum_{t\in\mathcal{T}_d} \langle x, Ay^t - Ay\rangle - \sum_{t\in\mathcal{T}_d} \langle y, A^\top x^t - A^\top x\rangle \ge 0, \end{aligned}$$

where the last inequality follows since the product distribution (x, y) is the Nash equilibrium of A. □

Lemma 5. *For any $\eta \in (0, 1/2)$ and $d \in [D]_+$, it holds that*

$$\begin{aligned} \mathcal{R}_x(d) &\le \frac{\log m_x}{\eta} + 4\eta + \frac{\log(m_x m_y)}{\frac{1}{4\eta} - \eta}, \\ \mathcal{R}_y(d) &\le \frac{\log m_y}{\eta} + 4\eta + \frac{\log(m_x m_y)}{\frac{1}{4\eta} - \eta}. \end{aligned}$$

Proof. Combining Lemma 3 with the inequalities $\|g^t - g^{\tau_t}\|_\infty = \|A(y^t - y^{\tau_t})\|_\infty \le \|y^t - y^{\tau_t}\|_1$ and $\|\ell^t - \ell^{\tau_t}\|_\infty =$

$\|A^\top(x^t - x^{\tau_t})\|_\infty \le \|x^t - x^{\tau_t}\|_1$, we have

$$\mathcal{R}_x(d) \le \frac{\log m_x}{\eta} + \eta \sum_{t\in\mathcal{T}_d} \|y^t - y^{\tau_t}\|_1^2 - \frac{1}{4\eta} \sum_{t\in\mathcal{T}_d^-} \|x^t - x^{\tau_t}\|_1^2 , \tag{5}$$

$$\mathcal{R}_y(d) \le \frac{\log m_y}{\eta} + \eta \sum_{t\in\mathcal{T}_d} \|x^t - x^{\tau_t}\|_1^2 - \frac{1}{4\eta} \sum_{t\in\mathcal{T}_d^-} \|y^t - y^{\tau_t}\|_1^2 . \tag{6}$$

Summing up the last two upper bounds, we have

$$\begin{aligned}
&\mathcal{R}_x(d) + \mathcal{R}_y(d) \\
&\le \frac{\log(m_x m_y)}{\eta} + 4\eta \\
&\quad + \left(\eta - \frac{1}{4\eta}\right) \sum_{t\in\mathcal{T}_d^-} \left(\|x^t - x^{\tau_t}\|_1^2 + \|y^t - y^{\tau_t}\|_1^2\right).
\end{aligned}$$

Hence, using Lemma 4 and the assumption that $\eta \in (0, 1/2)$,

$$\sum_{t\in\mathcal{T}_d^-} \left(\|x^t - x^{\tau_t}\|_1^2 + \|y^t - y^{\tau_t}\|_1^2\right) \le \frac{\log(m_x m_y) + 4\eta^2}{\eta\left(\frac{1}{4\eta} - \eta\right)} . \tag{7}$$

Finally, combining (5) and (6) with (7), we obtain the desired upper bounds. □

Finally, we are ready to prove Theorem 2.

Proof of Theorem 2. From the definitions of the regrets,

$$\begin{aligned}
\mathsf{Reg}_x^T &= \max_{x\in\Delta_{m_x}} \sum_{t=1}^T \langle x, g^t\rangle - \sum_{t=1}^T \langle x^t, g^t\rangle \\
&\le \sum_{d\in[D]_+} \max_{x\in\Delta_{m_x}} \sum_{t\in\mathcal{T}_d} \langle x - x^t, g^t\rangle = \sum_{d\in[D]_+} \mathcal{R}_x(d) .
\end{aligned}$$

Similarly, we also have $\mathsf{Reg}_y^T \le \sum_{d\in[D]_+} \mathcal{R}_y(d)$. Combining these inequalities with Lemma 5, we complete the proof of Theorem 2. □

3.4 Adversarial Robustness and Adaptive Bounds

Here, we discuss how the above learning dynamic can handle adversarial scenarios in which the opponent may deviate from the outputs of optimistic Hedge.

Common approach In the non-delayed setting, the most well-known approach is to monitor the observed vectors induced by the opponent's strategies, and to switch to an algorithm that guarantees a sublinear regret in the worst case once a certain threshold is exceeded. We show that a similar approach can also be applied in our setting of delayed feedback.

From (7) with $\eta = 1/4$, if both players completely follow Algorithm 1, for each $d \in [D]_+$ we have

$$\begin{aligned}
&\max\left\{ \sum_{t\in\mathcal{T}_d^-} \|g^t - g^{\tau_t}\|_\infty^2 , \sum_{t\in\mathcal{T}_d^-} \|\ell^t - \ell^{\tau_t}\|_\infty^2 \right\} \\
&\le \sum_{t\in\mathcal{T}_d^-} \left(\|x^t - x^{\tau_t}\|_1^2 + \|y^t - y^{\tau_t}\|_1^2\right) \\
&\le \frac{16}{3}\left(\log(m_x m_y) + \frac{1}{4}\right).
\end{aligned}$$

From this observation, the x-player may monitor $M_x(d) = \sum_{t\in\mathcal{T}_d^-} \|g^t - g^{\tau_t}\|_\infty^2$, and if for some $d \in [D]_+$ it holds that $M_x(d) > (16/3)(\log(m_x m_y) + 1/4)$, then switch its learning rate in (3) to $\Theta(\sqrt{(D+1)\log m_x / T})$. This guarantees that as long as $M_x(d) \le (16/3)(\log(m_x m_y) + 1/4)$, the individual regret upper bound of Theorem 2 is maintained, while otherwise it still ensures the regret bound of $\mathsf{Reg}_x^T = O(\sqrt{DT\log m_x})$ simultaneously. The y-player can similarly monitor $M_y(d) = \sum_{t\in\mathcal{T}_d^-} \|\ell^t - \ell^{\tau_t}\|_\infty^2$, and by applying the same procedure, robustness in adversarial scenarios can be guaranteed.

Adaptive approach In the above approach, once the inequality that $M_x(d) > (16/3)(\log(m_x m_y) + 1/4)$ is satisfied for some $d \in [D]_+$ due to the opponent's deviation, the individual regret deteriorates from constant to $\sqrt{T}$, which is undesirable behavior. Moreover, such methods that exploit the property of the strategy's squared path-length to guarantee adversarial robustness implicitly assume knowledge of the opponent's number of actions, which fails to preserve the strong-uncoupledness of the learning dynamics.

To address this, we consider achieving regret upper bounds and convergence rates that adapt to the degree of deviation of each player based on OFTRL with *adaptive* learning rate, as recently investigated by Tsuchiya et al. (2025). They considered the following more general problem setup. Under delayed feedback, their setting can be formulated as follows. At each round $t = 1, 2, \dots, T$,

1. Prescribed algorithms (*e.g.,* optimistic Hedge) suggest strategies $\widehat{x}^t \in \Delta_{m_x}$ and $\widehat{y}^t \in \Delta_{m_y}$ to each player;
2. The x-player selects a strategy $x^t \leftarrow \widehat{x}^t + \widehat{c}_x^t \in \Delta_{m_x}$ and the y-player selects $y^t \leftarrow \widehat{y}^t + \widehat{c}_y^t \in \Delta_{m_y}$;
3. The x-player gains a payoff of $\langle x^t, g^t\rangle$ for $g^t = Ay^t$ and the y-player incurs a loss of $\langle y^t, \ell^t\rangle$ for $\ell^t = A^\top x^t$;
4. The x-player observes $\widetilde{g}^{t-D} = g^{t-D} + \widetilde{c}_x^t \in [-1, 1]^{m_x}$ and the y-player observes $\widetilde{\ell}^{t-D} = \ell^{t-D} + \widetilde{c}_y^t \in [-1, 1]^{m_y}$ if $t - D \ge 1$;

Here, $\widehat{c}_x^t$ and $\widetilde{c}_x^t$ are corruption vectors for strategies and observation of the x-player at round t, respectively, and can depend on their own observations up to round $t-1$. Note that it holds that $\sum_{t=1}^T \|\widehat{c}_x^t\|_1 = \sum_{t=1}^T \|x^t - \widehat{x}^t\|_1 \le \widehat{C}_x$, $\sum_{t=1}^T \|\widetilde{c}_x^t\|_\infty = \sum_{t=1}^T \|g^t - \widetilde{g}^t\|_\infty \le \widetilde{C}_x$, and $C_x = \widehat{C}_x +$

$2\widetilde{C}_x$. Similarly, $\widehat{c}_y^t$ and $\widetilde{c}_y^t$ are corruption levels of strategies and observations of the y-player, respectively, and satisfy $\sum_{t=1}^T \|\widehat{c}_y^t\|_1 = \sum_{t=1}^T \|y^t - \widehat{y}^t\|_1 \le \widehat{C}_y$, $\sum_{t=1}^T \|\widetilde{c}_y^t\|_\infty = \sum_{t=1}^T \|\ell^t - \widetilde{\ell}^t\|_\infty \le \widetilde{C}_y$, and $C_y = \widehat{C}_y + 2\widetilde{C}_y$.

In this regime, a time-varying (or adaptive) learning rate is a promising approach. We choose η_x^t in (3) (and the corresponding learning rate η_y^t for the y-player) as follows: For each $d \in [D]_+$ and $t \in \mathcal{T}_d$,

$$\eta_x^t = \sqrt{\frac{\log_+(m_x)/2}{\log_+(m_x) + \sum_{s \in \mathcal{T}_d : s < t} \|\widetilde{g}^s - \widetilde{g}^{\tau_s}\|_\infty^2}}, \quad \eta_y^t = \sqrt{\frac{\log_+(m_y)/2}{\log_+(m_y) + \sum_{s \in \mathcal{T}_d : s < t} \|\widetilde{\ell}^s - \widetilde{\ell}^{\tau_s}\|_\infty^2}}, \tag{8}$$

where $\log_+(z) = \max\{\log z, 4\}$ and $\widetilde{g}^t = \widetilde{\ell}^t = \mathbf{0}$ for $t \le 0$. This choice of learning rate allows us to prove the following bounds:

Theorem 6. *With the learning dynamic in Algorithm 1 with η_x^t, η_y^t in (8), the regret of each player satisfies*

$$\mathsf{Reg}_x^T = \widetilde{O}\left((D+1) + \sqrt{(D+1)(C_x + C_y)} + C_x\right),$$
$$\mathsf{Reg}_y^T = \widetilde{O}\left((D+1) + \sqrt{(D+1)(C_x + C_y)} + C_y\right).$$

This result implies that if a player completely follows the prescribed algorithm while the opponent deviates, the regret can be bounded by $O(D + 1 + \sqrt{(D+1)C})$, where $C \in [0, 6T]$ denotes the degree of deviation of the opponent. This bound interpolates between the $\widetilde{O}(D+1)$ upper bound when all players follow the prescribed algorithm and the $\widetilde{O}(\sqrt{(D+1)T})$ upper bound in adversarial scenarios. Moreover, when there is no delay ($D = 0$), this result (partially) recovers the result of Tsuchiya et al. (2025) for two-player zero-sum games.

3.5 Reducing Space and Time Complexities

By exploiting the fact that we are in a game setting rather than a general online learning setting, we can improve the space and time complexities of the learning dynamic in Algorithm 1 by a factor of D, at the cost of sacrificing the adversarial robustness discussed above. In this learning dynamic, each player maintains a single instance of the optimistic Hedge algorithm. The algorithm is updated only when $t \in \mathcal{T}_1 \setminus \{1\}$, using the gradients observed in $\{s \in \mathcal{T}_1 \setminus \{1\} : s < t\}$, and otherwise simply outputs the same strategy as in the previous round. This learning dynamic is summarized in Algorithm 2. (The algorithm for the y-player is omitted, but it is defined analogously.)

This learning dynamic achieves the following regret upper bound.

Theorem 7. *With the learning dynamic in Algorithm 2 with $\eta = 1/4$, the regret of each player is upper bounded as $\mathsf{Reg}_x^T = \widetilde{O}(D+1)$ and $\mathsf{Reg}_y^T = \widetilde{O}(D+1)$.*

Algorithm 2 Efficient learning dynamic in two-player zero-sum games under delayed feedback

(# Algorithm for the x-player)
Initialize an optimistic Hedge algorithm.
Let $x^0 = \frac{1}{m_x}\mathbf{1}$.
for $t = 1, 2, \ldots, T$ **do**
  Compute $d = t \bmod (D+1)$.
  if $t > 1$ and $d = 1$ **then**
    Compute a strategy $x^t \in \Delta_{m_x}$ by $x^t(i) \propto \exp\big(\eta\big(g^{t-(D+1)}(i) + \sum_{s \in \mathcal{T}_1 : s < t} g^s(i)\big)\big)$.
  else
    Choose $x^t = x^{t-1}$.
  Observe a gain vector $g^{t-D} = Ay^{t-D} \in [-1, 1]^{m_x}$.

This regret upper bound can be proven by noting that, as long as the opponent follows the prescribed algorithm, the same output is repeated for $D+1$ consecutive rounds, and by applying an analysis analogous to that in Theorem 2. In contrast to Algorithm 1, Algorithm 2 requires each player to maintain only a single instance of optimistic Hedge, and the number of updates is reduced by a factor of $D+1$. Still, as mentioned above, this approach has the drawback that it cannot guarantee robustness against deviations by the opponent. However, in more complex games such as convex games, OFTRL can employ sophisticated regularizers (*e.g.,* Farina et al. 2022), in which case this becomes a significant advantage.

4 REGRET LOWER BOUNDS

In this section, we provide an $\Omega(D+1)$ regret lower bound for two-player zero-sum games under delayed feedback, thereby showing that the $\widetilde{O}(D+1)$ upper bound in Section 3 is optimal with respect to D. When feedback is delayed by D rounds, it is natural to expect that the social regret, *i.e.,* the sum of all players' regrets, is $\Omega(D)$. By carefully constructing an appropriate payoff matrix, we further show that the individual regret of each player is also $\Omega(D)$.

Theorem 8. *Consider the problem of learning in two-player zero-sum games. Then, for any learning dynamic, there exists a payoff matrix $A \in [-1, 1]^{2\times 2}$ such that the regret of each player is lower bounded by*

$$\mathbb{E}\left[\mathsf{Reg}_x^T\right] \ge \frac{D}{2}, \quad \mathbb{E}\left[\mathsf{Reg}_y^T\right] \ge \frac{D}{2},$$

where the expectations are taken with respect to the internal randomness of the learning dynamic.

Proof. Let $c = 1/2$ and let $\xi_x, \xi_y \in \{-1, 1\}$ be constants specified later, and consider the payoff matrix $A \in [-1, 1]^{2\times 2}$ given by

$$A = c\,\xi_x(e_1 - e_2)\mathbf{1}^\top + c\,\xi_y \mathbf{1}(e_1 - e_2)^\top = c\begin{pmatrix} \xi_x + \xi_y & \xi_x - \xi_y \\ -\xi_x + \xi_y & -\xi_x - \xi_y \end{pmatrix}. \tag{9}$$

Then, for any $x \in \Delta_{m_x}$ and $y \in \Delta_{m_y}$, it holds that

$$\begin{aligned}\langle x, Ay\rangle &= c\xi_x x^\top (e_1 - e_2)\mathbf{1}^\top y + c\xi_y x^\top \mathbf{1}(e_1 - e_2)^\top y \\ &= c\xi_x x^\top (e_1 - e_2) + c\xi_y (e_1 - e_2)^\top y \\ &= c\xi_x (x(1) - x(2)) + c\xi_y (y(1) - y(2))\,, \quad (10)\end{aligned}$$

where we used $\mathbf{1}^\top y = x^\top \mathbf{1} = 1$. From this observation, the comparator $x^* \in \arg\max_{x\in\Delta_{m_x}} \sum_{t=1}^T \langle x, Ay^t\rangle$ satisfies $x^*(1) - x^*(2) = \xi_x$. Hence, the regret of the row player is evaluated as

$$\begin{aligned}\mathsf{Reg}_x^T &= \max_{x\in\Delta_{m_x}} \sum_{t=1}^T \langle x, Ay^t\rangle - \sum_{t=1}^T \langle x^t, Ay^t\rangle \\ &= cT + \sum_{t=1}^T c\xi_y (y^t(1) - y^t(2)) \\ &\quad - \sum_{t=1}^T \big(c\xi_x (x^t(1) - x^t(2)) + c\xi_y (y^t(1) - y^t(2))\big) \\ &= cD - c\xi_x \sum_{t=1}^D (x^t(1) - x^t(2)) \\ &\quad + c(T - D) - c\xi_x \sum_{t=D+1}^T (x^t(1) - x^t(2)) \\ &\geq cD - c\xi_x \sum_{t=1}^D (x^t(1) - x^t(2))\,.\end{aligned}$$

Since no feedback is observed before round $D+1$, the random variables $x^1, \dots, x^D$ and $y^1, \dots, y^D$ are independent of A, and hence independent of ξ_x and ξ_y. Therefore, choosing

$$\xi_x = -\mathrm{sgn}\left(\mathbb{E}\left[\sum_{t=1}^D (x^t(1) - x^t(2))\right]\right)$$

in the last inequality, we obtain $\mathbb{E}[\mathsf{Reg}_x^T] \geq cD = \frac{D}{2}$ as desired. For the regret of the column player, we can follow a similar analysis as above by choosing $\xi_y = \mathrm{sgn}(\mathbb{E}[\sum_{t=1}^D (y^t(1) - y^t(2))])$, we obtain the desired lower bound, and we have completed the proof. □

5 EXTENSION TO MULTIPLAYER GENERAL-SUM GAMES

In the preceding sections, we have focused only on simple two-player zero-sum games in order to convey the core ideas more clearly. Due to the simplicity of the learning dynamic, similar results can be readily obtained for more general multiplayer general-sum games. In this section, we begin by introducing the problem setting of multiplayer general-sum games. We then discuss how, by combining the ideas from Section 3 with learning dynamics based on OFTRL with a log-barrier regularizer (Anagnostides et al., 2022b), one can obtain the $O((D+1)\log T)$ regret upper bound.

5.1 Multiplayer General-Sum Games under Delayed Feedback

Here we define the setting of multiplayer general-sum games under delayed feedback. Let $n \geq 2$ denote the number of players, and write $[n] = \{1, \dots, n\}$ for the player set. Each player $i \in [n]$ is equipped with m_i actions and a utility function $u_i \colon [m_1] \times \cdots \times [m_n] \to [-1, 1]$. We define the *expected utility* $u_i^t \in [-1,1]^{m_i}$ as the vector whose a_i-th component is $u_i^t(a_i) = \mathbb{E}_{a_{-i}\sim x_{-i}^t}[u_i(a_i, a_{-i})]$, where $x_{-i}^t = (x_1^t, \dots, x_{i-1}^t, x_{i+1}^t, \dots, x_n^t)$.

The procedure of the game is as follows: At each round $t = 1, 2, \dots, T$,

1. Each player $i \in [n]$ chooses a strategy $x_i^t \in \Delta_{m_i}$;
2. Each player i gains a payoff of $\langle x_i^t, u_i^t\rangle$;
3. If $t - D \geq 1$, each player i observes an expected utility $u_i^{t-D} \in [-1,1]^{m_i}$.

As in the case of two-player zero-sum games, a natural goal is to minimize the external regret given by $\mathsf{Reg}_i^T = \max_{x_i^*\in\Delta_{m_i}} \mathsf{Reg}_i^T(x_i^*)$ for $\mathsf{Reg}_i^T(x_i^*) = \sum_{t=1}^T \langle x_i^* - x_i^t, u_i^t\rangle$. It is well known that in multiplayer general-sum games, if each player minimizes the external regret, an (approximate) coarse correlated equilibrium (CCE) is obtained (see *e.g.,* Roughgarden 2016 for the relationship between this and the following regret notions and the corresponding equilibrium concepts in multiplayer general-sum games). Since a CCE is somewhat undesirable, much research has focused on minimizing *swap regret* in order to obtain the more desirable (approximate) correlated equilibrium. The swap regret of player $i \in [n]$ is given by $\mathsf{SwapReg}_i^T = \max_{M\in\mathcal{M}_{m_i}} \mathsf{SwapReg}_i^T(M)$ for $\mathsf{SwapReg}_i^T(M) = \sum_{t=1}^T \langle x_i^t, Mu_i^t - u_i^t\rangle$. Here $\mathcal{M}_m = \{M \in [0,1]^{m\times m} \colon M(k,\cdot) \in \Delta_m \text{ for } k \in [m]\}$ is the set of all $m \times m$ row stochastic matrices. From the definitions, it holds that $\mathsf{Reg}_i^T \leq \mathsf{SwapReg}_i^T$.

5.2 Learning Dynamics and Regret Upper Bounds

In two-player zero-sum games, the proposed learning dynamic runs optimistic Hedge separately on each subset of rounds $\mathcal{T}_d$. Analogously, in multiplayer general-sum games, one can assign each $\mathcal{T}_d$ to the state-of-the-art swap regret minimization algorithm Anagnostides et al. (2022b); Tsuchiya et al. (2025). These approaches reduce swap regret minimization to external regret minimization over an expanded action space (Blum and Mansour, 2007), and implement each external regret minimizer using OFTRL in (1) with a log-barrier regularizer (see the papers on swap regret minimization in games for details). This learning dynamic allows us to prove the following upper bounds on external and swap regret:

Proposition 9. *Consider the problem of multiplayer general-sum games. Then, there exists a learning dynamic*

such that for each player $i \in [n]$,

$$\mathsf{Reg}_i^T \le \mathsf{SwapReg}_i^T = O\left(nm^{5/2}(D+1)\log\left(\frac{T}{D+1}\right)\right),$$

where $m = \max_{i\in[n]} m_i$.

This improves upon the results of Fujimoto et al. (2025) in two ways. Although not stated explicitly, the algorithm of Fujimoto et al. (2025) has an external regret upper bound with a $T^{1/4}$ dependence on T. In contrast, our approach improves this dependence to logarithmic in T. Moreover, we provide logarithmic regret upper bounds not only for external regret but also for swap regret, which is much more challenging to upper bound.

6 NUMERICAL EXPERIMENTS

We numerically compare the proposed learning dynamic with the WOFTRL-based dynamic of Fujimoto et al. (2025) in two-player zero-sum games. The code for the experiments is publicly available on GitHub[2]. Additional experimental results are provided in Appendix C.

Setup For each instance, we generated a 2×2 payoff matrix A by drawing each entry independently from $\mathrm{Unif}([-1,1])$ and then normalizing the matrix so that $\max_{i,j}|A(i,j)| = 1$. For the proposed method, we considered learning rates $\eta \in \{0.025, 0.05, 0.1, 0.25\}$, and for WOFTRL, we used the theoretically prescribed learning rate $\eta = 1/(\sqrt{2}(D+1)(D+2))$. To evaluate the dependence on the delay parameter, we measured the social and individual regrets at $T = 200{,}000$ while varying the delay D from 0 to 200 in increments of 40. To examine how the regrets evolve over time, we plotted the regret trajectories up to $T = 50,000$ for a fixed delay $D = 20$.

Results The upper-left and lower panels of Figure 1 show the social and individual regrets as a function of the delay D. The social regret of the proposed method grew approximately linearly in D, which was consistent with the theoretical $\Theta(D+1)$ dependence. In contrast, the social regret of WOFTRL increased much faster, empirically supporting the suboptimality of its $O(D^2+1)$ dependence on delay. The individual regrets of both players exhibited the same linear scaling in D under the proposed learning dynamic. The upper-right panel of Figure 1 shows the time evolution of the regrets for $D = 20$. The social regret of the proposed method quickly stabilized and remained bounded as T increased, providing empirical evidence for the $\widetilde{O}(1)$ dependence on the time horizon.

7 CONCLUSION

In this study, we investigated the optimal dependence on the delay parameter D in the setting of learning from delayed feedback in games. First, in two-player zero-sum games, we showed that by preparing $(D+1)$ independent copies of optimistic Hedge for the non-delayed feedback setting, one can achieve a regret upper bound of $\widetilde{O}(D+1)$. In doing so, we also provided a detailed discussion of adversarial robustness, which has not been sufficiently addressed in prior work. We further provided a dynamic method that improves computational efficiency by a factor of D, at the cost of sacrificing adversarial robustness. We then established the optimality of this dependence on D by proving individual regret lower bounds of $\Omega(D+1)$ for all players. Finally, using an approach analogous to that for two-player zero-sum games, we derived $O((D+1)\log T)$ external and swap regret upper bounds, thereby significantly improving upon the existing $T^{1/4}$ dependence on T.

One promising direction for future work is to investigate learning dynamics under more complex delayed feedback structures. A particularly interesting direction for future research is to investigate regret upper bounds in the setting where the delay may differ across players and is unknown to each player. Note that, if each player knows their own delay in advance, then they can simply wait until the corresponding feedback arrives, which yields an $O(\max\{D_x, D_y\}+1)$ regret bound, where D_x and D_y denote the delays of the x-player and the y-player, respectively.

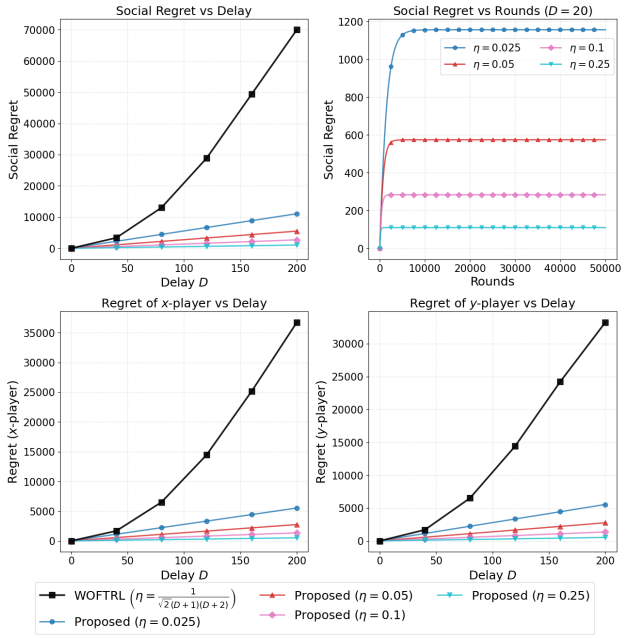

Figure 1: Empirical comparison of the proposed learning dynamic against WOFTRL. Top left: final social regret at $T = 200,000$ as a function of the delay D. Top right: social regret trajectory over rounds up to $T = 50,000$ for $D = 20$. Bottom left and bottom right: final individual regrets of the x-player and y-player, respectively, as a function of D.

[2] https://github.com/zhuangrt/Optimal-Learning-in-Games-under-Delayed-Feedback

Acknowledgments

TT is supported by JSPS KAKENHI Grant Number JP24K23852, and SI is supported by JSPS KAKENHI Grant Number JP25K03184 and by JST PRESTO, Grant Number JPMJPR2511.

References

Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 736–749. Association for Computing Machinery, 2022a.

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. In *Advances in Neural Information Processing Systems*, volume 35, pages 3292–3304. Curran Associates, Inc., 2022b.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. In *Advances in Neural Information Processing Systems*, volume 33, pages 2159–2170. Curran Associates, Inc., 2020.

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(47): 1307–1324, 2007.

Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold' em poker is solved. *Science*, 347(6218):145–149, 2015.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49, pages 605–622. PMLR, 2016.

Gabriele Farina, Ioannis Anagnostides, Haipeng Luo, Chung-Wei Lee, Christian Kroer, and Tuomas Sandholm. Near-optimal no-regret learning dynamics for general convex games. In *Advances in Neural Information Processing Systems*, volume 35, pages 39076–39089. Curran Associates, Inc., 2022.

Genevieve E. Flaspohler, Francesco Orabona, Judah Cohen, Soukayna Mouatadid, Miruna Oprescu, Paulo Orenstein, and Lester Mackey. Online learning with optimism and delay. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 3363–3373. PMLR, 2021.

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.

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

Yuma Fujimoto, Abe Kenshi, and Kaito Ariu. Learning from delayed feedback in games via extra prediction. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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

Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Delay and cooperation in nonstochastic linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pages 4872–4883. Curran Associates, Inc., 2020.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 1453–1461. PMLR, 2013.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, pages 1744–1750, 2016.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and Computation*, 108 (2):212–261, 1994.

Naresh Manwani and Mudit Agarwal. Delaytron: Efficient learning of multiclass classifiers with delayed bandit feedbacks. In *2023 International Joint Conference on Neural Networks*, pages 1–10, 2023.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 11752–11762. Curran Associates, Inc., 2022.

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.

Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 36743–36768. PMLR, 2024.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 993–1019, 2013a.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, volume 26, pages 3066–3074. Curran Associates, Inc., 2013b.

Tim Roughgarden. *Twenty Lectures on Algorithmic Game Theory*. Cambridge University Press, 2016.

Yuki Shibukawa, Taira Tsuchiya, Shinsaku Sakaue, and Kenji Yamanishi. Bandit and delayed feedback in online structured prediction. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, volume 28, pages 2989–2997. Curran Associates, Inc., 2015.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, volume 32, pages 6541–6550. Curran Associates, Inc., 2019.

Taira Tsuchiya, Shinji Ito, and Haipeng Luo. Corrupted learning dynamics in games. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291, pages 5506–5552. PMLR, 2025.

Dirk van der Hoeven, Lukas Zierahn, Tal Lancewicki, Aviv Rosenberg, and Nicolò Cesa-Bianchi. A unified analysis of nonstochastic delayed feedback for combinatorial semi-bandits, linear bandits, and MDPs. In *Proceedings of the 36th Conference on Learning Theory*, volume 195, pages 1285–1321. PMLR, 2023.

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 *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 4259–4299. PMLR, 2021.

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

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pages 3285–3294. PMLR, 2020.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

   The problem setting is detailed in Section 2. The complexity analysis is omitted since the algorithms are the well-known optimistic Hedge and optimistic FTRL. The source code for the numerical experiments in the appendix is included in the supplemental material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

   The assumptions are fully provided for each proposition. The complete proofs are fully provided in Sections 3 and 4 and the appendix.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

   The source code for the experiments in the appendix is included in the supplemental material. The experimental setup is fully described in Section 6 and the appendix. Since we are considering a deterministic algorithm, error bars are not necessary. The computing infrastructure used for the experiments is described in the appendix.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

   This study is primarily theoretical and did not use existing assets.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

   This study is primarily theoretical and did not use crowdsourcing or conduct research with human subjects.

Optimal Learning in Games Under Delayed Feedback: Supplementary Materials

A ADDITIONAL RELATED WORK

Here we review the literature on delayed feedback in online learning that could not be covered in the main text.

In contrast to the game-theoretic literature, delayed feedback has attracted substantial attention in the contexts of online learning and bandit problems. To the best of our knowledge, research on delayed feedback in these settings began with Weinberger and Ordentlich (2002). Subsequent research in the full-information setting (*i.e.,* when gradient feedback is available) has been extensively studied (Joulani et al. 2013, 2016; Flaspohler et al. 2021, to name a few). In this setting, the optimal regret in terms of the delay parameter D and the number of rounds T is known to be $\widetilde{\Theta}(\sqrt{DT})$. Of particular relevance to this paper is the approach of Weinberger and Ordentlich (2002); Joulani et al. (2013), which maintains $D+1$ independent copies of online learning algorithms. By contrast, the approach of Fujimoto et al. (2025) can be viewed as related to approaches that run follow-the-regularized-leader based only on observed gradients (*e.g.,* Flaspohler et al. 2021).

Delayed feedback has also been studied for multi-armed bandits, as well as in more structured settings such as linear bandits and Markov decision processes (Cesa-Bianchi et al., 2016; Zimmert and Seldin, 2020; Ito et al., 2020; Masoudian et al., 2022; van der Hoeven et al., 2023), often motivated by applications such as online advertising, where outcomes are not observed immediately. It is also worth noting that, more recently, delayed feedback has been investigated in online bandit classification (Manwani and Agarwal, 2023) and ins generalized formulation of bandit structured prediction (Shibukawa et al., 2025). While this paper focuses on the full-information setting, leveraging these insights to explore delayed feedback under bandit feedback, where each player observes only the payoff associated with the actions actually taken, is a promising direction for future work. Several studies have also explored settings with stochastic and adversarial delays that vary across rounds, as well as situations in which certain observations experience delays as large as $\Omega(T)$. Investigating such regimes is another direction for future work.

B OMITTED PROOFS

This section provides the omitted proofs.

B.1 Proof of Theorem 6

Here we provide the proof of Theorem 6.

Proof of Theorem 6. For each $d \in [D]_+$, define $\widehat{C}_x(d) = \sum_{t\in\mathcal{T}_d}\|\widehat{c}_x^t\|_1$, $\widetilde{C}_x(d) = \sum_{t\in\mathcal{T}_d}\|\widetilde{c}_x^t\|_\infty$, and $C_x(d) = \widehat{C}_x(d) + 2\widetilde{C}_x(d)$. Similarly, define $\widehat{C}_y(d) = \sum_{t\in\mathcal{T}_d}\|\widehat{c}_y^t\|_1$, $\widetilde{C}_y(d) = \sum_{t\in\mathcal{T}_d}\|\widetilde{c}_y^t\|_\infty$, and $C_y(d) = \widehat{C}_y(d) + 2\widetilde{C}_y(d)$. Then, from Tsuchiya et al. (2025, Theorem 10), we have

$$\mathcal{R}_x(d) = O\left(\sqrt{(\log(m_x m_y) + C_x(d) + C_y(d))\log m_x} + C_x(d)\right),$$
$$\mathcal{R}_y(d) = O\left(\sqrt{(\log(m_x m_y) + C_x(d) + C_y(d))\log m_y} + C_y(d)\right).$$

Since

$$\mathsf{Reg}_x^T = \max_{x\in\Delta_{m_x}}\sum_{t=1}^T\langle x, g^t\rangle - \sum_{t=1}^T\langle x^t, g^t\rangle \le \sum_{d\in[D]_+}\max_{x\in\Delta_{m_x}}\sum_{t\in\mathcal{T}_d}\langle x - x^t, g^t\rangle = \sum_{d\in[D]_+}\mathcal{R}_x(d),$$

Algorithm 3 Learning dynamic in multiplayer general-sum games under delayed feedback

(# Algorithm for player $i \in [n]$)
Initialize $(D+1)$ independent copies of optimistic FTRL with a log-barrier regularizer $(\mathsf{Alg}_i(d))_{d=0}^{D}$ in Anagnostides et al. (2022b) or Tsuchiya et al. (2025).
for $t = 1, 2, \dots, T$ **do**
  Compute $d = t \bmod (D+1)$.
  Compute a strategy $x_i^t \in \Delta_{m_i}$ by $\mathsf{Alg}_i(d)$.
  Observe a utility vector $u_i^{t-D} \in [-1,1]^{m_i}$.
  Update $\mathsf{Alg}_i((t-D) \bmod (D+1))$ based on past observations.

we have

$$\begin{aligned}\mathsf{Reg}_x^T &= O\left(\sum_{d\in[D]_+}\left(\sqrt{(\log(m_x m_y) + C_x(d) + C_y(d))\log m_x} + C_x(d)\right)\right)\\ &\le O\left(\sqrt{(D+1)\sum_{d\in[D]_+}(\log(m_x m_y) + C_x(d) + C_y(d))\log m_x} + C_x\right)\\ &\le O\left((D+1)\log(m_x m_y) + \sqrt{(D+1)(C_x+C_y)\log m_x} + C_x\right),\end{aligned}$$

where the second line follows from the Cauchy–Schwarz inequality. This is the desired upper bound for the regret of the x-player. The regret of the y-player can be upper bounded by the same argument, and we completed the proof of Theorem 6. □

B.2 Proof of Theorem 7

Here we provide the proof of Theorem 7.

Proof of Theorem 7. Note that all players use the same strategy within each set $\{t, \dots, t+D\}$ for every $t \in \mathcal{T}_1$. Then,

$$\mathsf{Reg}_x^T \le (D+1)\left(\max_{x\in\Delta_{m_x}}\sum_{t\in\mathcal{T}_1}\langle x, g^t\rangle - \sum_{t\in\mathcal{T}_1}\langle x^t, g^t\rangle\right) \le (D+1)\left(\frac{4}{3}\log(m_x m_y) + 4\log m_x + 16\right),$$

where the last inequality can be proven by exactly the same argument as in the proof of Theorem 2. The regret of the y-player can be upper bounded by the same argument. □

B.3 Proof of Proposition 9

Here we provide the missing details from Section 5 and the proof of Proposition 9. We consider the learning dynamics summarized in Algorithm 3.

Proof of Proposition 9. For each player $i \in [n]$, define the swap regret of $\mathsf{Alg}_i(d)$ by

$$\mathcal{S}_i(d) = \max_{M\in\mathcal{M}_{m_i}}\sum_{t\in\mathcal{T}_d}\langle x_i^t, Mu_i^t - u_i^t\rangle,$$

where we recall $\mathcal{T}_d = \{t \in [T] \colon t \bmod (D+1) = d\}$. Then, from Anagnostides et al. (2022b, Corollary 4.5) or Tsuchiya et al. (2025, Theorem 13), for each $d \in [D]_+$ we have

$$\mathcal{S}_i(d) = O\left(nm^{5/2}\log\left(1 + \left\lfloor\frac{T}{D+1}\right\rfloor\right)\right)$$

for $m = \max_{i\in[n]} m_i$. Hence, using this upper bound, we have

$$\begin{aligned}\mathsf{SwapReg}_i^T &= \max_{M\in\mathcal{M}_{m_i}} \sum_{d\in[D]_+}\sum_{t\in\mathcal{T}_d}\langle x_i^t, Mu_i^t - u_i^t\rangle \\ &\le \sum_{d\in[D]_+}\max_{M\in\mathcal{M}_{m_i}}\sum_{t\in\mathcal{T}_d}\langle x_i^t, Mu_i^t - u_i^t\rangle = \sum_{d\in[D]_+}\mathcal{S}_i(d) \\ &\le O\left(nm^{5/2}(D+1)\log\left(1+\left\lfloor\frac{T}{D+1}\right\rfloor\right)\right),\end{aligned}$$

which is the desired bound. □

As can be seen from these analyses, the approach of preparing $(D+1)$ independent copies (Weinberger and Ordentlich, 2002; Joulani et al., 2013) can be directly applied to more general classes of games, for example, to convex games (Farina et al., 2022). Moreover, although Proposition 9 considers only the setting where all players fully follow the prescribed OFTRL algorithm with a log-barrier regularizer, by employing the adaptive learning rate for swap regret minimization (Tsuchiya et al., 2025), it is possible to obtain regret guarantees that adapt to deviations by the opponents as in Theorem 6.

C ADDITIONAL NUMERICAL EXPERIMENTS

This appendix provides additional plots that complement the results in Section 6. Unless otherwise stated, all experimental settings are the same as those in Section 6. We report results across multiple random seeds and additional delay settings not included in the main text. Figure 2 shows the final social regret at $T = 200,000$ for eight random seeds. Figures 3 and 4 show the corresponding individual regrets for the x- and the y-players. Figures 5 to 7 show regret trajectories up to $T = 50,000$ for $D \in \{5, 10, 15, 20\}$. These additional plots are consistent with the trends reported in Section 6.

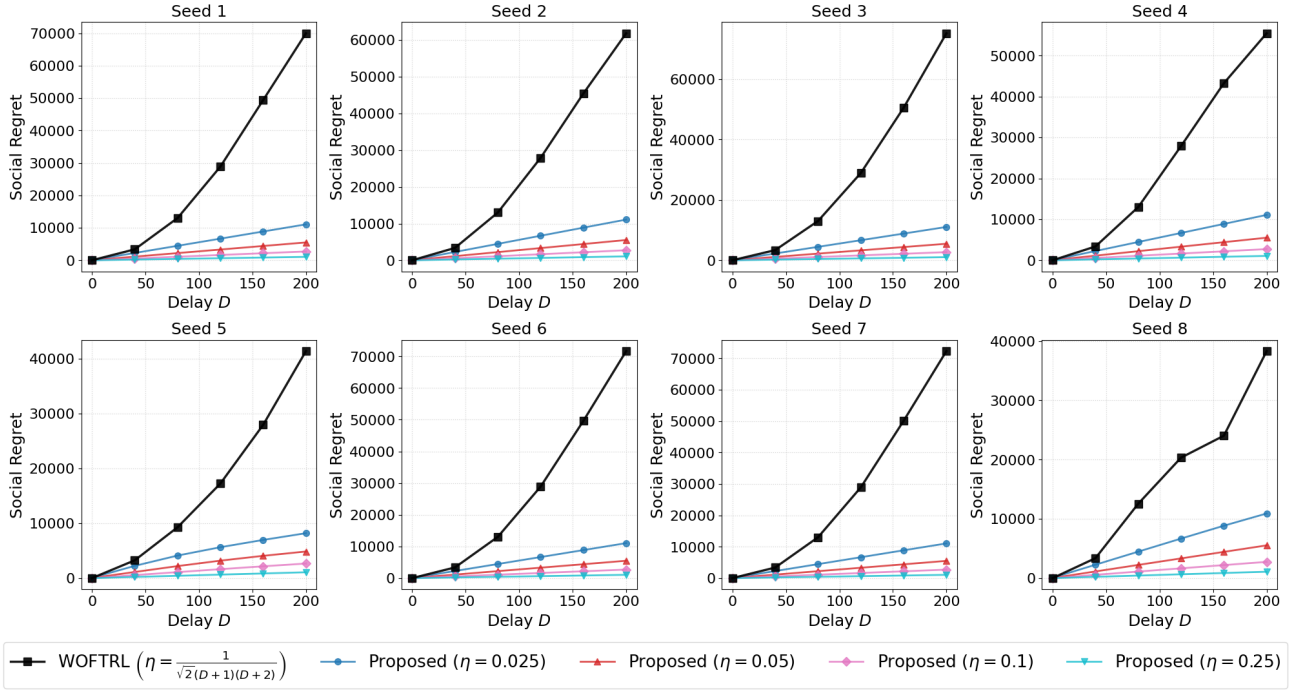

Figure 2: Final social regret (vertical axis) as a function of the delay D (horizontal axis) at $T = 200,000$ for eight different random seeds. The colored lines correspond to the proposed learning dynamic with different learning rates, and the black line corresponds to the learning dynamic based on the weighted optimistic follow-the-regularized-leader (WOFTRL).

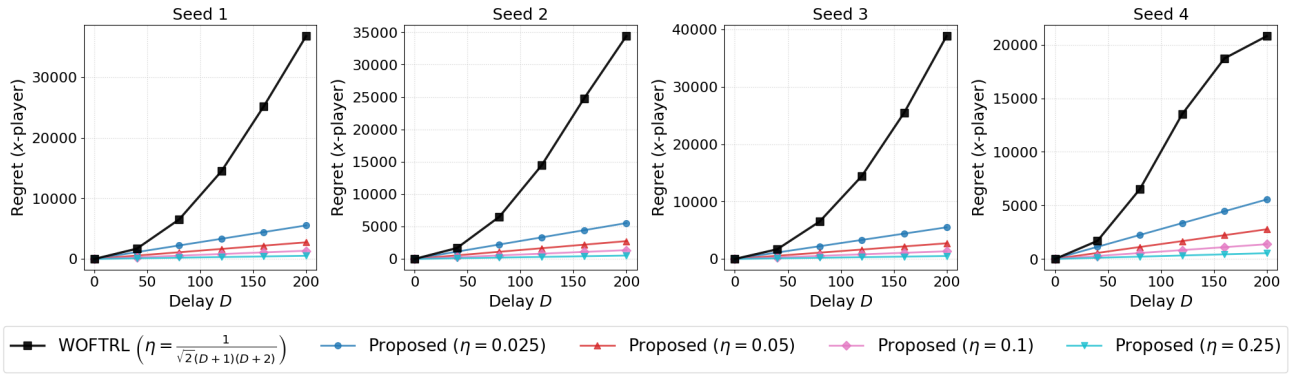

Figure 3: The individual regret of the x-player as a function of the delay D.

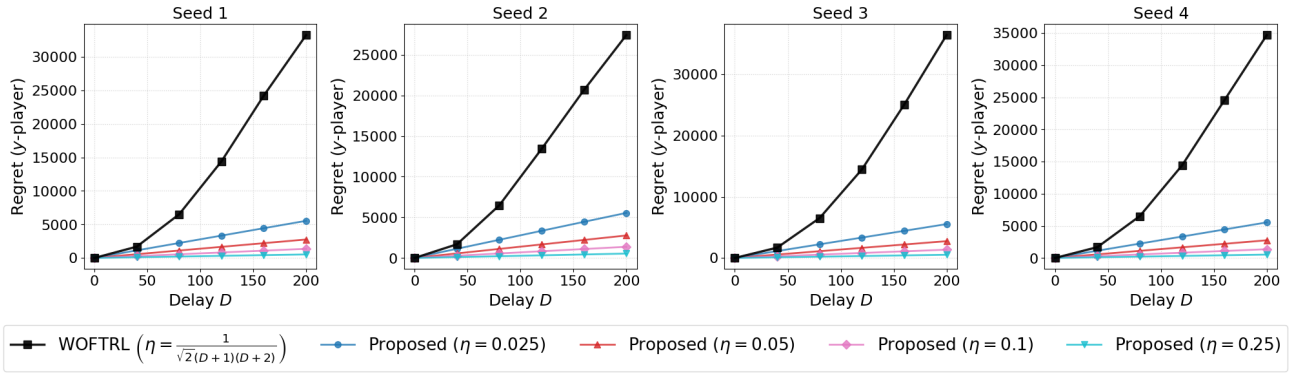

Figure 4: The individual regret of the y-player as a function of the delay D.

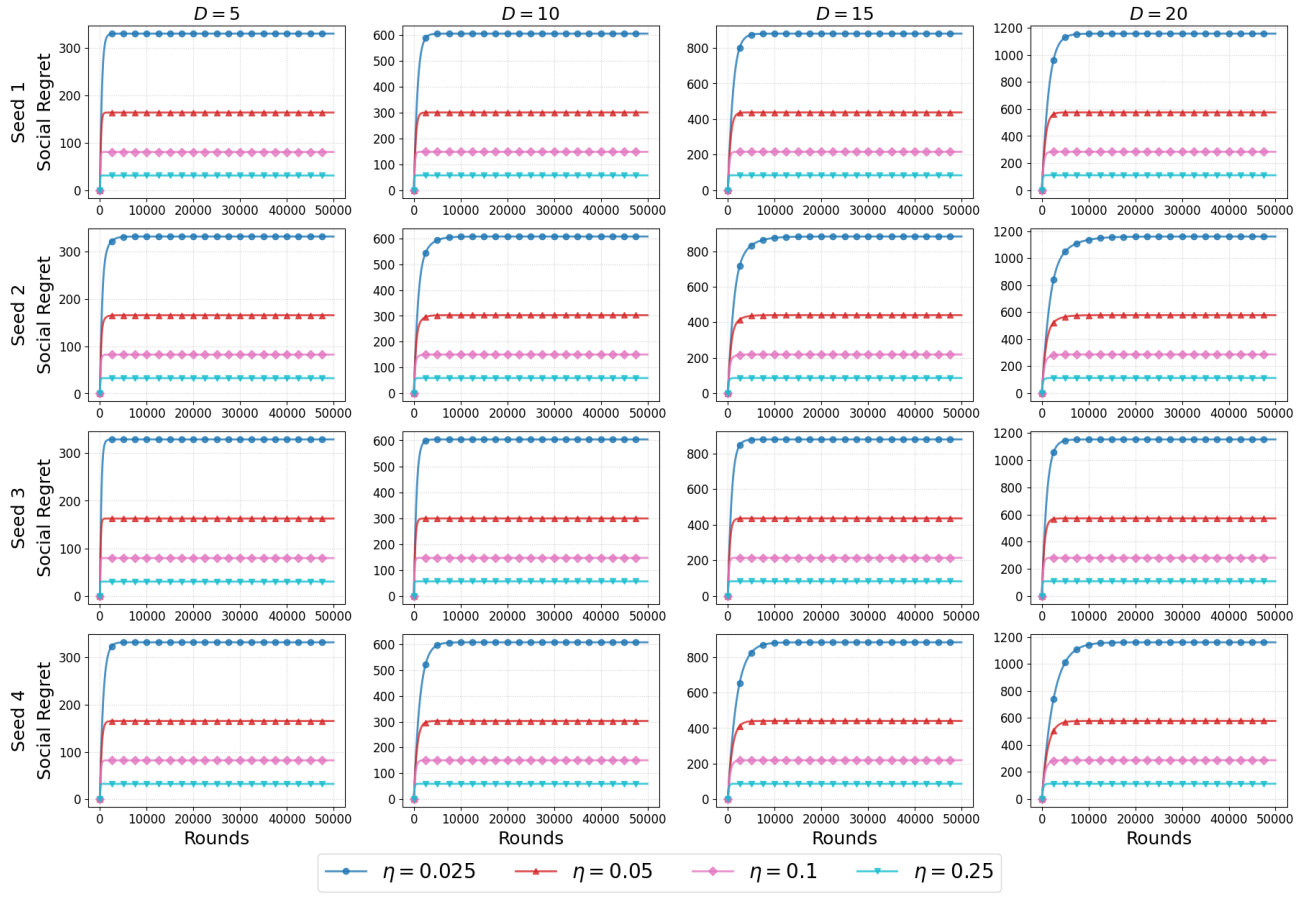

Figure 5: Social regret (vertical axis) as a function of the number of rounds T (horizontal axis) for the proposed learning dynamic with different learning rates η. Each subplot corresponds to a specific random seed (row) and delay parameter D (column).

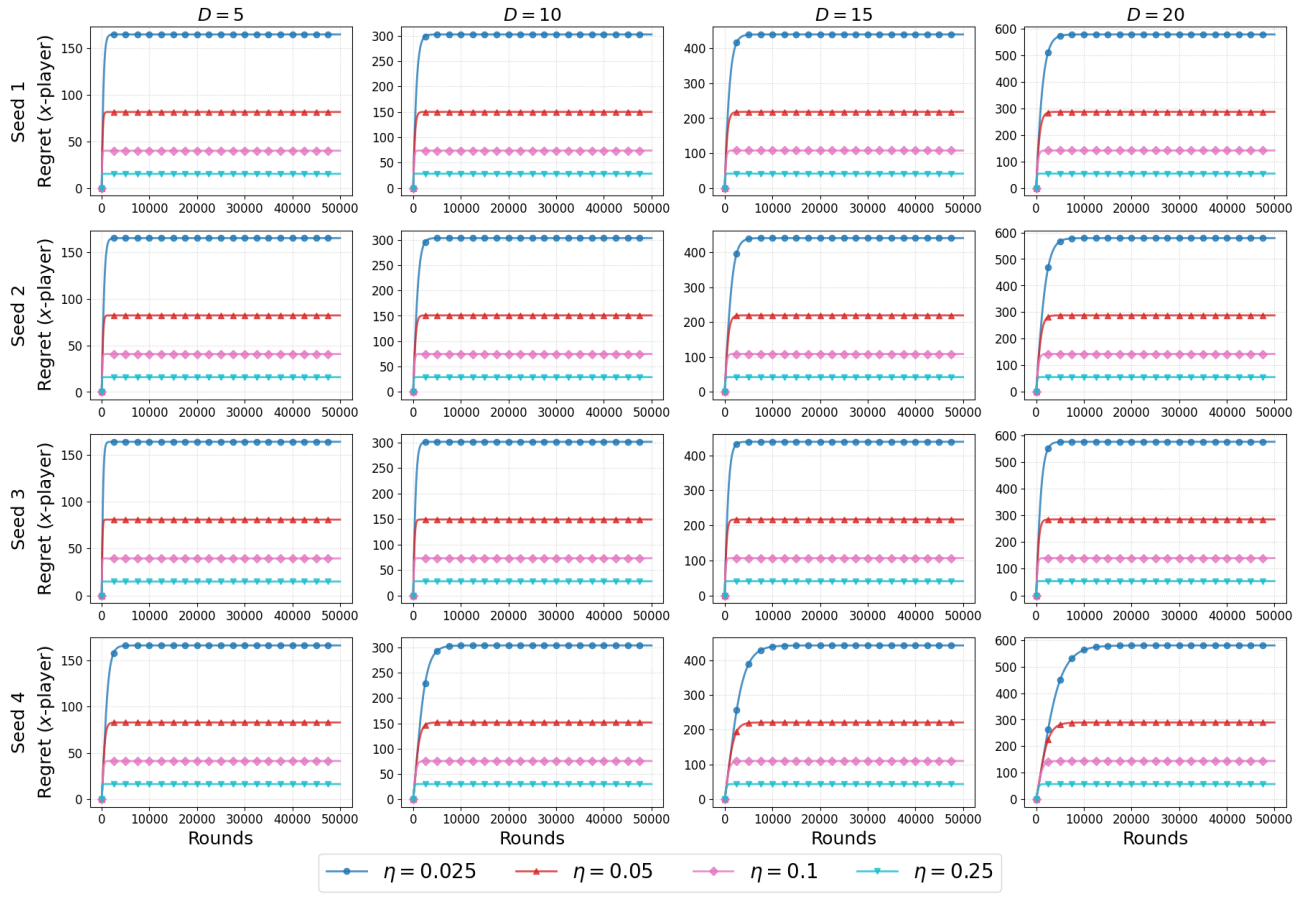

Figure 6: Individual regret of the x-player as a function of the number of rounds T.

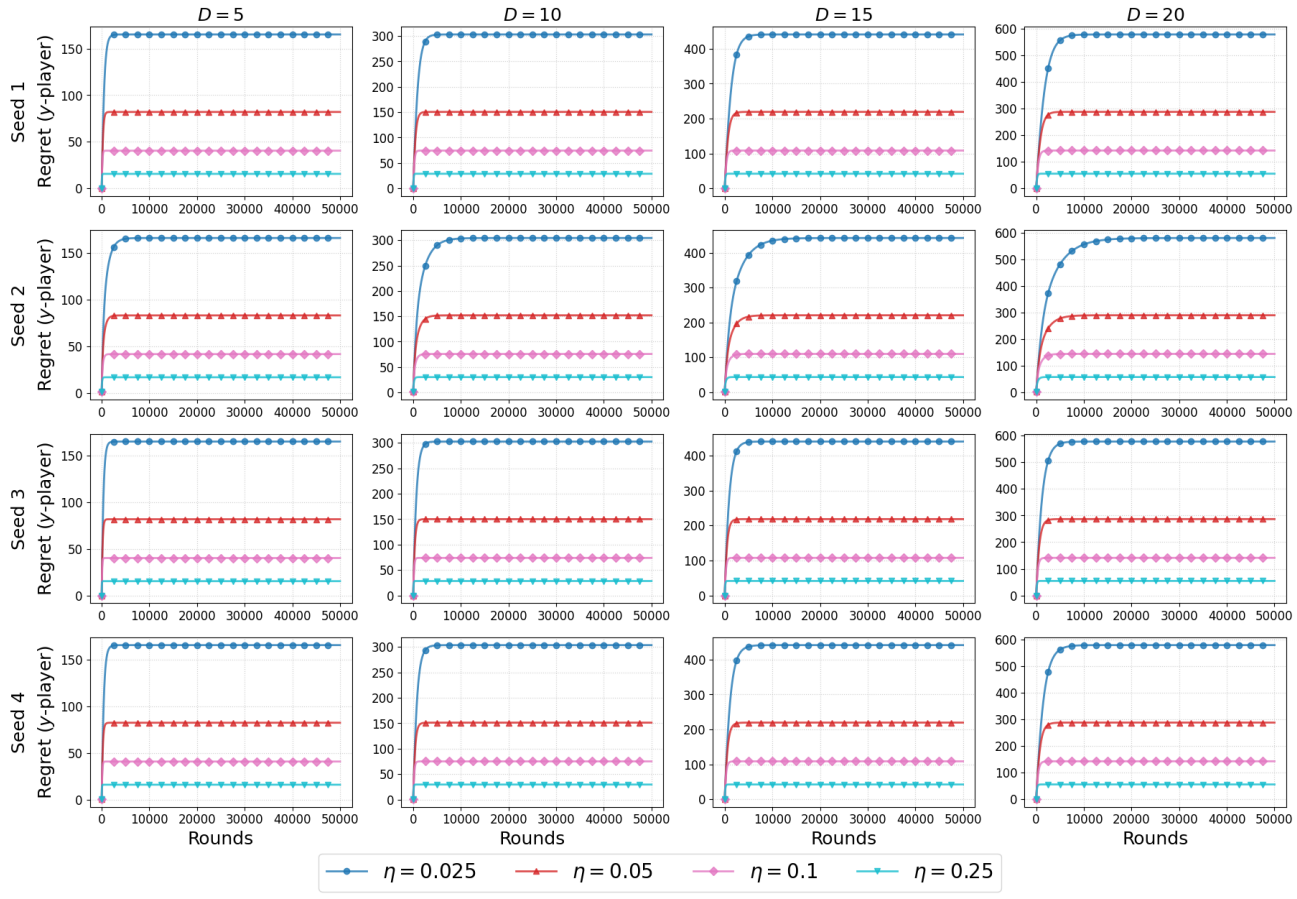

Figure 7: Individual regret of the y-player as a function of the number of rounds T.