Learning from Delayed Feedback in Games via Extra Prediction

Yuma Fujimoto
CyberAgent
fujimoto.yuma1991@gmail.com

Kenshi Abe
CyberAgent
abe_kenshi@cyberagent.co.jp

Kaito Ariu
CyberAgent
kaito_ariu@cyberagent.co.jp

Abstract

This study raises and addresses the problem of time-delayed feedback in learning in games. Because learning in games assumes that multiple agents independently learn their strategies, a discrepancy in optimization often emerges among the agents. To overcome this discrepancy, the prediction of the future reward is incorporated into algorithms, typically known as Optimistic Follow-the-Regularized-Leader (OFTRL). However, the time delay in observing the past rewards hinders the prediction. Indeed, this study firstly proves that even a single-step delay worsens the performance of OFTRL from the aspects of social regret and convergence. This study proposes the weighted OFTRL (WOFTRL), where the prediction vector of the next reward in OFTRL is weighted n times. We further capture an intuition that the optimistic weight cancels out this time delay. We prove that when the optimistic weight exceeds the time delay, our WOFTRL recovers the good performances that social regret is constant in general-sum normal-form games, and the strategies last-iterate converge to the Nash equilibrium in poly-matrix zero-sum games. The theoretical results are supported and strengthened by our experiments.

1 Introduction

Normal-form games involve multiple agents who independently choose their actions from a finite set of options, while their rewards depend on the joint actions of all agents. Thus, when they learn their strategies, it matters to take into account the temporal change in the others' strategies. One of the representative methods in such multi-agent learning is "optimistic" algorithms, where an agent not only updates its strategy naively by the current rewards, such as Follow the Regularized Leader (FTRL) [1, 2] and Mirror Descent (MD) [3, 4], but also predicts its future reward. These algorithms are called optimistic FTRL (OFTRL) [5] and MD (OMD) [6] and are known to achieve good performance in two metrics. The first metric is regret, which evaluates how close to the optimum the time series of their strategies is. When all agents adopt OFTRL or OMD, their social regret is constant regardless of the final time T [6, 5], which outperforms the regret of $O(\sqrt{T})$ by vanilla FTRL and MD [7, 8]. The second metric is convergence, which judges whether or not their strategies converge to the Nash equilibrium. OFTRL and OMD are also known to converge to the Nash equilibrium in poly-matrix zero-sum games [9, 10, 11, 12], whereas vanilla FTRL and MD frequently exhibit cycling and fail to converge [13, 14, 15]. Therefore, the prediction of future reward plays a key role in multi-agent learning.

In such optimistic algorithms, the prediction of future reward is based on observed previous rewards. In the real world, however, various factors can hinder the observation of previous rewards. One of the

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

most likely causes of this unobservability is time delay (or latency) to observe the previous rewards. A typical situation where such time delays in learning in games matter is a market (i.e., the Cournot competition [16, 17, 18]), where each firm determines the amount of its products as its strategy, but sales of the products can be observed with some time lags. In addition, multi-agent recommender systems are often used and analyzed [19, 20], where multiple recommenders equipped with different perspectives cooperatively introduce an item to buyers. Such recommender systems also face the problem of time-delayed feedback [21] because some delays occur until the buyers actually purchase the item. Such delayed feedback perhaps makes it difficult to predict future rewards and worsens the performance of multi-agent learning.

This study proposes and addresses the problem that time-delayed feedback significantly impacts both social regret and convergence in learning in games. Our contributions are summarized as follows.

- **An example is provided where any small time delay worsens both social regret and convergence.** This example is Matching Pennies under the unconstrained setting with the Euclidean regularizer. In detail, Thm. 6 shows that OFTRL suffers from $O(\sqrt{T})$-regret, corresponding to the worst-case regret obtained in the adversarial setting. In addition, Thm. 7 shows that OFTRL cannot converge to the equilibrium but rather diverges.
- **An algorithm tolerant to time delay is proposed and interpreted.** We formulate weighted OFTRL (WOFTRL), which weights the optimism in OFTRL by n times. Based on the Taylor expansion, we find that the time delay m and the optimistic weight n cancel each other out. This cancel-out is also confirmed in Thms. 6 and 7.
- **Our algorithm is proved to achieve the constant social regret and last-iterate convergence.** In Cor. 11, we prove that when the optimistic weight one-step exceeds the time delay ($n = m + 1$), the regret is $O(m^2)$. Furthermore, Cor. 13 shows that the strategies converge to the Nash equilibrium in any poly-matrix zero-sum games. Our experiments (Figs. 1-3) also support and reinforce these corollaries.

Related works: Delayed feedback has been frequently studied in the context of online learning. The delays are known to worsen regret in the full feedback [22, 23, 24, 25, 26] and bandit feedback [27, 28, 29, 30, 31, 32, 33, 34] settings. In the context of learning in games, abnormal feedback has recently attracted attention and been studied. A representative example is time-varying games, where a gap between the true reward and feedback always exists as the game changes over time. Such a gap affects the properties of regret [35, 36, 37, 38] and convergence [39, 40, 41, 42].

2 Setting

We consider N players which is labeled by $i \in \{1, \cdots, N\}$. Let $\mathcal{A}_i$ denote the set of player i's actions. Each player i's payoff depends on its own action and the actions of the other players. Each player i's strategy is in what probabilities the player chooses its actions and given by the probability distribution $\boldsymbol{x}_i \in \mathcal{X}_i := \Delta^{|\mathcal{A}_i|-1}$. The expected payoff is defined as $U_i(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$. Here, U_i is multi-linear; in other words, the following equation holds

$$U_i(\boldsymbol{x}_1, \cdots, a\boldsymbol{x}_j + a'\boldsymbol{x}'_j, \cdots, \boldsymbol{x}_N) = aU_i(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_j, \cdots, \boldsymbol{x}_N) + a'U_i(\boldsymbol{x}_1, \cdots, \boldsymbol{x}'_j, \cdots, \boldsymbol{x}_N), \tag{1}$$

for all $\boldsymbol{x}_j, \boldsymbol{x}'_j \in \mathcal{X}_i$ and all $a, a' \in \mathbb{R}$ and all $j \in \{1, \cdots, N\}$.

We define the gradient of the expected payoff as $\boldsymbol{u}_i(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) = \partial U_i(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)/\partial \boldsymbol{x}_i$. (Here, we remark that $\boldsymbol{u}_i(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$ is independent of $\boldsymbol{x}_i$ because of the multi-linearity of U_i.) We use the concatenation of the strategies and rewards as $\boldsymbol{x} := (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) \in \mathcal{X} := (\mathcal{X}_1, \cdots, \mathcal{X}_N)$ and $\boldsymbol{u}(\boldsymbol{x}) = (\boldsymbol{u}_1(\boldsymbol{x}), \cdots, \boldsymbol{u}_N(\boldsymbol{x}))$, respectively. Based on the multi-linearity of U_i, the L-Lipschitz continuity is satisfied with $L := 2\max_{(i,\boldsymbol{x})} |U_i(\boldsymbol{x})|$ as

$$\|\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{u}(\boldsymbol{x}')\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|_2. \tag{2}$$

2.1 Online Learning with Time Delay

Every round $t \in \{1, \cdots, T\}$, each player sequentially determines its strategy $\boldsymbol{x}_i^t \in \mathcal{X}_i$. As a result, the player observes the rewards for each action, i.e., $\boldsymbol{u}_i^t = \boldsymbol{u}_i(\boldsymbol{x}_1^t, \cdots, \boldsymbol{x}_N^t)$ (called full feedback

setting). Here, u_{ij}^t indicates player i's reward by choosing action a_j. Thus, in online learning, each player determines the next strategy $\boldsymbol{x}_i^{t+1}$ by using the algorithm $\boldsymbol{f}_i$ of the past rewards $\{\boldsymbol{u}_i^s\}_{1\leq s\leq t}$, which is denoted as

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{f}_i(\{\boldsymbol{u}_i^s\}_{1\leq s\leq t}). \tag{3}$$

Let us further define a time delay in the observation of past rewards. When the time delay of $m \in \mathbb{N}$ steps occurs, only the rewards of $\{\boldsymbol{u}_i^{s-m}\}_{1\leq s\leq t}$ are observable in Eq. (3). Here, we assumed that rewards before the initial time give no information, i.e., $\boldsymbol{u}_i^t = \boldsymbol{0}$ for all $t \leq 0$. Thus, the time delay modifies online learning algorithms as follows.

Definition 1 (Online learning with time delay)**.** *With the time delay of $m \in \mathbb{N}$ steps, online learning is modified as*

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{f}_i(\{\boldsymbol{u}_i^{s-m}\}_{1\leq s\leq t}), \tag{4}$$

where $\boldsymbol{u}_i^t = \boldsymbol{0}$ for all $t \leq 0$.

2.2 Follow the Regularized Leader

Under the full feedback setting, one of the most successful algorithms is Follow the Regularized Leader (FTRL). The generalization of this FTRL is formulated as follows.

Definition 2 (Generalized FTRL with time delay)**.** *With the time delay of $m \in \mathbb{N}$ steps, generalized FTRL [5] is formulated as follows*

$$\boldsymbol{x}_i^{t+1} = \arg\max_{\boldsymbol{x}_i\in\mathcal{X}} \eta \left\langle \boldsymbol{x}_i, \sum_{s=1}^{t-m} \boldsymbol{u}_i^s + \boldsymbol{m}_i^t \right\rangle - h(\boldsymbol{x}_i). \tag{5}$$

Here, $h(\boldsymbol{x}_i)$ is 1-strongly convex and is called regularizer.

Here, $\eta \in \mathbb{R}$ is learning rate, and h is a regularizer. $\boldsymbol{m}_i^t$ depends on the details of FTRL variants, typically, vanilla [1, 2] and optimistic FTRL [5].

Algorithm 3 (Vanilla FTRL with time delay)**.** *When $\boldsymbol{m}_i^t = \boldsymbol{0}$, generalized FTRL corresponds to vanilla FTRL.*

Algorithm 4 (Optimistic FTRL with time delay)**.** *When $\boldsymbol{m}_i^t = \boldsymbol{u}_i^{t-m}$, generalized FTRL corresponds to optimistic FTRL (OFTRL).*

3 Issue by Time Delay

In this section, we introduce an example in which any small time delay worsens the performance of OFTRL. As an example, we define Matching Pennies as follows. This is the simplest game that provides cycling behavior by vanilla FTRL.

Example 5 (Matching Pennies)**.** *Matching Pennies considers two players ($N = 2$) with their rewards;*

$$\boldsymbol{u}_1 = +\boldsymbol{A}\boldsymbol{x}_2, \quad \boldsymbol{u}_2 = -\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_1, \quad \boldsymbol{A} = \begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}. \tag{6}$$

Let us focus on two aspects: regret and convergence. First, individual ($\textsc{Reg}_i$) and social ($\textsc{RegTot}$) regret are defined as

$$\textsc{Reg}_i(T) := \max_{\boldsymbol{x}_i\in\mathcal{X}_i} \sum_{t=1}^{T} \left\langle \boldsymbol{x}_i - \boldsymbol{x}_i^t, \boldsymbol{u}_i^t \right\rangle, \quad \textsc{RegTot}(T) := \sum_{i=1}^{N} \textsc{Reg}_i(T). \tag{7}$$

It has already been known that without any time delay, OFTRL achieves constant social regret (called "fast convergence" [5]). However, if a slight time delay exists, OFTRL suffers from the social regret of $\Omega(\sqrt{T})$ at least even in such a simple game (see Appendix B for the full proof). This regret is also obtained under an adversarial environment [5] and thus is worst-case for OFTRL.

Theorem 6 (Social regret of $\Omega(\sqrt{T})$ by time delay)**.** *Suppose that the strategy space is unconstrained ($\boldsymbol{x}_i \in \mathbb{R}^{|\mathcal{A}_i|}$) in Exm. 5 with the Euclidean regularizer ($h(\boldsymbol{x}_i) = \|\boldsymbol{x}_i\|_2^2/2$). Then, when both players use OFTRL ($n = 1$) with any time delay ($m \geq 1$), their social regret is no less than $\Omega(\sqrt{T})$, which is achieved with $\eta = 1/\sqrt{T}$.*

Proof Sketch. In the proof, we consider an extended class of OFTRL, where the weight of the optimistic prediction of the future is generalized. By direct calculations, we obtain the recurrence formula of the dynamics of both players' strategies. We prove a lemma that this recurrence formula is approximately solved by circular functions, whose radius varies with time depending on the time delay m and optimistic weight n. Here, we emphasize that the rate of the radius change is $e^{\alpha t}$ with $\alpha = m - n + 1/2$. Thus, if the players use OFTRL ($n = 1$), the radius always grows ($\alpha > 0$) for any time delay $m \geq 1$. In conclusion, OFTRL performs the same as vanilla FTRL and suffers from the social regret of $\Omega(\sqrt{T})$ at least. □

Second, the Nash equilibrium is defined as $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_N^*)$ which satisfies

$$\boldsymbol{x}_i^* \in \arg\max_{\boldsymbol{x}_i \in \mathcal{X}_i} U_i(\boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N^*), \tag{8}$$

for all $i = 1, \cdots, N$. We also denote the set of Nash equilibria as $\mathcal{X}^*$. In poly-matrix zero-sum games, where the payoff U_i is divided into the zero-sum matrix games, it has been known that OFTRL achieves convergence to the Nash equilibrium. This convergence is measured by the distance from the Nash equilibria, which is formulated as

$$\textsc{Dis}(T) := \sqrt{\min_{\boldsymbol{x}^* \in \mathcal{X}^*} \sum_{i=1}^{N} \|\boldsymbol{x}_i^T - \boldsymbol{x}_i^*\|_2^2}. \tag{9}$$

Under the existence of any small time delay ($m \geq 1$), the distance ($\textsc{Dis}(T)$) does not converge but rather diverges as follows (see Appendix C for the full proof).

Theorem 7 (Divergence by time delay)**.** *Suppose that the strategy space is unconstrained ($\boldsymbol{x}_i \in \mathbb{R}^{|\mathcal{A}_i|}$) in Exm. 5 with the Euclidean regularizer ($h(\boldsymbol{x}_i) = \|\boldsymbol{x}_i\|_2^2/2$). Then, when both the players use OFTRL ($n = 1$) with any time delay ($m \geq 1$), their strategies diverge from the equilibrium ($\lim_{T\to\infty} \textsc{Dis}(T) \to \infty$).*

Proof Sketch. We reuse the lemma in the proof of Thm. 6, showing that the dynamics of the strategies of both players are approximately solved by circular functions, whose radius varies with the exponential rate of $\alpha = m - n + 1/2$. In OFTRL ($n = 1$) with some delay ($m \geq 1$), the rate is positive ($\alpha > 0$), meaning that the radius is amplified with time. Because the equilibrium is at the center of the circular function, the divergence from the equilibrium is proven. □

Interpretation of Thms. 6 and 7: The crux of their proofs is that the exponential rate of the radius change (α) is divided into three terms: 1) expansion by the time delay ($+m$), 2) contraction by optimism ($-n$), and 3) expansion by the accumulation of the discretization errors. These terms intuitively explain the previous and present results. First, in the case of vanilla FTRL ($m = n = 0$), the radius grows with the exponential rate of $\alpha = 1/2$. This means that the strategies diverge from the Nash equilibrium over time, and social regret converges only slowly. Second, in the case of OFTRL ($m = 0$, $n = 1$), the radius shrinks with the exponential rate of $\alpha = -1/2$. This means that the strategies converge to the equilibrium over time, and social regret becomes constant. Finally, if OFTRL is accompanied by a time delay ($m \geq 1$, $n = 1$), the radius grows with the exponential rate of $\alpha \geq 1/2$, resulting in the social regret of $\Omega(\sqrt{T})$ again.

Remark on unconstrained setting: Note that Thm. 6 considers the unconstrained setting. In other words, the dynamics assume no boundary condition in the strategy space. This unconstrained setting is necessary to analyze the global behavior of the dynamics, which differs on the boundary of the strategy space in the constrained setting. However, this unconstrained setting is sufficient to capture the worsening of the performance of OFTRL. At least, our experiments observe the same results as Thms. 6 and 7 even in the constrained setting.

4 Algorithm

So far, we understand an issue that OFTRL becomes useless due to the effect of time delay. The next question is how the property of constant social regret is recovered, even under time delay. In the following, we propose an extension of OFTRL, where the optimistic prediction of the future reward is added $n \in \mathbb{N}$ times.

Algorithm 8 (Weighted Optimistic Follow the Regularized Leader)**.** *Weighted Optimistic Follow The Regularized Leader (WOFTRL) is given by* $\boldsymbol{m}_i^t = n\boldsymbol{u}_i^{t-m}$ *for* $n \in \mathbb{N}$ *in generalized FTRL.*

Interpretation of WOFTRL: WOFTRL can be rewritten as

$$\boldsymbol{x}_i^t = \arg\max_{\boldsymbol{x}_i \in \mathcal{X}_i} \eta \left\langle \boldsymbol{x}_i, \tilde{\boldsymbol{x}}_i^t \right\rangle - h(\boldsymbol{x}_i), \quad \tilde{\boldsymbol{x}}_i^t = \sum_{s=1}^{t-m-1} \boldsymbol{u}_i^s + n\boldsymbol{u}_i^{t-m-1}. \tag{10}$$

For arbitrary time series $\{v^t\}_t$, we define the finite difference of the time series as

$$\delta v^t := v^{t+1} - v^t, \quad \delta\delta v^t = \delta v^{t+1} - \delta v^t. \tag{11}$$

In Eq. (10), the following relation is satisfied;

$$\delta\tilde{\boldsymbol{x}}_i^t = \boldsymbol{u}_i^{t-m} + n\delta\boldsymbol{u}_i^{t-m-1}. \tag{12}$$

By definition, we evaluate the time-delayed terms as

$$\boldsymbol{u}_i^{t-m} = \boldsymbol{u}_i^t - m\delta\boldsymbol{u}_i^t - \sum_{s=t-m}^{t-1}\sum_{r=s}^{t-1} \delta\delta\boldsymbol{u}_i^r, \quad \delta\boldsymbol{u}_i^{t-m-1} = \delta\boldsymbol{u}_i^t - \sum_{s=t-m-1}^{t-1} \delta\delta\boldsymbol{u}_i^s. \tag{13}$$

By substituting these into Eq. (12), we obtain

$$\delta\tilde{\boldsymbol{x}}_i^t = \boldsymbol{u}_i^t + (n-m)\delta\boldsymbol{u}_i^t - \left(\sum_{s=t-m}^{t-1}\sum_{r=s}^{t-1} \delta\delta\boldsymbol{u}_i^r + n \sum_{s=t-m-1}^{t-1} \delta\delta\boldsymbol{u}_i^s \right). \tag{14}$$

In the RHS, the first term corresponds to the vanilla FTRL without time delay and is $O(1)$. Here, it is known that predicting future rewards enhances performance in terms of social regret and convergence. Since $\delta\boldsymbol{u}_i^t$ in the second term shows the time derivative of rewards and is $O(\eta)$, it pushes ahead the time t in $\boldsymbol{u}_i^t$. Thus, if its coefficient is positive ($n - m > 0$), the second term contributes to predicting a future reward and is expected to improve performance. In addition, we remark that an optimistic weight n and time delay m conflict with each other. Finally, the third term is a prediction error which consists of the accumulation of only $\delta\delta\boldsymbol{u}_i^t$ and is $O(\eta^2)$. Thus, this term is negligible if η is sufficiently small.

5 Constant Social Regret

This section shows that our WOFTRL can achieve constant social regret. For the regret analysis, let us introduce the following Regret bounded by Variation in Utilities (RVU) property [5].

Definition 9 (RVU property)**.** *The time series of* $\{\boldsymbol{x}_i^t\}_{1\le t\le T}$ *is defined to satisfy the RVU property if there exists* $0 < \alpha$ *and* $0 < \beta \le \gamma$ *such that*

$$\mathrm{REG}_i(T) \le \alpha + \beta \sum_{t=1}^{T} \|\boldsymbol{u}_i^t - \boldsymbol{u}_i^{t-1}\|_2^2 - \gamma \sum_{t=1}^{T} \|\boldsymbol{x}_i^t - \boldsymbol{x}_i^{t-1}\|_2^2. \tag{15}$$

We can prove that WOFTRL satisfies this property as follows (see Appendix D for the full proof).

Theorem 10 (RVU property when $n = m + 1$)**.** *When* $n = m + 1$*, WOFTRL using learning rate* η *satisfies the RVU property with constraints* $\alpha = h_{\max}/\eta$*,* $\beta = \lambda^2\eta$*, and* $\gamma = 1/(4\eta)$*. Here, we defined* $\lambda := (m+1)(m+2)/2$*.*

Proof sketch. The basic procedure of the proof is the same as the previous study [5]. We show that when $n = m + 1$, the cumulative payoff ($\tilde{\boldsymbol{x}}_i^t$ in Eq. (10)) well approximates the gold reward until time t, defined as

$$\tilde{\boldsymbol{g}}_i^t := \sum_{s=1}^{t} \boldsymbol{u}_i^s. \tag{16}$$

Here, note that β is λ^2 times larger (worse) than without the time delay. λ is interpreted as accumulated errors in predicting the gold reward. Indeed, the difference between the gold and predicted rewards is described as

$$\tilde{\boldsymbol{g}}_i^t - \tilde{\boldsymbol{x}}_i^t = \sum_{s=t-m}^{t} \boldsymbol{u}_i^s - (m+1)\boldsymbol{u}_i^{t-m-1}, \tag{17}$$

where the total time gap in its RHS is evaluated as $\sum_{s=t-m}^{t}\{s-(t-m-1)\} = (m+1)(m+2)/2 = \lambda$. □

Corollary 11 (Constant social regret with time delay)**.** *When $n = m+1$, WOFTRL using the learning rate of $\eta = 1/(2\lambda L)$ achieves the social regret of $O(m^2)$, i.e.,* $\mathrm{RegTot}(T) = O(m^2)$.

Proof. Sum up for $i \in \{1, \cdots, N\}$ the RVU property in Thm. 10, then the second term of the payoff is lower-bounded by the L-Lipschitz continuity as

$$\|\boldsymbol{u}^t - \boldsymbol{u}^{t-1}\|_2^2 \le L^2 \|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|_2^2. \tag{18}$$

By directly applying $\eta = 1/(2\lambda L)$, the second and third terms are canceled out, and we obtain

$$\mathrm{RegTot}(T) \le 2\lambda N L h_{\max} = O(m^2). \tag{19}$$

□

Remark on Cor. 11: An important remark is that the time delay m worsens social regret. One demerit is that we must use a $\lambda = (m+1)(m+2)/2$ times smaller learning rate. This leads to another demerit that the social regret becomes $\lambda = (m+1)(m+2)/2$ times larger, too. In addition, we remark that if no time delay exists ($m = 0$), the previous result (Cor. 6 in [5]) is restored as $\lambda = 1$. We also obtain the individual regret of $O(m^{3/2}T^{1/4})$ by utilizing Thm. 10 (see Appendix F for details).

5.1 Experiment for Matching Pennies

We consider Matching Pennies defined in Exm. 5. Now Fig. 1 provides the experiments for WOFTRL with the Euclidean regularizer, i.e., $h(\boldsymbol{x}_i) = \|\boldsymbol{x}_i\|_2^2/2$.

Constant social regret when $n > m$: First, see the upper left triangle region of Panel A. This region corresponds to $n > m$, where the optimistic weight is greater than the time delay. It shows that the regret is sufficiently small. In detail, Panel B shows the time series of the regrets for various optimistic weights n with a fixed time delay ($m = 10$). In the cases of $n > m$ (i.e., $n = 11, 13, 15$), we see that the regrets converge to sufficiently small values. Finally, Panel C shows the two ways to take the learning rate η. When we keep η constant, the regret is also constant ($O(1)$), independent of the final time T. This is completely different from the case of $\eta = 1/\sqrt{T}$, where the regret grows with $O(\sqrt{T})$.

$O(\sqrt{T})$-social regret when $n \le m$: Next, see the lower right triangle region of Panel A. This region corresponds to $n \le m$, where the optimistic weight is smaller than the time delay. It shows that the regret takes various values and is relatively large. In detail, Panel B shows that the regret oscillates and is large in the case of $n \le m$ (i.e., $n = 1, 3, 5, 7, 9$ against $m = 10$). It also shows that a transition in the behavior of the regret occurs at $n = m$.

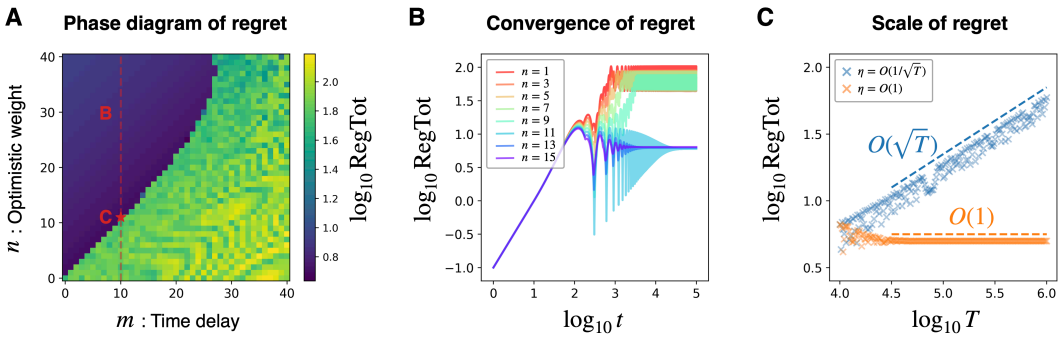

Figure 1: Regret analysis for Matching Pennies (Exm. 5). **A**. The phase diagram of social regrets for various time delays m (horizontal) and optimistic weights n (vertical). The deep blue color indicates that the regret is small ($O(1)$-regret), while the green and yellow ones indicate that the regret is large ($O(\sqrt{T})$-regret). A transition is clearly shown between $O(1)$- and $O(\sqrt{T})$-regret. We set the parameters as $T = 10^5$ and $\eta = 10^{-2}$. **B**. The convergence of social regrets for various optimistic weights n and a fixed time-delay $m = 10$ (corresponding to the red broken line in Panel A). We see the regret oscillates and is relatively large for $n = 1, 3, 5, 7, 9$ ($O(\sqrt{T})$-regret) but converges to a small value for $n = 11, 13, 15$ ($O(1)$-regret). A transition is clearly observed again in $m = n$. We set the parameters as $\eta = 10^{-2}$. **C**. The scale of social regrets in the case of $m = 10$ and $n = 11$ (corresponding to the red star in Panel A). We plot the two ways to take learning rate: $\eta = 1/\sqrt{T}$ (blue dots) and $\eta = O(1)$ (orange ones). The regrets for $\eta = O(1/\sqrt{T})$ follow the broken blue line, which has a slope of $1/2$ (meaning that the regrets are $O(\sqrt{T})$). On the other hand, the regrets for $\eta = O(1)$ follow the broken orange line, which has a slope of 0 (meaning that the regrets are $O(1)$). We set $\eta = 1/\sqrt{T}$ for the blue dots and $\eta = 10^{-2}$ for the orange ones.

Exceptional $O(\sqrt{T})$-social regret when $n > m$: Finally, see the particular region where m is large but $n - m$ is sufficiently small (e.g., $m = 30$ and $n = 35$) in Panel A. Note that the regret is $O(\sqrt{T})$ even though this region satisfies $n > m$. This $O(\sqrt{T})$-regret is due to the finiteness of the learning rate η. When there is a sufficiently large time delay (e.g., $m = 30$) against a finite learning rate (e.g., $\eta = 10^{-2}$ in Panel A), it becomes difficult to estimate the current status from the time-delayed rewards. This difficulty is related to the accumulation of estimation errors following $\lambda = (m+1)(m+2)/2$ in Cor. 11. If we take sufficiently small η, the fast convergence occurs in the wider region of $n > m$.

5.2 Experiment for Sato's Game

We further discuss an example of a nonzero-sum game. We pick up a Sato's game [43], which is Rock-Paper-Scissors, but some scores are also generated in draw cases, as follows.

$$\boldsymbol{u}_1 = +(\boldsymbol{A}+\boldsymbol{D})\boldsymbol{x}_2, \quad \boldsymbol{u}_2 = -\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_1, \quad \boldsymbol{A} = \begin{pmatrix} 0 & -1 & +1 \\ +1 & 0 & -2 \\ -1 & +2 & 0 \end{pmatrix}, \quad \boldsymbol{D} = \frac{1}{10}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}. \tag{20}$$

In such nonzero-sum Sato's games, heteroclinic cycles with complex oscillations that diverge from the Nash equilibrium are observed [43].

From our experiments (Fig. 2), we obtain similar results to those in Fig. 1: (i) Constant social regret when $n > m$, (ii) $O(\sqrt{T})$-social regret when $n \leq m$, and (iii) Exceptional $O(\sqrt{T})$-social regret when $n > m$.

6 Convergence in Poly-Matrix Zero-Sum Games

We also show that WOFTRL convergences to the Nash equilibrium in poly-matrix zero-sum games, which are separable into zero-sum games between a couple of players, as defined below.

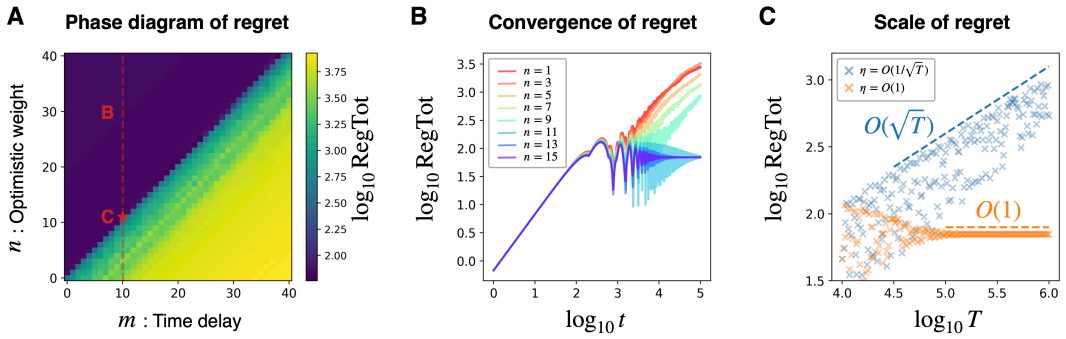

Figure 2: Regret analysis for a Sato's game (Eqs. 20) with the entropic regularizer, i.e., $h(\boldsymbol{x}) = \langle \boldsymbol{x}, \log \boldsymbol{x} \rangle$. The results and parameter settings for all the panels are the same as those in Fig. 1. **A**. The phase diagram of social regrets for various time delays m (horizontal) and optimistic weights n (vertical). **B**. The convergence of social regrets for various optimistic weights n and a fixed time-delay $m = 10$ (corresponding to the red broken line in Panel A). **C**. The scale of social regrets in the case of $m = 10$ and $n = 11$ (corresponding to the red star in Panel A).

Definition 12 (Poly-matrix zero-sum games)**.** *Poly-matrix zero-sum games are*

$$\boldsymbol{u}_i = \sum_{i' \neq i} \boldsymbol{A}_{(ii')} \boldsymbol{x}_{i'}, \quad \boldsymbol{A}_{(i'i)} = -\boldsymbol{A}_{(ii')}^{\mathrm{T}}. \tag{21}$$

Matching Pennies is a special case of poly-matrix games. In poly-matrix zero-sum games, the strategy converges to the Nash equilibrium for a sufficiently small learning rate (see Appendix E for the full proof).

Corollary 13 (Last-Iterate Convergence)**.** *Suppose h is a convex function of Legendre type [44, 45]. When $n = m + 1$, WOFTRL using the learning rate of $\eta \leq 1/(\sqrt{8}n^2 L)$ converges to the Nash equilibrium, i.e., there exists $\boldsymbol{x}^* \in \mathcal{X}^*$ such that $\lim_{T\to\infty} \|\overline{\boldsymbol{x}^T} - \boldsymbol{x}^*\|_2 = 0$.*

Proof Sketch. We follow but extend the method by the prior study [10]. First, we prove that when the regularizer is Legendre type, generalized FTRL is equivalent to its MD (mirror descent) variant [6]. (We remark that this is an extension of the already-known equivalence between vanilla FTRL and MD [46, 45].) We thereafter show the last-iterate convergence via this MD variant. Furthermore, for each Nash equilibrium $\boldsymbol{x}^* \in \mathcal{X}^*$, we find a divergence $V^t(\boldsymbol{x}^*)$ defined as

$$V^t(\boldsymbol{x}^*) := D(\boldsymbol{x}^*, \hat{\boldsymbol{x}}^t) + \frac{1}{2}\Big\{ D(\hat{\boldsymbol{x}}^t, \boldsymbol{x}^{t-1}) + \sum_{r=1}^{n-1} \frac{n-r}{n} (D(\boldsymbol{x}^{t-r}, \hat{\boldsymbol{x}}^{t-r}) + D(\hat{\boldsymbol{x}}^{t-r}, \boldsymbol{x}^{t-r-1})) \Big\}, \tag{22}$$

which monotonically decreases with time t. We prove that this divergence decreases to 0 for one equilibrium $\boldsymbol{x}^*$, meaning that WOFTRL converges to the equilibrium. □

6.1 Experiment for Rock-Paper-Scissors

Now, Fig. 3 shows the simulations for weighted Rock-Paper-Scissors, formulated as Eqs. (20) with the nonzero-sum term $\boldsymbol{D}$ eliminated.

Last-iterate convergence when $n > m$: See the two right panels, where the optimistic weight ($n = 5, 6$) dominates the time delay ($m = 4$). Thus, the trajectories converge to the Nash equilibrium, meaning last-iterate convergence. Here, we also observe that when the optimistic weight is larger ($n = 6$ than $n = 5$), convergence is faster.

Non-convergence when $n \leq m$: See the two left panels, where the optimistic weight ($n = 3, 4$) does not dominate the time delay $m = 4$. Thus, the trajectories fail to converge but rather diverge from the equilibrium. Here, we also see that when the optimistic weight is smaller ($n = 3$ than $n = 4$), the divergence is faster.

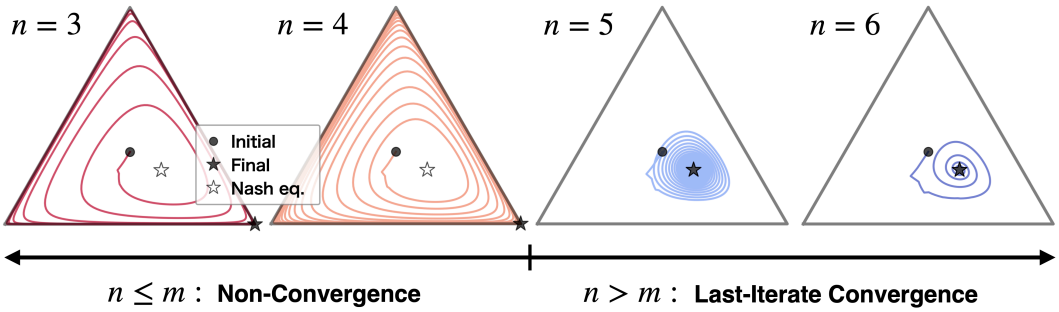

Figure 3: Convergence analysis for a weighted Rock-Paper-Scissors with the entropic regularizer, i.e., $h(\boldsymbol{x}_i) = \langle \boldsymbol{x}_i, \log \boldsymbol{x}_i \rangle$. We set the parameters as $\eta = 10^{-1}$ and $m = 4$. The colored lines are the trajectories of learning. The black dot, black star, and white star indicate the initial state, final state, and Nash equilibrium, respectively. From left to right, optimistic weights are $n = 3, 4, 5, 6$. In the left two panels ($n \leq m$), the black star does not overlap the white one, meaning non-convergence. On the other hand, in the right two panels ($n > m$), the black star overlaps the white one, indicating last-iterate convergence.

7 Conclusion

This study tackled the problem of time-delayed full feedback in learning in games. First, we demonstrated, in a simple example of Matching Pennies with the unconstrained setting, that any slight time delay makes the performance of OFTRL worse from the perspectives of social regret (Thm. 6) and convergence (Thm. 7). To overcome these impossibility theorems, we introduced WOFTRL, which weights the optimism of OFTRL n times. The Taylor expansion for the learning rate revealed that the optimistic weight n cancels the time delay m out (Eq. (14)). We proved that when the optimistic weight even slightly dominates the time delay ($n = m + 1$), both constant social regret (Cor. 11) and last-iterate convergence (Cor. 13) hold. Our experiments (Figs. 1-3) strengthened the results of these corollaries.

One future direction is to consider other types of time delay. Although this study only considered that the delay m is a constant value, real-world delays are often more complicated. For example, previous studies on online learning have discussed situations where m is given stochastically [31, 32] or adversarially [24, 25, 26]. We remark that the theory of this study is to some extent applicable to such stochastic delays; It is sufficient to take a larger n than the possible delays and use the reward before $n - 1$ steps every time, i.e., use GFTRL with $\boldsymbol{m}_i^t = n\boldsymbol{u}_i^{t-n+1}$. Another direction is to obtain the convergence rate for WOFTRL. This study opens the door to time-delay problems and offers various straightforward open questions in the field of algorithmic game theory.

Acknowledgments and Disclosure of Funding

Kaito Ariu was supported by JSPS KAKENHI Grant Numbers 23K19986 and 25K21291.

References

[1] Shai Shalev-Shwartz and Yoram Singer. Convex repeated games and fenchel duality. In *NeurIPS*, 2006.

[2] Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *COLT*, 2008.

[3] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

[4] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[5] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *NeurIPS*, 2015.

[6] Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *NeurIPS*, 2013.

[7] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, 2003.

[8] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[9] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *ICLR*, 2018.

[10] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR*, 2019.

[11] Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *NeurIPS*, 2020.

[12] Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *NeurIPS*, 2022.

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

[14] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *SODA*, 2018.

[15] James P Bailey and Georgios Piliouras. Multi-agent learning in network zero-sum games is a hamiltonian system. In *AAMAS*, 2019.

[16] Ludo Waltman and Uzay Kaymak. Q-learning agents in a cournot oligopoly model. *Journal of Economic Dynamics and Control*, 32(10):3275–3293, 2008.

[17] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolo, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.

[18] Jason D Hartline, Sheng Long, and Chenhao Zhang. Regulation of algorithmic collusion. In *CSLAW*, 2024.

[19] Joaquim Neto, A Jorge Morais, Ramiro Gonçalves, and António Leça Coelho. Multi-agent-based recommender systems: a literature review. In *ICICT*, 2022.

[20] Afef Selmi, Zaki Brahmi, and M Gammoudi. Multi-agent recommender system: State of the art. In *ICICS*, 2014.

[21] Jiaqi Yang and De-Chuan Zhan. Generalized delayed feedback model with post-click information in recommender systems. In *NeurIPS*, 2022.

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

[23] Martin Zinkevich, John Langford, and Alex Smola. Slow learners are fast. In *NeurIPS*, 2009.

[24] Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. In *NeurIPS*, 2015.

[25] Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *AAAI*, 2016.

[26] Ohad Shamir and Liran Szlak. Online learning with local permutations and delayed feedback. In *ICML*, 2017.

[27] Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. In *NeurIPS*, 2010.

[28] Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *ICML*, 2013.

[29] Thomas Desautels, Andreas Krause, and Joel W Burdick. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *J. Mach. Learn. Res.*, 15(1):3873–3923, 2014.

[30] Nicol'o Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *COLT*, 2016.

[31] Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. In *UAI*, 2017.

[32] Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *ICML*, 2018.

[33] Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Nonstochastic bandits with composite anonymous feedback. In *COLT*, 2018.

[34] Bingcong Li, Tianyi Chen, and Georgios B Giannakis. Bandit online learning with unknown delays. In *AISTATS*, 2019.

[35] Mengxiao Zhang, Peng Zhao, Haipeng Luo, and Zhi-Hua Zhou. No-regret learning in time-varying zero-sum games. In *ICML*, 2022.

[36] Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On the convergence of no-regret learning dynamics in time-varying games. In *NeurIPS*, 2023.

[37] Benoit Duvocelle, Panayotis Mertikopoulos, Mathias Staudigl, and Dries Vermeulen. Multiagent online learning in time-varying games. *Mathematics of Operations Research*, 48(2):914–941, 2023.

[38] Yu-Hu Yan, Peng Zhao, and Zhi-Hua Zhou. Fast rates in time-varying strongly monotone games. In *ICML*, 2023.

[39] Tanner Fiez, Ryann Sim, Stratis Skoulakis, Georgios Piliouras, and Lillian Ratliff. Online learning in periodic zero-sum games. In *NeurIPS*, 2021.

[40] Yi Feng, Hu Fu, Qun Hu, Ping Li, Ioannis Panageas, Xiao Wang, et al. On the last-iterate convergence in time-varying zero-sum games: Extra gradient succeeds where optimism fails. In *NeurIPS*, 2023.

[41] Yi Feng, Ping Li, Ioannis Panageas, and Xiao Wang. Last-iterate convergence separation between extra-gradient and optimism in constrained periodic games. In *UAI*, 2024.

[42] Yuma Fujimoto, Kaito Ariu, and Kenshi Abe. Synchronization in learning in periodic zero-sum games triggers divergence from nash equilibrium. In *AAAI*, 2025.

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

[44] Tyrrell R Rockafellar. *Convex analysis*. Princeton University Press, 1970.

[45] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[46] Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In *AISTATS*, 2011.

Appendix

A Dynamics in Unconstrained, Euclidean, Matching Pennies

Suppose the unconstrained setting for the Euclidean regularizer $h(\boldsymbol{x}_i) = \|\boldsymbol{x}_i\|_2^2/2$, then the convex-conjugate

$$\boldsymbol{x}_i^t = \arg\max_{\boldsymbol{x}_i \in \mathbb{R}^d} \eta \left\langle \boldsymbol{x}_i, \tilde{\boldsymbol{x}}_i^t \right\rangle - h(\boldsymbol{x}_i), \tag{A1}$$

is trivially solved as

$$\boldsymbol{x}_i^t = \tilde{\boldsymbol{x}}_i^t. \tag{A2}$$

We further consider Matching Pennies, where the payoff matrix is

$$\boldsymbol{A} = \begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix} = \mathbf{c} \otimes \mathbf{c}, \quad \mathbf{c} := \begin{pmatrix} +1 \\ -1 \end{pmatrix}. \tag{A3}$$

By substituting Eq. (A2) into Eq. (12), the dynamics of the strategies are described as

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t + \eta \boldsymbol{u}_i^{t-m} + n\eta(\boldsymbol{u}_i^{t-m} - \boldsymbol{u}_i^{t-m-1}), \tag{A4}$$

To simplify Eq. (A4), let us define

$$\Delta x_i^t := \left\langle \boldsymbol{x}_i^t, \mathbf{c} \right\rangle. \tag{A5}$$

Then, $\boldsymbol{u}_i^t$ is calculated as

$$\boldsymbol{u}_1^t = +\boldsymbol{A}\boldsymbol{x}_2^t = +\Delta x_2^t \mathbf{c}, \quad \boldsymbol{u}_2^t = -\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_1^t = -\Delta x_1^t \mathbf{c}. \tag{A6}$$

Finally, the following recurrence formulas are obtained;

$$\Delta x_1^{t+1} = \Delta x_1^t + 2(n+1)\eta \Delta x_2^{t-m} - 2n\eta \Delta x_2^{t-m-1}, \tag{A7}$$

$$\Delta x_2^{t+1} = \Delta x_2^t - 2(n+1)\eta \Delta x_1^{t-m} + 2n\eta \Delta x_1^{t-m-1}. \tag{A8}$$

These recurrence formulas are approximately solved as follows.

Lemma A1 (Solution of recurrence formulas)**.** *Eqs.* (A7) *and* (A8) *are solved as*

$$\Delta x_1^t = e^{\lambda_t} \cos\theta_t, \quad \Delta x_2^t = -e^{\lambda_t} \sin\theta_t, \tag{A9}$$

$$\lambda_{t+1} = \lambda_t + \left(m - n + \frac{1}{2}\right)(2\eta)^2 + \Theta(\eta^4), \quad \theta_{t+1} = \theta_t + 2\eta + \Theta(\eta^3), \tag{A10}$$

Proof. For readability, we scale the learning rate in Eqs. (A7) and (A8) as $2\eta \leftarrow \eta$, we obtain

$$\Delta x_1^{t+1} = \Delta x_1^t + (n+1)\eta \Delta x_2^{t-m} - n\eta \Delta x_2^{t-m-1}, \tag{A11}$$

$$\Delta x_2^{t+1} = \Delta x_2^t - (n+1)\eta \Delta x_1^{t-m} + n\eta \Delta x_1^{t-m-1}. \tag{A12}$$

By using the relation of Eqs. (A9), we obtain

$$e^{\lambda_{t+1}+i\theta_{t+1}} = \Delta x_1^{t+1} - i\Delta x_2^{t+1} \tag{A13}$$

$$= \left\{\Delta x_1^t + (n+1)\eta \Delta x_2^{t-m} - n\eta \Delta x_2^{t-m-1}\right\} - i\left\{\Delta x_2^t - (n+1)\eta \Delta x_1^{t-m} + n\eta \Delta x_1^{t-m-1}\right\} \tag{A14}$$

$$= \underbrace{(\Delta x_1^t - i\Delta x_2^t)}_{=e^{\lambda_t+i\theta_t}} + i(n+1)\eta \underbrace{(\Delta x_1^{t-m} - i\Delta x_2^{t-m})}_{=e^{\lambda_{t-m}+i\theta_{t-m}}} - in\eta \underbrace{(\Delta x_1^{t-m-1} - i\Delta x_2^{t-m-1})}_{=e^{\lambda_{t-m-1}+i\theta_{t-m-1}}} \tag{A15}$$

$$= e^{\lambda_t+i\theta_t} + i(n+1)\eta e^{\lambda_{t-m}+i\theta_{t-m}} - in\eta e^{\lambda_{t-m-1}+i\theta_{t-m-1}}. \tag{A16}$$

Let us prove Eq. (A10) by induction. Assume Eqs. (A10) up to (λ_t, θ_t), and then the time delay is evaluated as

$$e^{\lambda_{t-m}+i\theta_{t-m}} = e^{\lambda_t - m(m-n+\frac{1}{2})\eta^2 + \Theta(\eta^4)} e^{i(\theta_t - m\eta + \Theta(\eta^3))} \tag{A17}$$

$$= e^{\lambda_t}(1 + \Theta(\eta^2)) e^{i\theta_t}(1 - im\eta + \Theta(\eta^2) + i\Theta(\eta^3)) \tag{A18}$$

$$= e^{\lambda_t+i\theta_t}(1 - im\eta + \Theta(\eta^2) + i\Theta(\eta^3)). \tag{A19}$$

Using this, we obtain the LHS of Eq. (A16) from the RHS as

$$
\begin{aligned}
e^{\lambda_{t+1}+i\theta_{t+1}} &= e^{\lambda_t+i\theta_t} + i(n+1)\eta e^{\lambda_{t-m}-i\theta_{t-m}} + in\eta e^{\lambda_{t-m-1}+i\theta_{t-m-1}} && \text{(A20)}\\
&= e^{\lambda_t+i\theta_t} + i(n+1)\eta e^{\lambda_t+i\theta_t}(1 - im\eta + \Theta(\eta^2) + i\Theta(\eta^3)) \\
&\quad - in\eta e^{\lambda_t+i\theta_t}(1 - i(m+1)\eta + \Theta(\eta^2) + i\Theta(\eta^3)) && \text{(A21)}\\
&= e^{\lambda_t+i\theta_t}(1 + i\eta + (m-n)\eta^2 + i\Theta(\eta^3) + \Theta(\eta^4)) && \text{(A22)}\\
&= e^{\lambda_t+i\theta_t} e^{i\eta+(m-n+\frac{1}{2})\eta^2+i\Theta(\eta^3)+\Theta(\eta^4)} && \text{(A23)}\\
&= e^{\{\lambda_t+(m-n+\frac{1}{2})\eta^2+\Theta(\eta^4)\}+i\{\theta_t+\eta+\Theta(\eta^3)\}}. && \text{(A24)}
\end{aligned}
$$

By comparing the real and imaginary parts in Eq. (A24), we obtain

$$
\lambda_{t+1} = \lambda_t + \left(m - n + \frac{1}{2}\right)\eta^2 + \Theta(\eta^4), \quad \theta_{t+1} = \theta_t + \eta + \Theta(\eta^3), \tag{A25}
$$

corresponding to Eq. (A10) by rescaling the learning rate $\eta \leftarrow 2\eta$. Because we derived Eq. (A10) for time $t+1$ under the assumption of Eq. (A10) until time t, we have proved the lemma by induction. □

B Proof of Thm. 6

Proof. To introduce meaningful regret in the unconstrained setting, we assume that $\boldsymbol{x}_i \in \mathbb{R}^d$ is bounded as

$$
\|\boldsymbol{x}_i\|_1 \leq \frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{x}_i^t\|_1. \tag{A26}
$$

However, this bound is not essential in the following discussion. Now, social regret is lower-bounded as

$$
\begin{aligned}
\textsc{RegTot}(T) &= \max_{\boldsymbol{x}_i}\sum_{i=1,2}\sum_{t=1}^{T}\left\langle \boldsymbol{x}_i - \boldsymbol{x}_i^t, \boldsymbol{u}_i^t\right\rangle && \text{(A27)}\\
&= \max_{\boldsymbol{x}_1}\sum_{t=1}^{T}\underbrace{\left\langle \boldsymbol{x}_1, \boldsymbol{u}_1^t\right\rangle}_{=\langle \boldsymbol{x}_1, \mathbf{c}\rangle \Delta x_2^t} + \max_{\boldsymbol{x}_2}\sum_{t=1}^{T}\underbrace{\left\langle \boldsymbol{x}_2, \boldsymbol{u}_2^t\right\rangle}_{=\langle \boldsymbol{x}_2, \mathbf{c}\rangle \Delta x_1^t} - \sum_{t=1}^{T}\underbrace{\sum_{i=1,2}\left\langle \boldsymbol{x}_i^t, \boldsymbol{u}_i^t\right\rangle}_{=0} && \text{(A28)}\\
&= \max_{\boldsymbol{x}_1}\langle \boldsymbol{x}_1, \mathbf{c}\rangle \sum_{t=1}^{T}\Delta x_2^t + \max_{\boldsymbol{x}_2}\langle \boldsymbol{x}_2, \mathbf{c}\rangle \sum_{t=1}^{T}\Delta x_1^t && \text{(A29)}\\
&= \|\boldsymbol{x}_1\|_1\left|\sum_{t=1}^{T}\Delta x_2^t\right| + \|\boldsymbol{x}_2\|_1\left|\sum_{t=1}^{T}\Delta x_1^t\right| && \text{(A30)}\\
&\geq D\sum_{i=1,2}\left|\sum_{t=1}^{T}\Delta x_i^t\right| && \text{(A31)}\\
&\geq D\sqrt{\sum_{i=1,2}\left|\sum_{t=1}^{T}\Delta x_i^t\right|^2}. && \text{(A32)}
\end{aligned}
$$

Here, we define the scale of their strategy space $D := \min_{i=1,2}\|\boldsymbol{x}_i\|_1$.

By Lem. A1, we immediately obtain

$$
\lambda_t = \lambda_1 + \underbrace{\left(m - n + \frac{1}{2}\right)}_{=:\alpha}\eta^2(t-1) + \Theta(\eta^4(t-1)), \quad \theta_t = \theta_1 + \eta(t-1) + \Theta(\eta^3(t-1)). \tag{A33}
$$

Using this, we obtain

$$\sum_{t=1}^{T}\Delta x_1^t - i\sum_{t=1}^{T}\Delta x_2^t = \sum_{t=1}^{T} e^{\lambda_t+i\theta_t} \tag{A34}$$

$$= \frac{e^{\lambda_1+i\theta_1} - e^{\lambda_{T+1}+i\theta_{T+1}}}{1-e^{i\eta+\alpha\eta^2+i\Theta(\eta^3)+\Theta(\eta^4)}} \tag{A35}$$

$$= e^{\lambda_1+i\theta_1}\frac{1-e^{(\alpha\eta^2+i\eta)T+i\Theta(\eta^3 T)+\Theta(\eta^4 T)}}{1-e^{\alpha\eta^2+i\eta+i\Theta(\eta^3)+\Theta(\eta^4)}}. \tag{A36}$$

Here, we used

$$e^{\lambda_1+i\theta_1} - e^{\lambda_{T+1}+i\theta_{T+1}} = \sum_{t=1}^{T}(e^{\lambda_t+i\theta_t} - e^{\lambda_{t+1}+i\theta_{t+1}}) \tag{A37}$$

$$= \sum_{t=1}^{T} e^{\lambda_t+i\theta_t}(1-e^{i\eta+\alpha\eta^2+i\Theta(\eta^3)+\Theta(\eta^4)}) \tag{A38}$$

$$= (1-e^{i\eta+\alpha\eta^2+i\Theta(\eta^3)+\Theta(\eta^4)})\sum_{t=1}^{T} e^{\lambda_t+i\theta_t}. \tag{A39}$$

Finally, we evaluate

$$\sum_{i=1,2}\left|\sum_{t=1}^{T}\Delta x_i^t\right|^2 = \left(\sum_{t=1}^{T}\Delta x_1^t - i\sum_{t=1}^{T}\Delta x_2^t\right)\left(\sum_{t=1}^{T}\Delta x_1^t + i\sum_{t=1}^{T}\Delta x_2^t\right) \tag{A40}$$

$$= \frac{e^{\lambda_1+i\theta_1} - e^{\lambda_{T+1}+i\theta_{T+1}}}{1-e^{i\eta+\alpha\eta^2+i\Theta(\eta^3)+\Theta(\eta^4)}}\frac{e^{\lambda_1-i\theta_1} - e^{\lambda_{T+1}-i\theta_{T+1}}}{1-e^{-i\eta+\alpha\eta^2-i\Theta(\eta^3)+\Theta(\eta^4)}} \tag{A41}$$

$$= e^{\lambda_1+i\theta_1}\frac{1-e^{(i\eta+\alpha\eta^2)T+i\Theta(\eta^3 T)+\Theta(\eta^4 T)}}{1-e^{i\eta+\alpha\eta^2+i\Theta(\eta^3)+\Theta(\eta^4)}} \times e^{\lambda_1-i\theta_1}\frac{1-e^{(-i\eta+\alpha\eta^2)T-i\Theta(\eta^3 T)+\Theta(\eta^4 T)}}{1-e^{-i\eta+\alpha\eta^2-i\Theta(\eta^3)+\Theta(\eta^4)}} \tag{A42}$$

$$= e^{2\lambda_1}\frac{1-2e^{\alpha\eta^2 T+\Theta(\eta^4 T)}\cos(\eta T+\Theta(\eta^3 T)) + e^{2\alpha\eta^2 T+\Theta(\eta^4 T)}}{1-2e^{\alpha\eta^2+\Theta(\eta^4)}\cos(\eta+\Theta(\eta^3)) + e^{2\alpha\eta^2+\Theta(\eta^4)}} \tag{A43}$$

$$= e^{2\lambda_1}\frac{1+\Theta(1)e^{\alpha\eta^2 T+\Theta(\eta^4 T)} + e^{2\alpha\eta^2 T+\Theta(\eta^4 T)}}{\eta^2+\Theta(\eta^4)} \tag{A44}$$

$$= \Omega(\frac{\max\{1, e^{2\alpha\eta^2 T}\}}{\eta^2}) \tag{A45}$$

$$\geq \Omega(T). \tag{A46}$$

Here, we used

$$1-2e^{\alpha\eta^2+\Theta(\eta^4)}\cos(\eta+\Theta(\eta^3)) + e^{2\alpha\eta^2+\Theta(\eta^4)} \tag{A47}$$

$$= 1-2\left(1+\alpha\eta^2+\Theta(\eta^4)\right)\left(1-\frac{1}{2}\eta^2+\Theta(\eta^4)\right)+\left(1+2\alpha\eta^2+\Theta(\eta^4)\right) \tag{A48}$$

$$= \eta^2+\Theta(\eta^4). \tag{A49}$$

The equality holds when and only when $\eta = \Omega(1/\sqrt{T})$. In conclusion, we obtain the lower bound of social regret as $\textsc{RegTot}(T) \geq \Omega(D\sqrt{T})$, leading to $\textsc{RegTot}(T) \geq \Omega(\sqrt{T})$ by ignoring the scale of the strategy space. □

C Proof of Thm. 7

Proof. In Matching Pennies, the Nash equilibrium $(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*)$ satisfies

$$\boldsymbol{A}\boldsymbol{x}_1^* = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_2^* = \boldsymbol{0} \;\Leftrightarrow\; \Delta x_1^* = \Delta x_2^* = 0. \tag{A50}$$

We discuss the distance from the Nash equilibria, which is formulated as

$$\textsc{Dis}(T) = \min_{\boldsymbol{x}^*} \frac{1}{2} \sum_{i=1,2} \|\boldsymbol{x}_i^T - \boldsymbol{x}_i^*\|_2^2 \overset{\text{a}}{=} \frac{1}{2} \sum_{i=1,2} \frac{1}{4} |\Delta x_i^T|^2 \underbrace{\|\mathbf{c}\|_2^2}_{=2} = \frac{1}{4} \sum_{i=1,2} |\Delta x_i^T|^2 \tag{A51}$$

$$\overset{\text{b}}{=} \frac{1}{4} e^{\lambda_T} = \frac{1}{4} e^{\lambda_0 + \alpha \eta^2 T + \Theta(\eta^4 T)}. \tag{A52}$$

In (a), we used the nearest Nash equilibrium from $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ as

$$\arg\min_{\boldsymbol{x}_i^*} \|\boldsymbol{x}_i - \boldsymbol{x}_i^*\|_2^2 = \boldsymbol{x}_i - \frac{1}{2} \Delta x_i \mathbf{c}. \tag{A53}$$

In (b), we applied Lem. A1. For OFTRL $n = 1$ with a time delay $m \geq 1$, $\alpha = n - m + \frac{1}{2} > 0$ holds. For sufficiently small η, the term of $\Theta(\eta^4 T)$ is negligible, and D^T diverges with the final time T. We have proved Thm. 7. □

D Proof of Thm. 10

Proof. First, for all $\tilde{\mathbf{f}} \in \mathbb{R}^d$, we define $\mathbf{f}$ and F as

$$\mathbf{f} := \arg\max_{\boldsymbol{x}_i \in \mathcal{X}_i} \eta \langle \boldsymbol{x}_i, \tilde{\mathbf{f}} \rangle - h(\boldsymbol{x}_i), \quad \mathrm{F} := \max_{\boldsymbol{x}_i \in \mathcal{X}_i} \eta \langle \boldsymbol{x}_i, \tilde{\mathbf{f}} \rangle - h(\boldsymbol{x}_i). \tag{A54}$$

In this notation, we define

$$\tilde{\boldsymbol{g}}_i^t := \sum_{s=1}^{t} \boldsymbol{u}_i^s, \quad \tilde{\boldsymbol{x}}_i^t := \sum_{s=1}^{t-m-1} \boldsymbol{u}_i^s + (m+1) \boldsymbol{u}_i^{t-m-1}. \tag{A55}$$

Here, $\tilde{\boldsymbol{g}}_i^t$ represents i's cumulative rewards until time t. Also, $\tilde{\boldsymbol{x}}_i^t$ is a prediction of $\tilde{\boldsymbol{g}}_i^t$ by delayed feedback because all the unobservable rewards $\boldsymbol{u}_i^{t-m}, \cdots, \boldsymbol{u}_i^t$ are replaced by the latest observable reward $\boldsymbol{u}_i^{t-m-1}$. According to $\tilde{\boldsymbol{g}}_i^t$ and $\tilde{\boldsymbol{x}}_i^t$, we use $(\boldsymbol{g}_i^t, G_i^t)$ and $(\boldsymbol{x}_i^t, X_i^t)$. Note that $\boldsymbol{x}_i^t$ corresponds to the WOFTRL algorithm for the case of $n = m + 1$.

Now, between different $\tilde{\mathbf{f}}, \tilde{\mathbf{f}}' \in \mathbb{R}^d$, the following lemma holds.

Lemma A2 (First-order optimality condition)**.** *For all $\tilde{\mathbf{f}}, \tilde{\mathbf{f}}' \in \mathbb{R}^d$ and their corresponding $(\mathbf{f}, \mathrm{F}), (\mathbf{f}', \mathrm{F}')$, the following inequalities hold*

$$\frac{1}{2} \|\mathbf{f}' - \mathbf{f}\|_2^2 \leq -\mathrm{F}' + \mathrm{F} + \eta \langle \mathbf{f}', \tilde{\mathbf{f}}' - \tilde{\mathbf{f}} \rangle, \tag{FO1}$$

$$\|\mathbf{f}' - \mathbf{f}\|_2^2 \leq \eta \langle \mathbf{f}' - \mathbf{f}, \tilde{\mathbf{f}}' - \tilde{\mathbf{f}} \rangle, \tag{FO2}$$

$$\|\mathbf{f}' - \mathbf{f}\|_2 \leq \eta \|\tilde{\mathbf{f}}' - \tilde{\mathbf{f}}\|_2. \tag{FO3}$$

Now, the payoff at time t can be divided as follows;

$$-\left\langle \boldsymbol{x}_i^t, \boldsymbol{u}_i^t \right\rangle = \left\langle \boldsymbol{g}_i^t - \boldsymbol{x}_i^t, \tilde{\boldsymbol{g}}_i^t - \tilde{\boldsymbol{x}}_i^t \right\rangle - \left\langle \boldsymbol{x}_i^t, \tilde{\boldsymbol{x}}_i^t - \tilde{\boldsymbol{g}}_i^{t-1} \right\rangle - \left\langle \boldsymbol{g}_i^t, \tilde{\boldsymbol{g}}_i^t - \tilde{\boldsymbol{x}}_i^t \right\rangle. \tag{A56}$$

The first term is upper-bounded as

$$\sum_{t=1}^{T} \left\langle \boldsymbol{g}_i^t - \boldsymbol{x}_i^t, \tilde{\boldsymbol{g}}_i^t - \tilde{\boldsymbol{x}}_i^t \right\rangle \tag{A57}$$

$$\leq \sum_{t=1}^{T} \|\boldsymbol{g}_i^t - \boldsymbol{x}_i^t\|_2 \|\tilde{\boldsymbol{g}}_i^t - \tilde{\boldsymbol{x}}_i^t\|_2 \tag{A58}$$

$$\leq \eta \sum_{t=1}^{T} \|\tilde{\boldsymbol{g}}_i^t - \tilde{\boldsymbol{x}}_i^t\|_2^2 \tag{A59}$$

$$= \eta \sum_{t=1}^{T} \left\| \sum_{s=t-m}^{t} \boldsymbol{u}_i^s - (m+1)\boldsymbol{u}_i^{t-m-1} \right\|_2^2 \tag{A60}$$

$$= \eta \sum_{t=1}^{T} \left\| \sum_{s=t-m}^{t} (t-s+1)(\boldsymbol{u}_i^s - \boldsymbol{u}_i^{s-1}) \right\|_2^2 \tag{A61}$$

$$\leq \eta \sum_{t=1}^{T} \left(\sum_{s=t-m}^{t} (t-s+1) \|\boldsymbol{u}_i^s - \boldsymbol{u}_i^{s-1}\|_2 \right)^2 \tag{A62}$$

$$= \eta \sum_{t=1}^{T} \sum_{s=t-m}^{t} (t-s+1) \sum_{s'=t-m}^{t} (t-s'+1) \|\boldsymbol{u}_i^s - \boldsymbol{u}_i^{s-1}\|_2 \|\boldsymbol{u}_i^{s'} - \boldsymbol{u}_i^{s'-1}\|_2 \tag{A63}$$

$$\leq \eta \sum_{t=1}^{T} \sum_{s=t-m}^{t} (t-s+1) \sum_{s'=t-m}^{t} (t-s'+1) \frac{1}{2} \left(\|\boldsymbol{u}_i^s - \boldsymbol{u}_i^{s-1}\|_2^2 + \|\boldsymbol{u}_i^{s'} - \boldsymbol{u}_i^{s'-1}\|_2^2 \right) \tag{A64}$$

$$\leq \eta \sum_{t=1}^{T} \sum_{s=t-m}^{t} (t-s+1) \sum_{s'=t-m}^{t} (t-s'+1) \|\boldsymbol{u}_i^s - \boldsymbol{u}_i^{s-1}\|_2^2 \tag{A65}$$

$$= \lambda\eta \sum_{t=1}^{T} \sum_{s=t-m}^{t} (t-s+1) \|\boldsymbol{u}_i^s - \boldsymbol{u}_i^{s-1}\|_2^2 \tag{A66}$$

$$\leq \lambda^2 \eta \sum_{t=1}^{T} \|\boldsymbol{u}_i^t - \boldsymbol{u}_i^{t-1}\|_2^2. \tag{A67}$$

In the final two lines, we twice used

$$\sum_{s=t-m}^{t} (t-s+1) = \frac{(m+1)(m+2)}{2} =: \lambda. \tag{A68}$$

Furthermore, the second and third terms are also upper-bounded as

$$-\sum_{t=1}^{T} \left\langle \boldsymbol{x}_i^t, \tilde{\boldsymbol{x}}_i^t - \tilde{\boldsymbol{g}}_i^{t-1} \right\rangle - \sum_{t=1}^{T} \left\langle \boldsymbol{g}_i^t, \tilde{\boldsymbol{g}}_i^t - \tilde{\boldsymbol{x}}_i^t \right\rangle \tag{A69}$$

$$\overset{\text{a}}{\leq} \sum_{t=1}^{T} \frac{1}{\eta} \left(-X_i^t + G_i^{t-1} - \frac{1}{2} \|\boldsymbol{x}_i^t - \boldsymbol{g}_i^{t-1}\|_2^2 \right) + \sum_{t=1}^{T} \frac{1}{\eta} \left(-G_i^t + X_i^t - \frac{1}{2} \|\boldsymbol{g}_i^t - \boldsymbol{x}_i^t\|_2^2 \right) \tag{A70}$$

$$= \frac{-G_i^T + G_i^0}{\eta} - \frac{1}{2\eta} \sum_{t=1}^{T} (\|\boldsymbol{x}_i^t - \boldsymbol{g}_i^{t-1}\|_2^2 + \|\boldsymbol{g}_i^t - \boldsymbol{x}_i^t\|_2^2) \tag{A71}$$

$$\overset{\text{b}}{\leq} -\max_{\boldsymbol{x}_i \in \mathcal{X}_i} \left\langle \boldsymbol{x}_i, \sum_{t=1}^{T} \boldsymbol{u}_i^t \right\rangle + \frac{h_{\max}}{\eta} - \frac{1}{2\eta} \sum_{t=1}^{T} (\|\boldsymbol{x}_i^t - \boldsymbol{g}_i^{t-1}\|_2^2 + \|\boldsymbol{g}_i^t - \boldsymbol{x}_i^t\|_2^2) \tag{A72}$$

$$\overset{\text{c}}{\leq} -\max_{\boldsymbol{x}_i \in \mathcal{X}_i} \left\langle \boldsymbol{x}_i, \sum_{t=1}^{T} \boldsymbol{u}_i^t \right\rangle + \frac{h_{\max}}{\eta} - \frac{1}{4\eta} \sum_{t=1}^{T} \|\boldsymbol{x}_i^t - \boldsymbol{x}_i^{t-1}\|_2^2. \tag{A73}$$

In (a), we used Eq. (FO1) for $(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') = (\tilde{\boldsymbol{x}}_i^t, \tilde{\boldsymbol{g}}_i^{t-1})$ and $(\tilde{\boldsymbol{g}}_i^t, \tilde{\boldsymbol{x}}_i^t)$. In (b), we used

$$0 \le G_i^0, \quad -\frac{G_i^T}{\eta} \le -\max_{\boldsymbol{x}_i \in \mathcal{X}_i} \left\langle \boldsymbol{x}_i, \sum_{t=1}^T \boldsymbol{u}_i^t \right\rangle + \frac{h_{\max}}{\eta}, \quad h_{\max} := \max_{\boldsymbol{x}_i \in \mathcal{X}_i} h(\boldsymbol{x}_i). \tag{A74}$$

In (c), we used

$$\frac{1}{2}\sum_{t=1}^T \|\boldsymbol{x}_i^t - \boldsymbol{x}_i^{t-1}\|^2 \le \sum_{t=1}^T (\|\boldsymbol{x}_i^t - \boldsymbol{g}_i^{t-1}\|^2 + \|\boldsymbol{g}_i^{t-1} - \boldsymbol{x}_i^{t-1}\|^2) \tag{A75}$$

$$\le \sum_{t=1}^T (\|\boldsymbol{x}_i^t - \boldsymbol{g}_i^{t-1}\|^2 + \|\boldsymbol{g}_i^t - \boldsymbol{x}_i^t\|^2). \tag{A76}$$

Finally, we obtain the upper bound of the individual regret $\textsc{Reg}_i(T)$ as

$$\textsc{Reg}_i(T) = \max_{\boldsymbol{x}_i \in \mathcal{X}_i} \sum_{t=1}^T \left\langle \boldsymbol{x}_i - \boldsymbol{x}_i^t, \boldsymbol{u}_i^t \right\rangle \tag{A77}$$

$$= \max_{\boldsymbol{x}_i \in \mathcal{X}_i} \left\langle \boldsymbol{x}_i, \sum_{t=1}^T \boldsymbol{u}_i^t \right\rangle - \sum_{t=1}^T \left\langle \boldsymbol{x}_i^t, \boldsymbol{u}_i^t \right\rangle \tag{A78}$$

$$\le \frac{h_{\max}}{\eta} + \lambda^2 \eta \sum_{t=1}^T \|\boldsymbol{u}_i^t - \boldsymbol{u}_i^{t-1}\|_2^2 - \frac{1}{4\eta}\sum_{t=1}^T \|\boldsymbol{x}_i^t - \boldsymbol{x}_i^{t-1}\|_2^2. \tag{A79}$$

Thus, we have proven that the regret satisfies the RVU property with $\alpha = h_{\max}/\eta$, $\beta = \lambda^2\eta$, and $\gamma = 1/(4\eta)$. □

D.1 Proof of Lem. A2

Proof.

$$\mathrm{F} = \eta\langle \mathbf{f}, \tilde{\mathbf{f}} \rangle - h(\mathbf{f}) \tag{A80}$$

$$\ge \eta\langle \mathbf{f}', \tilde{\mathbf{f}} \rangle - h(\mathbf{f}') + \frac{1}{2}\|\mathbf{f}' - \mathbf{f}\|_2^2 \tag{A81}$$

$$= \eta\langle \mathbf{f}', \tilde{\mathbf{f}}' \rangle - h(\mathbf{f}') - \eta\langle \mathbf{f}', \tilde{\mathbf{f}}' - \tilde{\mathbf{f}} \rangle + \frac{1}{2}\|\mathbf{f}' - \mathbf{f}\|_2^2 \tag{A82}$$

$$= \mathrm{F}' - \eta\langle \mathbf{f}', \tilde{\mathbf{f}}' - \tilde{\mathbf{f}} \rangle + \frac{1}{2}\|\mathbf{f}' - \mathbf{f}\|_2^2. \tag{A83}$$

In the inequality, we used 1-strict convexity of h. This corresponds to Eq. (FO1).

Furthermore, by summing up

$$\left(\frac{1}{2}\|\mathbf{f}' - \mathbf{f}\|_2^2 \le -\mathrm{F}' + \mathrm{F} + \eta\langle \mathbf{f}', \tilde{\mathbf{f}}' - \tilde{\mathbf{f}} \rangle, \quad \frac{1}{2}\|\mathbf{f} - \mathbf{f}'\|_2^2 \le -\mathrm{F} + \mathrm{F}' + \eta\langle \mathbf{f}, \tilde{\mathbf{f}} - \tilde{\mathbf{f}}' \rangle\right) \tag{A84}$$

$$\Rightarrow \|\mathbf{f}' - \mathbf{f}\|_2^2 \le \eta\langle \mathbf{f}' - \mathbf{f}, \tilde{\mathbf{f}}' - \tilde{\mathbf{f}} \rangle. \tag{A85}$$

This corresponds to Eq. (FO2).

Finally, by the Cauchy–Schwarz inequality, we obtain

$$\|\mathbf{f}' - \mathbf{f}\|_2^2 \le \eta\langle \mathbf{f}' - \mathbf{f}, \tilde{\mathbf{f}}' - \tilde{\mathbf{f}} \rangle \le \eta\|\mathbf{f}' - \mathbf{f}\|_2 \|\tilde{\mathbf{f}}' - \tilde{\mathbf{f}}\|_2 \tag{A86}$$

$$\Rightarrow \|\mathbf{f}' - \mathbf{f}\|_2 \le \eta\|\tilde{\mathbf{f}}' - \tilde{\mathbf{f}}\|_2. \tag{A87}$$

This corresponds to Eq. (FO3). □

E Proof of Cor. 13

Proof. We prove the last-iterate convergence by the following five steps.

Step 1: Equivalence between generalized FTRL and generalized MD When h is a Legendre function, the strategies always exist in the interior of the strategy space, i.e., $\boldsymbol{x}^t \in \mathrm{int}(\mathcal{X})$ for all t. Furthermore, the dynamics of generalized FTRL become equivalent to those of generalized MD as follows.

Definition A3 (Generalized Mirror Descent)**.** *With the time delay of $m \in \mathbb{N}$ steps, generalized Mirror Descent is formulated based on the prox-mapping $\boldsymbol{P}$ as follows*

$$\boldsymbol{x}^{t+1} = \boldsymbol{P}(\hat{\boldsymbol{x}}^{t-m+1}, \boldsymbol{m}^t), \quad \hat{\boldsymbol{x}}^{t+1} = \boldsymbol{P}(\hat{\boldsymbol{x}}^t, \boldsymbol{u}^t), \quad \boldsymbol{P}(\hat{\boldsymbol{x}}, \boldsymbol{u}) := \arg\max_{\boldsymbol{x} \in \mathcal{X}} \eta \langle \boldsymbol{x}, \boldsymbol{u} \rangle - D(\boldsymbol{x}, \hat{\boldsymbol{x}}), \tag{A88}$$

where $D(\boldsymbol{x}, \boldsymbol{x}')$ is the Bregman divergence between $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, defined as

$$D(\boldsymbol{x}, \boldsymbol{x}') := h(\boldsymbol{x}) - \langle \boldsymbol{x} - \boldsymbol{x}', \nabla h(\boldsymbol{x}') \rangle - h(\boldsymbol{x}'). \tag{A89}$$

Lemma A4 (Equivalence between generalized FTRL and generalized MD)**.** *Suppose that h is a Legendre function, then the time series of $\{\boldsymbol{x}^t\}_{t=1,\dots}$ are equivalent between Defs. 2 and A3.*

Step 2: Existence of Lyapunov function In accordance with WOFTRL, suppose $\boldsymbol{m}^t = n\boldsymbol{u}^{t-m}$ for $n = m + 1$ in generalized MD, then there exists a Lyapunov function, which is defined for all $\boldsymbol{x}^* \in \mathcal{X}^*$ as

$$V^t(\boldsymbol{x}^*) := D(\boldsymbol{x}^*, \hat{\boldsymbol{x}}^t) + \frac{1}{2}\Big\{ D(\hat{\boldsymbol{x}}^t, \boldsymbol{x}^{t-1}) + \sum_{r=1}^{n-1} \frac{n-r}{n} (D(\boldsymbol{x}^{t-r}, \hat{\boldsymbol{x}}^{t-r}) + D(\hat{\boldsymbol{x}}^{t-r}, \boldsymbol{x}^{t-r-1})) \Big\}. \tag{A90}$$

For $\eta \leq 1/(\sqrt{8}n^2 L)$, we see that this $V^t(\boldsymbol{x}^*)$ monotonically decreases for all $\boldsymbol{x}^* \in \mathcal{X}^*$, as follows.

Lemma A5 (Lyapunov function under time delay)**.** *Suppose generalized MD with $n = m + 1$ and $\eta \leq 1/(\sqrt{8}n^2 L)$, then for all $\boldsymbol{x}^* \in \mathcal{X}^*$, $V^t(\boldsymbol{x}^*)$ is monotonic decreasing as*

$$V^{t+1}(\boldsymbol{x}^*) \leq V^t(\boldsymbol{x}^*) - \frac{1}{2} D(\hat{\boldsymbol{x}}^{t+1}, \boldsymbol{x}^t) - \frac{1}{2} D(\boldsymbol{x}^t, \hat{\boldsymbol{x}}^t). \tag{A91}$$

By summing up Eq. (A91), for all $\boldsymbol{x}^* \in \mathcal{X}^*$, we derive

$$V^1(\boldsymbol{x}^*) \geq \lim_{T \to \infty} \left[V^T(\boldsymbol{x}^*) + \frac{1}{2} \sum_{t=1}^{T} (D(\hat{\boldsymbol{x}}^{t+1}, \boldsymbol{x}^t) + D(\boldsymbol{x}^t, \hat{\boldsymbol{x}}^t)) \right] \tag{A92}$$

$$\geq \frac{1}{2} \sum_{t=1}^{\infty} (D(\hat{\boldsymbol{x}}^{t+1}, \boldsymbol{x}^t) + D(\boldsymbol{x}^t, \hat{\boldsymbol{x}}^t)) \tag{A93}$$

$$\geq \frac{1}{4} \sum_{t=1}^{\infty} (\|\hat{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^t\|_2^2 + \|\boldsymbol{x}^t - \hat{\boldsymbol{x}}^t\|_2^2) \tag{A94}$$

$$\geq \frac{1}{8} \sum_{t=1}^{\infty} (\|\hat{\boldsymbol{x}}^{t+1} - \hat{\boldsymbol{x}}^t\|_2^2). \tag{A95}$$

By Eqs. (A94) and (A95), we obtain

$$\lim_{t \to \infty} \|\hat{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^t\|_2 = \lim_{t \to \infty} \|\hat{\boldsymbol{x}}^{t+1} - \hat{\boldsymbol{x}}^t\|_2 = 0. \tag{A96}$$

Step 3: Existence of subsequence converging to equilibrium Since $\mathcal{X}$ is a compact space, the Bolzano-Weierstrass theorem can be applied, and there exists a subsequence $\{\boldsymbol{x}^{t_l}\}_{l=1,\dots}$ equipped with its limit. In other words, there exists $\tilde{\boldsymbol{x}} \in \mathcal{X}$ such that

$$\lim_{l \to \infty} \boldsymbol{x}^{t_l} = \tilde{\boldsymbol{x}}. \tag{A97}$$

By using Eqs. (A96), we further obtain

$$\lim_{l \to \infty} \hat{\boldsymbol{x}}^{t_l+1} = \tilde{\boldsymbol{x}}, \quad \lim_{l \to \infty} \hat{\boldsymbol{x}}^{t_l} = \tilde{\boldsymbol{x}}. \tag{A98}$$

Thus, by substituting Eqs. (A97) and (A98) into the prox-mapping (the second one in Eqs. (A88)), we show that this $\tilde{\boldsymbol{x}}$ satisfies the fixed point condition of the prox-mapping, i.e.,

$$\tilde{\boldsymbol{x}} = \lim_{l \to \infty} \hat{\boldsymbol{x}}^{t_l+1} = \lim_{l \to \infty} \boldsymbol{P}(\hat{\boldsymbol{x}}^{t_l}, \boldsymbol{u}(\boldsymbol{x}^{t_l})) \overset{\text{a}}{=} \boldsymbol{P}\left(\lim_{l \to \infty} \hat{\boldsymbol{x}}^{t_l}, \lim_{l \to \infty} \boldsymbol{u}(\boldsymbol{x}^{t_l}) \right) = \boldsymbol{P}(\tilde{\boldsymbol{x}}, \boldsymbol{u}(\tilde{\boldsymbol{x}})). \tag{A99}$$

In (a), we used the continuity of prox-mapping $\boldsymbol{P}$, as the following lemma shows.

Lemma A6 (Continuity of prox-mapping)**.** *The prox-mapping $\boldsymbol{P}(\hat{\boldsymbol{x}}, \boldsymbol{u})$ is continuous for $\hat{\boldsymbol{x}} \in \mathrm{int}(\mathcal{X})$ and $\boldsymbol{u} \in \mathbb{R}^d$.*

By the first-order optimality condition, this fixed point satisfies the Nash equilibrium condition [10], i.e., $\tilde{\boldsymbol{x}} \in \mathcal{X}^*$. In conclusion, we obtain a subsequence $\{\boldsymbol{x}^{t_l}\}_{l=1,\cdots}$ which converges to one Nash equilibrium $\boldsymbol{x}^* \in \mathcal{X}^*$.

Step 4: Convergence guarantee by Lyapunov function Let $\boldsymbol{x}^*$ denote the specific Nash equilibrium, to which the subsequence $\{t_l\}_{l=1,\cdots}$ converges. Because $V^t(\boldsymbol{x}^*)$ is both lower-bounded by 0 and monotonic decreasing, there exists its limits which corresponds to its limit inferior, and we obtain

$$0 \leq \lim_{t\to\infty} V^t(\boldsymbol{x}^*) = \liminf_{t\to\infty} V^t(\boldsymbol{x}^*) \leq \lim_{l\to\infty} V^{t_l}(\boldsymbol{x}^*) = 0. \tag{A100}$$

This trivially leads to

$$\lim_{t\to\infty} D(\boldsymbol{x}^*, \hat{\boldsymbol{x}}^t) \leq \lim_{t\to\infty} V^t(\boldsymbol{x}^*) = 0 \Rightarrow \lim_{t\to\infty} \hat{\boldsymbol{x}}^t = \boldsymbol{x}^*. \tag{A101}$$

By further using $\lim_{t\to\infty} \|\hat{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^t\|_2 = 0$, we ultimately obtain

$$\lim_{t\to\infty} \boldsymbol{x}^t = \boldsymbol{x}^*. \tag{A102}$$

indicating the last-iterate convergence to one of the Nash equilibria. □

E.1 Proof of Lem. A4

Proof. First, GFTRL is obviously rewritten as

$$\boldsymbol{x}^{t+1} = \arg\max_{\boldsymbol{x}\in\mathcal{X}} \eta\big\langle \boldsymbol{x}, \boldsymbol{m}^t \big\rangle - F(\boldsymbol{x}, \tilde{\boldsymbol{g}}^{t-m+1}). \tag{A103}$$

Here, $F(\boldsymbol{x}, \tilde{\boldsymbol{g}})$ is the Fenchel coupling between $\boldsymbol{x} \in \mathcal{X}$ and $\tilde{\boldsymbol{g}} \in \mathbb{R}^d$, which is defined as

$$F(\boldsymbol{x}, \tilde{\boldsymbol{g}}) := h(\boldsymbol{x}) - \big\langle \boldsymbol{x}, \tilde{\boldsymbol{g}} \big\rangle + h^*(\tilde{\boldsymbol{g}}), \quad h^*(\tilde{\boldsymbol{g}}) := \max_{\boldsymbol{x}\in\mathcal{X}} \big\langle \boldsymbol{x}, \tilde{\boldsymbol{g}} \big\rangle - h(\boldsymbol{x}). \tag{A104}$$

Hence, it is sufficient to see the equivalence between $D(\boldsymbol{x}, \hat{\boldsymbol{x}}^t)$ and $F(\boldsymbol{x}, \tilde{\boldsymbol{g}}^t)$, which is proven as

$$\begin{aligned}
D(\boldsymbol{x}, \hat{\boldsymbol{x}}^t) &= h(\boldsymbol{x}) - \big\langle \boldsymbol{x} - \hat{\boldsymbol{x}}^t, \nabla h(\hat{\boldsymbol{x}}^t) \big\rangle + h(\hat{\boldsymbol{x}}^t) && \text{(A105)}\\
&\overset{\text{a}}{=} h(\boldsymbol{x}) - \big\langle \boldsymbol{x} - \hat{\boldsymbol{x}}^t, \boldsymbol{y}^t \big\rangle + h(\hat{\boldsymbol{x}}^t) && \text{(A106)}\\
&= h(\boldsymbol{x}) - \big\langle \boldsymbol{x}, \tilde{\boldsymbol{g}}^t \big\rangle - (\big\langle \hat{\boldsymbol{x}}^t, \tilde{\boldsymbol{g}}^t \big\rangle - h(\hat{\boldsymbol{x}}^t)) && \text{(A107)}\\
&\overset{\text{b}}{=} h(\boldsymbol{x}) - \big\langle \boldsymbol{x}, \tilde{\boldsymbol{g}}^t \big\rangle - h^*(\tilde{\boldsymbol{g}}^t) && \text{(A108)}\\
&= F(\boldsymbol{x}, \tilde{\boldsymbol{g}}^t). && \text{(A109)}
\end{aligned}$$

In (a), we applied the first-order optimality condition for $\hat{\boldsymbol{x}}^{t+1}$ in Eq. (A88) as

$$\begin{aligned}
&\big\langle \eta \boldsymbol{u}^t - \nabla h(\hat{\boldsymbol{x}}^{t+1}) + \nabla h(\hat{\boldsymbol{x}}^t), \boldsymbol{x} - \boldsymbol{x}' \big\rangle = 0 && \text{(A110)}\\
&\Leftrightarrow \big\langle \nabla h(\hat{\boldsymbol{x}}^{t+1}), \boldsymbol{x} - \boldsymbol{x}' \big\rangle = \big\langle \nabla h(\hat{\boldsymbol{x}}^t) + \eta \boldsymbol{u}^t, \boldsymbol{x} - \boldsymbol{x}' \big\rangle && \text{(A111)}\\
&\Rightarrow \big\langle \nabla h(\hat{\boldsymbol{x}}^{t+1}), \boldsymbol{x} - \boldsymbol{x}' \big\rangle = \big\langle \tilde{\boldsymbol{g}}^{t+1}, \boldsymbol{x} - \boldsymbol{x}' \big\rangle && \text{(A112)}\\
&\Leftrightarrow \big\langle \nabla h(\hat{\boldsymbol{x}}^t), \boldsymbol{x} - \boldsymbol{x}' \big\rangle = \big\langle \tilde{\boldsymbol{g}}^t, \boldsymbol{x} - \boldsymbol{x}' \big\rangle, && \text{(A113)}
\end{aligned}$$

for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. This means that the projection of $\nabla h(\hat{\boldsymbol{x}}^{t+1})$ on $\mathcal{X}$ is equal to that of $\tilde{\boldsymbol{g}}^{t+1}$. In (b), we considered that $\hat{\boldsymbol{x}}^{t+1}$ satisfies the first-order optimality condition for $h^*(\tilde{\boldsymbol{g}}^{t+1})$ as

$$\begin{aligned}
h^*(\boldsymbol{y}^t) &= \max_{\boldsymbol{x}\in\mathcal{X}} \big\langle \boldsymbol{x}, \boldsymbol{y}^t \big\rangle - h(\boldsymbol{x}) && \text{(A114)}\\
&= \big\langle \hat{\boldsymbol{x}}^t, \boldsymbol{y}^t \big\rangle - h(\hat{\boldsymbol{x}}^t). && \text{(A115)}
\end{aligned}$$

Thus, it has been proven that the dynamics of $\boldsymbol{x}^t$ are equivalent between generalized FTRL and generalized MD. □

E.2 Proof of Lem. A5

Proof. We prove that $V^t(\boldsymbol{x}^*)$ monotonically decreases as

$$V^{t+1}(\boldsymbol{x}^*) = D(\boldsymbol{x}^*, \hat{\boldsymbol{x}}^{t+1}) + \frac{1}{2}\Big\{D(\hat{\boldsymbol{x}}^{t+1}, \boldsymbol{x}^t) + \sum_{r=1}^{n-1} \frac{n-r}{n}(D(\boldsymbol{x}^{t-r+1}, \hat{\boldsymbol{x}}^{t-r+1}) + D(\hat{\boldsymbol{x}}^{t-r+1}, \boldsymbol{x}^{t-r}))\Big\} \tag{A116}$$

$$\overset{\text{a}}{=} D(\boldsymbol{x}^*, \hat{\boldsymbol{x}}^t) + \frac{1}{2}\sum_{r=1}^{n-1} \frac{n-r}{n}(D(\boldsymbol{x}^{t-r+1}, \hat{\boldsymbol{x}}^{t-r+1}) + D(\hat{\boldsymbol{x}}^{t-r+1}, \boldsymbol{x}^{t-r})) + \big\langle \nabla h(\hat{\boldsymbol{x}}^{t+1}) - \nabla h(\boldsymbol{x}^t), \hat{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^t \big\rangle - \frac{1}{2}D(\hat{\boldsymbol{x}}^{t+1}, \boldsymbol{x}^t) - D(\boldsymbol{x}^t, \hat{\boldsymbol{x}}^t) \tag{A117}$$

$$\overset{\text{b}}{\leq} D(\boldsymbol{x}^*, \hat{\boldsymbol{x}}^t) + \frac{1}{2}\sum_{r=1}^{n-1} \frac{n-r}{n}(D(\boldsymbol{x}^{t-r+1}, \hat{\boldsymbol{x}}^{t-r+1}) + D(\hat{\boldsymbol{x}}^{t-r+1}, \boldsymbol{x}^{t-r})) + \frac{1}{2n}\sum_{r=1}^{n}(D(\boldsymbol{x}^{t-r+1}, \hat{\boldsymbol{x}}^{t-r+1}) + D(\hat{\boldsymbol{x}}^{t-r+1}, \boldsymbol{x}^{t-r})) - \frac{1}{2}D(\hat{\boldsymbol{x}}^{t+1}, \boldsymbol{x}^t) - D(\boldsymbol{x}^t, \hat{\boldsymbol{x}}^t) \tag{A118}$$

$$= D(\boldsymbol{x}^*, \hat{\boldsymbol{x}}^t) + \frac{1}{2}\Big\{D(\hat{\boldsymbol{x}}^t, \boldsymbol{x}^{t-1}) + \sum_{r=1}^{n-1} \frac{n-r}{n}(D(\boldsymbol{x}^{t-r}, \hat{\boldsymbol{x}}^{t-r}) + D(\hat{\boldsymbol{x}}^{t-r}, \boldsymbol{x}^{t-r-1}))\Big\} - \frac{1}{2}D(\hat{\boldsymbol{x}}^{t+1}, \boldsymbol{x}^t) - \frac{1}{2}D(\boldsymbol{x}^t, \hat{\boldsymbol{x}}^t) \tag{A119}$$

$$= V^t(\boldsymbol{x}^*) - \frac{1}{2}D(\hat{\boldsymbol{x}}^{t+1}, \boldsymbol{x}^t) - \frac{1}{2}D(\boldsymbol{x}^t, \hat{\boldsymbol{x}}^t). \tag{A120}$$

In (a), we used

$$D(\boldsymbol{x}^*, \hat{\boldsymbol{x}}^{t+1}) + D(\hat{\boldsymbol{x}}^{t+1}, \boldsymbol{x}^t) = D(\boldsymbol{x}^*, \hat{\boldsymbol{x}}^t) - D(\boldsymbol{x}^t, \hat{\boldsymbol{x}}^t) + \big\langle \nabla h(\hat{\boldsymbol{x}}^{t+1}) - \nabla h(\boldsymbol{x}^t), \hat{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^t \big\rangle, \tag{A121}$$

which is obtained by summing up the following generalized law of cosines

$$D(\boldsymbol{x}^*, \hat{\boldsymbol{x}}^{t+1}) = D(\boldsymbol{x}^*, \hat{\boldsymbol{x}}^t) - D(\hat{\boldsymbol{x}}^{t+1}, \hat{\boldsymbol{x}}^t) + \big\langle \nabla h(\hat{\boldsymbol{x}}^{t+1}) - \nabla h(\hat{\boldsymbol{x}}^t), \hat{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^* \big\rangle, \tag{A122}$$

$$D(\hat{\boldsymbol{x}}^{t+1}, \boldsymbol{x}^t) = D(\hat{\boldsymbol{x}}^{t+1}, \hat{\boldsymbol{x}}^t) - D(\boldsymbol{x}^t, \hat{\boldsymbol{x}}^t) + \big\langle \nabla h(\boldsymbol{x}^t) - \nabla h(\hat{\boldsymbol{x}}^t), \boldsymbol{x}^t - \hat{\boldsymbol{x}}^{t+1} \big\rangle. \tag{A123}$$

and using the Nash equilibrium condition

$$\big\langle \nabla h(\hat{\boldsymbol{x}}^{t+1}) - \nabla h(\hat{\boldsymbol{x}}^t), \boldsymbol{x}^* - \boldsymbol{x}^t \big\rangle = \eta \big\langle \boldsymbol{u}^t, \boldsymbol{x}^* - \boldsymbol{x}^t \big\rangle = 0. \tag{A124}$$

In (b), for $\eta \leq 1/(8\sqrt{2}n^2L)$, we used

$$\begin{aligned}
&\left\langle \nabla h(\hat{\boldsymbol{x}}^{t+1}) - \nabla h(\boldsymbol{x}^t), \hat{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^t \right\rangle && \text{(A125)}\\
&\leq \|\nabla h(\hat{\boldsymbol{x}}^{t+1}) - \nabla h(\boldsymbol{x}^t)\|^2 && \text{(A126)}\\
&= \eta^2 \Big\| \sum_{r=1}^{n} \boldsymbol{u}^{t-r+1} - n\boldsymbol{u}^{t-n} \Big\|_2^2 && \text{(A127)}\\
&\leq \eta^2 L^2 \Big\| \sum_{r=1}^{n} \boldsymbol{x}^{t-r+1} - n\boldsymbol{x}^{t-n} \Big\|_2^2 && \text{(A128)}\\
&= \eta^2 L^2 \Big\| \sum_{r=1}^{n} \{r(\boldsymbol{x}^{t-r+1} - \hat{\boldsymbol{x}}^{t-r+1}) + r(\hat{\boldsymbol{x}}^{t-r+1} - \boldsymbol{x}^{t-r})\} \Big\|_2^2 && \text{(A129)}\\
&\leq 2n\eta^2 L^2 \sum_{r=1}^{n} r^2 (\|\boldsymbol{x}^{t-r+1} - \hat{\boldsymbol{x}}^{t-r+1}\|_2^2 + \|\hat{\boldsymbol{x}}^{t-r+1} - \boldsymbol{x}^{t-r}\|_2^2) && \text{(A130)}\\
&\leq 2n^3\eta^2 L^2 \sum_{r=1}^{n} (\|\boldsymbol{x}^{t-r+1} - \hat{\boldsymbol{x}}^{t-r+1}\|_2^2 + \|\hat{\boldsymbol{x}}^{t-r+1} - \boldsymbol{x}^{t-r}\|_2^2) && \text{(A131)}\\
&\leq 4n^3\eta^2 L^2 \sum_{r=1}^{n} (D(\boldsymbol{x}^{t-r+1}, \hat{\boldsymbol{x}}^{t-r+1}) + D(\hat{\boldsymbol{x}}^{t-r+1}, \boldsymbol{x}^{t-r})) && \text{(A132)}\\
&\leq \frac{1}{2n} \sum_{r=1}^{n} (D(\boldsymbol{x}^{t-r+1}, \hat{\boldsymbol{x}}^{t-r+1}) + D(\hat{\boldsymbol{x}}^{t-r+1}, \boldsymbol{x}^{t-r})). && \text{(A133)}
\end{aligned}$$

In the first inequality, we twice used

$$\begin{aligned}
\|\hat{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^t\|_2^2 &\leq \left\langle \nabla h(\hat{\boldsymbol{x}}^{t+1}) - \nabla h(\boldsymbol{x}^t), \hat{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^t \right\rangle && \text{(A134)}\\
&\leq \|\nabla h(\hat{\boldsymbol{x}}^{t+1}) - \nabla h(\boldsymbol{x}^t)\|_2 \|\hat{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^t\|_2. && \text{(A135)}
\end{aligned}$$

Here, we used the 1-strong convexity of h and the Cauchy-Schwarz inequality. □

E.3 Proof of Lem. A6

Proof. For all $\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}' \in \mathrm{int}(\mathcal{X})$ and all $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{R}^d$, we define

$$\boldsymbol{x} := \boldsymbol{P}(\hat{\boldsymbol{x}}, \boldsymbol{u}), \quad \boldsymbol{x}' := \boldsymbol{P}(\hat{\boldsymbol{x}}', \boldsymbol{u}'). \tag{A136}$$

We obtain

$$\begin{aligned}
\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2 &\overset{\mathrm{a}}{\leq} \left\langle \nabla h(\boldsymbol{x}) - \nabla h(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \right\rangle && \text{(A137)}\\
&\overset{\mathrm{b}}{=} \left\langle \nabla h(\hat{\boldsymbol{x}}) - \nabla h(\hat{\boldsymbol{x}}'), \boldsymbol{x} - \boldsymbol{x}' \right\rangle + \eta \left\langle \boldsymbol{u} - \boldsymbol{u}', \boldsymbol{x} - \boldsymbol{x}' \right\rangle && \text{(A138)}\\
&\overset{\mathrm{c}}{\leq} (\|\nabla h(\hat{\boldsymbol{x}}) - \nabla h(\hat{\boldsymbol{x}}')\|_2 + \eta\|\boldsymbol{u} - \boldsymbol{u}'\|_2)\|\boldsymbol{x} - \boldsymbol{x}'\|_2. && \text{(A139)}
\end{aligned}$$

In (a), we used 1-strong convexity of h. In (b), we summed up both the first-order optimality conditions for $\boldsymbol{x}$ and $\boldsymbol{x}'$, i.e.,

$$\left\langle \eta\boldsymbol{u} - \nabla h(\boldsymbol{x}) + \nabla h(\hat{\boldsymbol{x}}), \boldsymbol{x} - \boldsymbol{x}' \right\rangle = 0, \quad \left\langle \eta\boldsymbol{u}' - \nabla h(\boldsymbol{x}') + \nabla h(\hat{\boldsymbol{x}}'), \boldsymbol{x}' - \boldsymbol{x} \right\rangle = 0. \tag{A140}$$

In (c), we used the Cauchy-Schwarz inequality. In conclusion, we obtain

$$\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \leq \|\nabla h(\hat{\boldsymbol{x}}) - \nabla h(\hat{\boldsymbol{x}}')\|_2 + \eta\|\boldsymbol{u} - \boldsymbol{u}'\|_2. \tag{A141}$$

Here, ∇h is continuous on $\mathrm{int}(\mathcal{X})$. By the definition of a convex function of Legendre type, h is convex and differentiable. Moreover, $\mathrm{int}(\mathcal{X})$ is an open convex set, and $h(\hat{\boldsymbol{x}})$ is finite in $\hat{\boldsymbol{x}} \in \mathrm{int}(\mathcal{X})$. Thus, Cor. 25.5.1 in the study [44] is applicable, and h is continuously differentiable, meaning that ∇h is continuous on $\mathrm{int}(\mathcal{X})$. Thus, Eq. (A141) shows that $\boldsymbol{P}$ is continuous on $\mathrm{int}(\mathcal{X}) \times \mathbb{R}^{\mathrm{d}}$. □

F Analysis for Individual Regret

By using Eq. (FO3) for $(\tilde{\mathbf{f}}', \tilde{\mathbf{f}}) = (\tilde{\boldsymbol{x}}_i^t, \tilde{\boldsymbol{x}}_i^{t-1})$, we obtain the "stability" property as

$$\begin{aligned}
\|\boldsymbol{x}_i^t - \boldsymbol{x}_i^{t-1}\|_2 &\leq \eta \|\tilde{\boldsymbol{x}}_i^t - \tilde{\boldsymbol{x}}_i^{t-1}\|_2 && \text{(A142)}\\
&= \eta \|(m+2)\boldsymbol{u}^{t-m-1} - (m+1)\boldsymbol{u}^{t-m-2}\|_2 && \text{(A143)}\\
&\leq \eta \{(m+1)\|\boldsymbol{u}^{t-m-1} - \boldsymbol{u}^{t-m-2}\|_2 + \|\boldsymbol{u}^{t-m-1}\|_2\} && \text{(A144)}\\
&\leq (m+2)\eta L. && \text{(A145)}
\end{aligned}$$

This leads to the upper bound of individual regret as

$$\begin{aligned}
\mathrm{REG}_i(T) &\leq \frac{h_{\max}}{\eta} + \lambda^2 \eta \sum_{t=1}^{T} \|\boldsymbol{u}_i^t - \boldsymbol{u}_i^{t-1}\|_2^2 - \frac{1}{4\eta} \sum_{t=1}^{T} \|\boldsymbol{x}_i^t - \boldsymbol{x}_i^{t-1}\|_2^2 && \text{(A146)}\\
&\leq \frac{h_{\max}}{\eta} + \lambda^2 \eta \sum_{t=1}^{T} \|\boldsymbol{u}^t - \boldsymbol{u}^{t-1}\|_2^2 && \text{(A147)}\\
&\leq \frac{h_{\max}}{\eta} + \lambda^2 L^2 \eta \sum_{t=1}^{T} \|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|_2^2 && \text{(A148)}\\
&\leq \frac{h_{\max}}{\eta} + (m+2)^2 \lambda^2 N L^4 \eta^3 T. && \text{(A149)}
\end{aligned}$$

If we set $\eta = 1/(\sqrt{(m+2)\lambda} T^{1/4})$ there, we derive $\mathrm{REG}_i(T) \leq (h_{\max} + NL^4)\sqrt{(m+2)\lambda} T^{1/4} = O(m^{3/2} T^{1/4})$.

G Computational Environment

The codes are available at

`https://github.com/CyberAgentAILab/delayed_learning_games`

The simulations presented in this paper were conducted using the following computational environment.

- Operating System: macOS Monterey (version 12.4)
- Programming Language: Python 3.11.3
- Processor: Apple M1 Pro (10 cores)
- Memory: 32 GB

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We stated our contribution in the abstract and introduction (by itemizing).

   Guidelines:

   - The answer NA means that Abstract and Introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We stated the limitations in Conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All the proofs are in the main text or appendices.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We show the (hyper-)parameters to reproduce the experiments in their captions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is available at the URL in Appendix.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We show the (hyper-)parameters to reproduce the experiments in their captions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our experiments are deterministic and thus do not require statistical tests.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We clarify our computer resources in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper confirms the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The positive societal impacts are stated in Conclusion, while there are no negative ones.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets are attached in the supplement.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.