From Average-Iterate to Last-Iterate Convergence in Games: A Reduction and Its Applications*

Yang Cai
Yale University
yang.cai@yale.edu

Haipeng Luo
University of Southern California
haipengl@usc.edu

Chen-Yu Wei
University of Virginia
chenyu.wei@virginia.edu

Weiqiang Zheng
Yale University
weiqiang.zheng@yale.edu

Abstract

The convergence of online learning algorithms in games under self-play is a fundamental question in game theory and machine learning. Among various notions of convergence, last-iterate convergence is particularly desirable, as it reflects the actual decisions made by the learners and captures the day-to-day behavior of the learning dynamics. While many algorithms are known to converge in the average-iterate, achieving last-iterate convergence typically requires considerably more effort in both the design and the analysis of the algorithm. Somewhat surprisingly, we show in this paper that for a large family of games, there exists a simple black-box reduction that transforms the average iterates of an uncoupled learning dynamics into the last iterates of a new uncoupled learning dynamics, thus also providing a reduction from last-iterate convergence to average-iterate convergence. Our reduction applies to games where each player's utility is linear in both their own strategy and the joint strategy of all opponents. This family includes two-player bimatrix games and generalizations such as multi-player polymatrix games. By applying our reduction to the Optimistic Multiplicative Weights Update algorithm, we obtain new state-of-the-art last-iterate convergence rates for uncoupled learning dynamics in multi-player zero-sum polymatrix games: (1) an $O(\frac{\log d}{T})$ last-iterate convergence rate under gradient feedback, representing an exponential improvement in the dependence on the dimension d (i.e., the maximum number of actions available to either player); and (2) an $\tilde{O}(d^{\frac{1}{5}}T^{-\frac{1}{5}})$ last-iterate convergence rate under bandit feedback, improving upon the previous best rates of $\tilde{O}(\sqrt{d}T^{-\frac{1}{8}})$ and $\tilde{O}(\sqrt{d}T^{-\frac{1}{6}})$.

1 Introduction

The convergence of online learning algorithms in games under self-play is a fundamental question in both game theory and machine learning. Self-play methods for computing Nash equilibria have enabled the development of superhuman AI agents in competitive games such as Go [Sil+17], Poker [Bow+15; BS18; BS19], Stratego [Per+22], and Diplomacy [FAI+22]. More recently, self-play learning algorithms have also been applied to large language model (LLM) alignment with human feedback, which can be modeled as a two-player zero-sum game [Mun+23; Swa+24; Ye+24; Wu+24; Liu+24; Liu+25]. From a game-theoretic perspective, understanding the convergence behavior of self-play dynamics provides predictive insights into strategic multi-agent interactions and informs the design of more effective mechanisms.

*Authors are ordered alphabetically.

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

Among various notions of convergence, *last-iterate convergence* is particularly desirable, as it reflects the actual decisions made by the learners and captures the day-to-day behavior of the learning dynamics. Formally, we say an algorithm has a last-iterate convergence rate of $f(T)$ to Nash equilibria (where $f(T)$ is a decreasing function with $\lim_{T\to+\infty} f(T) = 0$) if the generated sequence of strategy profiles $\{x^t\}$ satisfies that, for any $T \geq 1$, the iterate x^T is an $f(T)$-approximate Nash equilibrium. This notion of *anytime* last-iterate convergence is stronger than those considered in some prior works, which do not provide the same guarantee for every T, but instead require knowing the time horizon in advance. See Section 1.1 for a detailed discussion.

Despite its importance, classical results on the convergence of self-play dynamics primarily concern *average-iterate* convergence. It is well known that when both players in a two-player zero-sum game employ no-regret online learning algorithms, the time-average of their strategies converges to a Nash equilibrium [FS99]. In contrast, a large class of online learning algorithms—including Mirror Descent (MD) and Follow-the-Regularized-Leader (FTRL)—fails to achieve last-iterate convergence in such settings. Worse yet, the sequence of actual decisions made by the learners can diverge, exhibiting cyclic or even chaotic behavior [MPP18; BP18; DP18].

Aside from last-iterate convergence, another desirable property of self-play learning dynamics with multiple self-interested agents is *uncoupledness* [DDK11]. In uncoupled learning dynamics, each player refines their strategies with minimal information about the game, while avoiding extensive inter-player coordination (e.g., shared randomness) and communication. Uncoupled learning dynamics are also studied under the names of independent or decentralized learning dynamics [OSZ21]. They are particularly appealing in large-scale games, where either the game is high-dimensional and full information is hard to obtain, or there are many players and coordination is difficult to enforce.

Due to their importance in large-scale games, designing and analyzing uncoupled learning dynamics with last-iterate convergence rates has received extensive attention recently due to its importance both in theory and practice. We will review this long line of work in Section 1.1, and only point out here that achieving last-iterate convergence typically requires considerably more effort in both the design and analysis of algorithms. On the design side, techniques such as *optimism* and *regularization* are often applied to ensure convergence of algorithms like OMD and FTRL. On the analysis side, establishing last-iterate convergence rates often demands problem-specific Lyapunov functions combined with tailored arguments. This stands in stark contrast to average-iterate convergence in two-player zero-sum games, which follows directly from the no-regret property of the algorithms.

Our contribution Somewhat surprisingly, we show in this paper that for a large family of games, there exists a simple black-box reduction, `A2L` (Algorithm 1), that transforms the average iterates of an uncoupled learning dynamic into the last iterates of another uncoupled learning dynamic, thereby providing a reduction from last-iterate convergence to average-iterate convergence. Our reduction applies to games where each player's utility is linear in both their own strategy and the joint strategy of all opponents. This family includes two-player bimatrix games and generalizations such as multi-player polymatrix games. Our reduction also works for games with nonlinear utilities but special structures such as the proportional response dynamics in Fisher markets (Appendix B).

Aside from its conceptual contribution, our reduction also yields concrete improvements for uncoupled learning in multi-player zero-sum polymatrix games (which include two-player zero-sum games as special cases) and leads to new state-of-the-art last-iterate convergence rates. Let d be the maximum number of actions available to each player. We apply our reduction to the Optimistic Multiplicative Weights Update (`OMWU`) algorithm [RS13; Syr+15], which is known to achieve an $O(\log d/T)$ average-iterate convergence rate in multi-player zero-sum polymatrix games under gradient feedback. As a result, our reduction yields uncoupled learning dynamics, `A2L-OMWU`, that enjoys an $O(\log d/T)$ last-iterate convergence rate under gradient feedback. This represents an exponential improvement in the dependence on d compared to the best previously known $O(\mathrm{poly}(d)/T)$ rate [CZ23b]. As an additional consequence, each player using `A2L-OMWU` incurs only $O(\log d \log T)$ dynamic regret.

We further extend our reduction to the more challenging setting where only *bandit feedback*, rather than full gradient feedback, is available. To this end, we design a new algorithm, `A2L-OMWU-Bandit` (Algorithm 2), which augments the reduction with a utility estimation procedure. We show that `A2L-OMWU-Bandit` achieves a $\widetilde{O}(T^{-1/5})$ last-iterate convergence rate with high probability. This result improves upon the previously best known rates of $\widetilde{O}(T^{-1/8})$ (high probability) and $\widetilde{O}(T^{-1/6})$ (in expectation) established in [Cai+23] for the special case of two-player zero-sum games.

1.1 Related Works

There is a vast literature on the regret guarantee and last-iterate convergence of learning algorithms in games. Here, we mainly review results applicable to two-player zero-sum games. For convergence under other conditions, such as strict equilibria and strong monotonicity, we refer readers to [GVM21; JLZ24; Ba+25] and the references therein.

Individual Regret Guarantee for Learning in Games While $\Theta(\sqrt{T})$ regret is fundamental for online learning against adversarial loss sequences, improved regret is possible for multi-agent learning in games. Starting from the pioneer work of [DDK11], a long line of works propose uncoupled learning dynamics with $O(1)$ regret in zero-sum games [RS13] and more generally variationally stable games [HAM21], and $O(\mathrm{poly}(\log T))$ (swap) regret even in general-sum normal-form games [DFG21; Ana+22a; Ana+22b; SPF25] and Markov games [Cai+24b; Mao+24]. However, guarantees for the stronger (worst-case) *dynamic regret* remain underexplored. It is worth noting that in the adversarial setting, achieving even sublinear dynamic regret is impossible. As a corollary of our $O(\log d \cdot T^{-1})$ last-iterate convergence rate, `A2L-OMWU` guarantees that each player suffers only $O(\log d \cdot \log T)$ dynamic regret, an exponential improvement on the dependence on d [CZ23b].

Last-Iterate Convergence with Gradient Feedback Two classic methods that exhibit last-iterate convergence are the Extra Gradient (EG) algorithm [Kor76] and the Optimistic Gradient (OG) algorithm [Pop80], which are optimistic variants of vanilla gradient descent. Both algorithms are known to converge asymptotically in the last iterate, but their convergence rates remained open until recently, highlighting the challenge of analyzing last-iterate convergence rates. [Wei+21] established problem-dependent linear convergence rates for OG. Tight $O(1/\sqrt{T})$ last-iterate convergence rates for both EG and OG were subsequently established by [COZ22; GTG22], matching the lower bounds from [Gol+20; GPD20].

Although the $O(1/\sqrt{T})$ rate is tight for EG and OG, these algorithms are not optimal among first-order methods. By leveraging anchoring-based acceleration techniques from the optimization literature [Hal67], accelerated versions of EG and OG have been proposed to achieve an $O(1/T)$ last-iterate convergence rate [Dia20; YR21; CZ23b; CZ23a; COZ24]. This matches the lower bounds for all first-order methods [OX21; YR21]. We note that all these results incur a $\mathrm{poly}(d)$ dependence. In contrast, our algorithm achieves exponentially better dependence on d.

In addition to optimism, regularization is another effective technique for achieving last-iterate convergence. By adding a regularization term, the original game becomes strongly monotone, enabling standard gradient-based algorithms to achieve linear last-iterate convergence—albeit on the modified game. Setting the regularization strength to $O(\varepsilon)$ yields an iteration complexity of $O(\log(1/\varepsilon)/\varepsilon)$ for computing an ε-approximate Nash equilibrium [CWC21; CWC24], corresponding to an $O(\log T/T)$ convergence rate. However, this approach has key limitations: it requires calibrating the regularization strength based on the total number of iterations T, and the resulting convergence guarantee applies only to the final iteration. It does not satisfy the stronger criterion of *anytime* last-iterate convergence. This distinction is crucial: without the anytime guarantee, a trivial workaround exists—one can run any no-regret algorithm for $T-1$ steps and output the average iterate at step T. Moreover, only anytime guarantees give individual dynamic regret guarantees.

While a diminishing regularization schedule does yield anytime last-iterate convergence, it leads to a slower convergence rate $\widetilde{O}(T^{-1/4})$ rate [PZO23]. With an additional assumption on the regularized Nash equilibrium, [ZDR22] obtained a $O(T^{-1/3})$ last-iterate convergence rate. One might wonder whether more aggressive schedules, such as the doubling trick, could improve this. However, it remains unclear whether such methods guarantee true anytime convergence.[2] In contrast, our algorithm achieves anytime last-iterate convergence *without* requiring prior knowledge of the time horizon, and it attains the optimal $O(1/T)$ rate, eliminating the extra logarithmic factor.

We also remark that there is a line of work on last-iterate convergence under *noisy gradient* feedback [Hsi+22]. [ASI22; Abe+23; Wu+25] design algorithms based on perturbation and show last-iterate convergence to a stationary point near a Nash equilibrium in both gradient and noisy

[2] [Liu+23] uses the doubling trick to establish a $\widetilde{O}(T^{-1})$ last-iterate convergence rate for the *output policy* of the algorithm (which generally differs from the *day-to-day policy* the players use to interact with each other). This is a weaker notion of last-iterate convergence than the one considered in this paper, which concerns the convergence of the day-to-day policy.

gradient feedback. [Abe+23] also show asymptotic convergence to an exact Nash equilibrium by adaptively adjusting the perturbation. For non-asymptotic convergence rates, [Abe+24] establish a $\tilde{O}(T^{-1/10})$ last-iterate convergence rate, which is improved to $\tilde{O}(T^{-1/7})$ by [Abe+25] recently.

Last-Iterate Convergence with Bandit Feedback Compared to gradient feedback, achieving last-iterate convergence under *bandit feedback* (i.e., payoff-based feedback) is significantly more challenging and remains less understood. [Cai+23] provides the first set of last-iterate convergence results for uncoupled learning dynamics in two-player zero-sum (Markov) games. They propose mirror-descent-based algorithms that achieve a high-probability convergence rate of $\widetilde{O}(T^{-1/8})$ and a rate of $\widetilde{O}(T^{-1/6})$ in expectation. Concurrently, [Che+23; Che+24] introduce best-response-based algorithms with an expected $\widetilde{O}(T^{-1/8})$ last-iterate convergence rate. A follow-up work [DWY24] presents extensions to general monotone games. Our work improves upon these results by achieving a high-probability $\widetilde{O}(T^{-1/5})$ rate, and our guarantees extend to the more general setting of multi-player zero-sum polymatrix games.

Last-Iterate Convergence of `OMWU` The Optimistic Multiplicative Weights Update (`OMWU`) algorithm [RS13; Syr+15] is a fundamental algorithm for learning in games. It achieves an $O(1/T)$ average-iterate convergence rate in two-player zero-sum games and guarantees $O(\log T)$ individual regret even in general-sum games [DFG21; Ana+22a; SPF25]. For last-iterate convergence in two-player zero-sum games, [DP19; HAM21] establish asymptotic convergence of `OMWU`, and [Wei+21] proves a problem-dependent linear convergence rate under the assumption of a unique Nash equilibrium. However, for worst-case uniform last-iterate convergence—without relying on problem-dependent constants—[Cai+24a] recently proved an $\Omega(1)$ lower bound, showing that `OMWU`'s last-iterate convergence can be arbitrarily slow.

2 Preliminaries

Notations We denote the d-dimensional probability simplex as $\Delta^d := \{x \in \mathbb{R}^d : 0 \le x_i \le 1, \sum_{i=1}^d x_i = 1\}$. The uniform distribution in Δ^d is denoted as $\mathrm{Uniform}(d)$.

2.1 Games

An n-player game $\mathcal{G} = ([n], \{\mathcal{X}_i\}, \{u_i\})$ consists of n players indexed by $[n] := \{1, 2, \ldots, n\}$. Each player i selects a strategy x_i from a closed convex set $\mathcal{X}_i \subseteq \mathbb{R}^{d_i}$. We refer to $d := \max_{i\in[n]} d_i$ as the dimensionality of the game. Given a strategy profile $x = (x_i, x_{-i})$, player i receives utility $u_i(x) \in [0, 1]$. A game is in normal-form if each player i has a finite number of d_i actions and given an action profile $a = (a_i, a_{-i})$, the utility for player i is $u_i(a)$. In normal-form games, the strategy set for each player u is the probability simplex $\mathcal{X}_i = \Delta^{d_i}$, and given a strategy profile x, the expected utility for player i is $\mathbb{E}_{\forall i\in[n], a_i\sim x_i}[u_i(a_1, \ldots, a_n)]$.

We focus on games with *linear utilities*.

Assumption 1. *A game $\mathcal{G}$ has* linear utilities *if for each player i,*

- *the utility function $u_i(x_i, x_{-i})$ is linear in $x_i \in \mathcal{X}_i$*
- *the utility function $u_i(x_i, x_{-i})$ is linear in $x_{-i} \in \times_{j\neq i}\mathcal{X}_j$.*

For such games, we define $u_i(\cdot, x_{-i}) \in \mathbb{R}^{d_i}$ to be the unique vector such that $u_i(x) = \langle x_i, u_i(\cdot, x_{-i})\rangle$. We remark that Assumption 1 is weaker than assuming $u_i(x)$ is linear in the whole strategy profile $x \in \times_i \mathcal{X}_i$. As an example, the bilinear function $u(x_1, x_2) = x_1^\top A x_2$ satisfies Assumption 1 since it is linear in x_1 and also linear in x_2, but it is not linear in (x_1, x_2). This structure includes several important game classes:

Example 1 (Two-Player Bimatrix Games)**.** *There are two players: the x-player chooses a mixed strategy $x \in \Delta^{d_1}$, and the y-player chooses $y \in \Delta^{d_2}$. Let $A, B \in \mathbb{R}^{d_1\times d_2}$ be the payoff matrices. The x-player's utility is $u_x(x, y) = x^\top Ay$, and the y-player's utility is $u_y(x, y) = x^\top By$. Two-player bimatrix games satisfy Assumption 1 due to the bilinear structure of the utilities.* A two-player zero-sum game *is a special case where $A + B = 0$.*

Example 2 (Multi-Player Polymatrix Games)**.** *Polymatrix games [Jan68; How72] generalize bimatrix games to multiple players by introducing networked interactions. The n players form the vertices of*

a graph $G = ([n], E)$, where each edge $(i, j) \in E$ represents a two-player bimatrix game between players i and j with payoff matrices $A_{i,j}$ and $A_{j,i}$. Each player i selects a strategy from Δ^{d_i} and plays the same strategy against all neighbors. Given a strategy profile $x \in \times_i \Delta^{d_i}$, player i's utility is defined as $\sum_{j:(i,j)\in E} x_i^\top A_{i,j} x_j$. A zero-sum polymatrix game *[CD11; Cai+16] is a special case where the total utility sums to zero, i.e., $\sum_{i=1}^n u_i(x) = 0$ for any strategy profile x.*

Nash Equilibrium and Total Gap An *ε-approximate Nash equilibrium* of $\mathcal{G}$ is a strategy profile $x \in \times_i \mathcal{X}_i$ such that no player i can deviate from x_i and improve her utility by more than $\varepsilon > 0$,

$$u_i(x_i, x_{-i}) \geq u_i(x_i', x_{-i}) - \varepsilon, \forall i \in [n], \forall x_i' \in \mathcal{X}_i.$$

When $\varepsilon = 0$, the strategy profile is a Nash equilibrium. We note that games with linear utilities have at least one Nash equilibrium [Nas50]. Given a strategy profile x, we use the total gap to evaluate its proximity to Nash equilibria. Specifically, we define $\text{TGAP}(x) := \sum_{i=1}^n (\max_{x_i' \in \mathcal{X}_i} u_i(x_i', x_{-i}) - u_i(x))$. We note that by definition, if $\text{TGAP}(x) \leq \varepsilon$, then x is an ε-approximate Nash equilibrium. In two-player zero-sum games, the total gap is also known as the duality gap.

2.2 Online Learning

Online Learning and Regret In an online learning problem, a learner repeatedly interacts with the environment. At each time $t \geq 1$, the learner chooses an action x^t from a closed convex set $\mathcal{X}$, while the environment simultaneously selects a linear utility function. The learner then receives a reward $u^t(x^t) := \langle u^t, x^t \rangle$ and observes some feedback. We focus on the *gradient feedback* setting, where the learner observes u^t; the more restricted *bandit feedback* setting is studied in Section 5.

The goal of the learner is to minimize the *(external) regret*, the difference between the cumulative utility and the utility of the best fixed action in hindsight, defined as $\text{Reg}(T) := \max_{x \in \mathcal{X}} \sum_{t=1}^T u^t(x) - \sum_{t=1}^T u^t(x^t)$. An online learning algorithm is *no-regret* if its regret is sublinear in T, that is, $\text{Reg}(T) = o(T)$. We note that classic results show that $\text{Reg}(T) = O(\sqrt{T})$ can be achieved and cannot be improved when utilities are chosen adversarially.

A stronger notion of regret is the (worst-case) *dynamic regret* [Zin03], which competes with the utility achieved by the best actions in each iteration, defined as $\text{DReg}(T) := \sum_{t=1}^T \max_{x \in \mathcal{X}} u^t(x) - \sum_{t=1}^T u^t(x^t)$. We remark that when utilities are chosen adversarially, it is impossible to achieve sublinear dynamic regret, that is, $\text{DReg}(T) = \Omega(T)$.

Convergence of Learning Dynamics For multi-agent learning dynamics in games, each player uses an online learning algorithm to repeatedly interact with other players. At each time t, player i chooses strategy x_i^t, gets utility $u_i(x_i^t, x_{-i}^t)$, and receives the utility vector $u_i^t := u_i(\cdot, x_{-i}^t)$ as feedback. For any $T \geq 1$, we denote the average iterate as $\overline{x}^T := (\overline{x}_1^T, \ldots, \overline{x}_n^T)$ where $\overline{x}_i^T = \frac{1}{T} \sum_{t=1}^T x_i^t$, and each player i's individual regret as $\text{Reg}_i(T)$. We show that the average iterate $\overline{x}^T = \frac{1}{T} \sum_{t=1}^T x^t$ has a total gap bounded by $\frac{1}{T} \sum_{i=1}^n \text{Reg}_i(T)$ for zero-sum polymatrix games (a generalization of the classic result of [FS99] for zero-sum bimatrix games). Thus, as long as each player has sublinear regret, average-iterate convergence to Nash equilibria is guaranteed. The proof is in Appendix A.

Lemma 1 (Average-Iterate Convergence by Bounding Regret)**.** *Let $\{x^t\}$ be the iterates of an online learning dynamics in a zero-sum polymatrix game. Define $\overline{x}^T = \frac{1}{T} \sum_{t=1}^T x^t$ to be the average iterate for all $T \geq 1$. Then the total gap of the average iterate $\overline{x}^T$ for any $T \geq 1$ is*

$$\text{TGAP}(\overline{x}^T) := \max_{x \in \times \Delta^{d_i}} \sum_{i=1}^n u_i(x_i, \overline{x}_{-i}^T) - u_i(\overline{x}^T) = \frac{1}{T} \sum_{i=1}^n \text{Reg}_i(T).$$

However, the convergence of the average-iterate sequence $\{\overline{x}^t\}$ does not imply the convergence of the last-iterate $\{x^t\}$, which is the real strategy played by the agents. Even worse, many online learning algorithms *diverge* and exhibit *cycling* or *chaotic behavior* even in simple two-player zero-sum games [MPP18; BP18; DP18].

Optimistic Multiplicative Weights Update (OMWU) The OMWU algorithm [RS13; Syr+15] is an optimistic variant of the classic Multiplicative Weights Update (MWU) algorithm [AHK12] with

the only modification that the most recent utility observation is counted twice. In each time $t \geq 1$, OMWU with steps size $\eta > 0$ chooses the strategy $x^t \in \Delta^d$ to be

$$x^t = \operatorname*{argmax}_{x \in \Delta^d} \left\{ \left\langle x, \sum_{k<t} u^k + u^{t-1} \right\rangle - \frac{1}{\eta}\phi(x) \right\},$$

where we define $u^0 := 0$ and ϕ is the negative entropy regularizer. The OMWU updates also admits a closed-form expression:

$$x^t[i] = \frac{\exp\left(\eta(\sum_{k<t} u^k[i] + u^{t-1}[i])\right)}{\sum_{j=1}^d \exp\left(\eta(\sum_{k<t} u^k[j] + u^{t-1}[j])\right)}, \forall i \in [d].$$

OMWU has been extensively studied in the online learning and game theory literature [RS13; Syr+15; Das+18; CP20; Wei+21; DFG21; Cai+24a; SPF25], with state-of-the-art individual regret guarantees in zero-sum games and general-sum games. We will use the *regret bounded by variation in utility* (RVU) property of OMWU throughout the paper.

Theorem 1 (Proposition 7 of [Syr+15])**.** *The regret of OMWU for a sequence of utilities $\{u^t\}$ satisfies*

$$\mathrm{Reg}(T) = \max_{x \in \Delta^d} \sum_{t=1}^T \langle u^t, x - x^t \rangle \leq \frac{\log d}{\eta} + \eta \sum_{t=1}^T \left\|u^t - u^{t-1}\right\|_\infty^2 - \frac{1}{4\eta} \sum_{t=1}^T \left\|x^t - x^{t-1}\right\|_1^2.$$

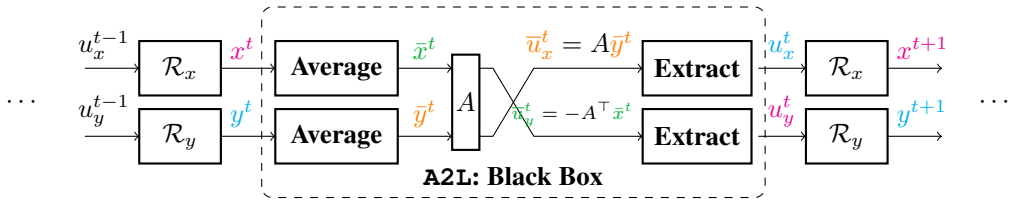
Lemma 2 (Adapted from Theorem 4 of [Syr+15])**.** *In a n-player normal-form game, let x and x' be two strategy profiles. Then $\sum_{i=1}^n \|u_i(\cdot, x_{-i}) - u_i(\cdot, x'_{-i})\|_\infty^2 \leq (n-1)^2 \sum_{i=1}^n \|x_i - x'_i\|_1^2$.*

3 A Reduction From Last-Iterate to Average-Iterate

In this section, we introduce a black-box reduction that transforms any online learning algorithm into a new one such that, when employed in a game with linear utilities (Assumption 1), the produced iterates exactly match the averaged iterates produced by the original algorithm.

Specifically, consider an n-player game $\mathcal{G}$, where each player i employs an online learning algorithm $\mathcal{R}_i$ and generate iterates $\{\hat{x}^t\}$. We propose a reduction, A2L (Algorithm 1), and let each player i employs $\texttt{A2L}(\mathcal{R}_i)$, generating a new sequence $\{\overline{x}^t\}$. We show that for any $t \geq 1$, $\overline{x}^t = \frac{1}{k}\sum_{k=1}^t \hat{x}^k$; that is, the last iterate of $\{\overline{x}^t\}$ equals the running average of $\{\hat{x}^t\}$.

The intuition of the reduction is as follows. Given any algorithm $\mathcal{R}$, $\texttt{A2L}(\mathcal{R})$ runs $\mathcal{R}$ as a subroutine and gets a strategy x^t in every iteration t. We note that the sequence $\{x^t\}$ is produced internally in the reduction and not played by the players. $\texttt{A2L}(\mathcal{R})$ plays the averaged strategy $\overline{x}^t = \frac{1}{t}\sum_{k=1}^t x^k$ and gets the gradient feedback $\overline{u}^t$. To show that the last iterate of $\{\overline{x}^t\}$ equals the running average of $\{\hat{x}^t\}$, it suffices to prove that $\hat{x}^t = x^t$. The next step is the key observation: when all players employ this reduction and play the running average, they can recover/extract the unseen utility vector $u^t = u_i(\cdot, x^t)$ under strategy x^t (which is not played) by the identity $u^t = t\overline{u}^t - (t-1)\overline{u}^{t-1}$, which holds due to the linearity of the utility (Assumption 1). Then $\texttt{A2L}(\mathcal{R})$ forwards this recovered utility to $\mathcal{R}$. Since $\mathcal{R}$ receives the same utilities when played alone or as a subroutine in $\texttt{A2L}(\mathcal{R})$, we can conclude that the iterates $\{\hat{x}^t\}$ equal the internal iterates of $\{x^t\}$. Thus $\overline{x}^t = \frac{1}{t}\sum_{k=1}^t x^k = \frac{1}{t}\sum_{k=1}^t \hat{x}^k$. A formal proof using induction is provided later in this section. We also provide an illustration of the A2L reduction for a two-player zero-sum game $\max_x \min_y xAy$ (Example 1) below, where we use the subscripts x and y to denote the corresponding quantities for the two players.



A key feature of this reduction is that it preserves uncoupledness: each player only observes their own utility, without access to other players' strategies, or any form of communication or shared randomness.

Algorithm 1: A2L: Black-Box Reduction from Average to Last-Iterate

Input: An online learning algorithm $\mathcal{R}$

1 **for** $t = 1, 2, \ldots,$ **do**

2 Get x^t from $\mathcal{R}$

3 Play $\bar{x}^t = \frac{1}{t}\sum_{k=1}^{t} x^k$ and receive $\overline{u}^t$ ▷ Play the running average iterates

4 Compute $u^t = t \cdot \overline{u}^t - \sum_{k=1}^{t-1} u^k$ ▷ Recover the unseen utility

5 Forward utility u^t to $\mathcal{R}$

Theorem 2. *Consider any n-player game $\mathcal{G}$ that satisfies Assumption 1). Let $\{\hat{x}_i^t : i \in [n], t \in [T]\}$ be the iterates generated by each player i running $\mathcal{R}_i$, and let $\{\bar{x}_i^t : i \in [n], t \in [T]\}$ be the iterates generated by running A2L $(\mathcal{R}_i)$. Then for all $i \in [n]$ and $t \in [T]$, we have $\bar{x}_i^t = \frac{1}{t}\sum_{k=1}^{t} \hat{x}_i^k$.*

Proof of Theorem 2. Note that A2L $(\mathcal{R}_i)$ uses $\mathcal{R}_i$ as a subroutine. Let $\{x_i^t\}$ denote the strategies produced by $\mathcal{R}_i$ during the execution of A2L $(\mathcal{R}_i)$. By Line 3, we have $\bar{x}_i^t = \frac{1}{t}\sum_{k=1}^{t} x_i^k, \forall i \in [n], t \in [T]$. Hence, it suffices to show that $x_i^t = \hat{x}_i^t$ for all $i \in [n], t \in [T]$.

We prove this by induction on t. Let the induction hypothesis be: for all $k \le t$ and all $i \in [n]$, we have $x_i^k = \hat{x}_i^k$ and $u_i^k = u_i(\cdot, x_{-i}^k)$.

Base Case: $t = 1$**.** Initialization ensures $x_i^1 = \hat{x}_i^1$ for all $i \in [n]$. By Lines 3 and 4, we also have $u_i^1 = \overline{u}_i^1 = u_i(\cdot, \overline{x}_{-i}^1) = u_i(\cdot, x_{-i}^1) = u_i(\cdot, \hat{x}_{-i}^1)$.

Induction Step: Suppose the hypothesis holds for t. Then for all $k \le t$ and $i \in [n]$, we have $u_i^k = u_i(\cdot, x_{-i}^k) = u_i(\cdot, \hat{x}_{-i}^k)$. Since $\mathcal{R}_i$'s output at time $t+1$ depends only on the sequence $\{u_i^k\}_{1 \le k \le t}$, we get $x_i^{t+1} = \hat{x}_i^{t+1}$ for all i.

At time $t+1$, each player i in A2L $(\mathcal{R}_i)$ plays $\bar{x}_i^{t+1} = \frac{1}{t+1}\sum_{k=1}^{t+1} x_i^k$. Then each player i receives the utility vector

$$\overline{u}_i^{t+1} = u_i\left(\cdot, \frac{1}{t+1}\sum_{k=1}^{t+1} x_{-i}^k\right) = \frac{1}{t+1}\sum_{k=1}^{t+1} u_i(\cdot, x_{-i}^k), \quad \text{(by linearity of } u_i \text{ (Assumption 1))}.$$

Then, $u_i^{t+1} = (t+1)\cdot\overline{u}_i^{t+1} - \sum_{k=1}^{t} u_i^k = \sum_{k=1}^{t+1} u_i(\cdot, x_{-i}^k) - \sum_{k=1}^{t} u_i(\cdot, x_{-i}^k) = u_i(\cdot, x_{-i}^{t+1})$. Thus, the induction hypothesis holds for $t+1$.

By induction, $x_i^t = \hat{x}_i^t$ for all i and t, and so $\bar{x}_i^t = \frac{1}{t}\sum_{k=1}^{t} x_i^k = \frac{1}{t}\sum_{k=1}^{t} \hat{x}_i^k$. □

Discussion on Limitations We discuss some limitations of our reduction compared to existing algorithms with last-iterate convergence guarantees, such as Extra Gradient (EG), Optimistic Gradient (OG), and their variants. Our reduction relies on the linear utility assumption (Assumption 1), which includes important classes of games such as multi-player polymatrix games, but does not extend to settings with concave utilities—such as convex-concave min-max optimization—where EG and OG are known to achieve last-iterate convergence [COZ22]. Moreover, since our approach reduces last-iterate convergence to average-iterate convergence, it inherently requires computing the running average of iterates. In contrast, algorithms like EG and OG do not involve such averaging procedures. Nonetheless, our reduction offers a simple and broadly applicable method for achieving last-iterate convergence in games with linear utilities, and it yields state-of-the-art convergence rates under both gradient and bandit feedback settings (see Sections 4 and 5).

Extension to Games with Nonlinear Utilities The core idea of A2L reduction lies in extracting the unseen feedback of u^t from the observed feedback $\overline{u}^t$, for which Assumption 1 is sufficient but not necessary. For example, in Appendix B, we show that A2L is also applicable to Proportional Response Dynamics (PRD) in Fisher markets where Assumption 1 does not hold. We obtain $O(1/T)$ last-iterate convergence rate for PRD in markets with Gross Substitutes utilities, improving [CCT25] which only shows an average-iterate convergence. We defer a detailed background on Fisher market and discussion on why A2L works to Appendix B. We believe that there are other similar examples where A2L applies even without Assumption 1.

Comparision with [Cut19] Our work is related to [Cut19], which proposes a reduction called *anytime online-to-batch (AO2B)* and achieves last-iterate convergence in offline convex optimization. While our A2L reduction and the AO2B reduction share the idea of playing the average of previous iterates, *they differ significantly in several key aspects*:

- **Problem setting:** AO2B is designed for static optimization, where the loss function remains fixed across rounds. A2L, on the other hand, applies to a dynamic multi-agent game setting, where each player's loss function changes over time due to the changing actions of other players.
- **Objective:** AO2B aims to solve a single offline optimization problem with improved convergence rates and does not guarantee equivalence between the last iterate of the new learning algorithm and the average iterate of the original algorithm (since the gradient feedbacks are different). In contrast, A2L aims to establish the equivalence between the last iterate of the new dynamics and the averaged iterate of the original dynamics.
- **Use of gradient:** In AO2B, the learner updates directly using the gradient at the average iterate $\overline{x}^t = \frac{1}{t}\sum_{k=1}^{t} x^k$. In contrast, A2L requires post-processing the gradient feedback to recover the true gradient at the original iterate x^t, as shown in Line 4 of Algorithm 1.

Weighted Version of A2L The A2L reduction naturally extends to the weighted averaging setting, provided all players agree on the weights $\{\alpha_t\}$ in advance and incorporate them into the reduction. More specifically, in each iteration t, each player plays the weighted average of their past iterates $\overline{x}_i^t = \frac{1}{\sum_{k=1}^{t}\alpha_k}\sum_{k=1}^{t}\alpha_k x_i^k$. Due to the linearity of utilities, the gradient feedback can still be used to recover the gradient corresponding to the original iterate x^k, and the A2L reduction in Algorithm 1 is the special case of $\{\alpha_t = 1\}$. As a result, the last iterate of the new dynamics corresponds to the weighted average in the original dynamics. The correctness follows the same steps as in Theorem 2. This weighted version of A2L provides additional flexibility and can be beneficial in practice. For example, for regret matching [Tam14], an algorithm widely used for solving large-scale games like poker [BS18; BS19], linear averaging (i.e., $\alpha_t = t$) often empirically outperforms uniform averaging, despite having the same theoretical rate.

4 Learning in Zero-Sum Games with Gradient Feedback

In this section, we show that our reduction gives uncoupled learning dynamics with improved last-iterate convergence rates in two-player zero-sum games and zero-sum polymatrix games.

We apply our reduction to the Optimistic Multiplicative Weights Update (OMWU) algorithm, which has been extensively studied in the literature [RS13; Syr+15; Das+18; CP20; Wei+21; DFG21; Cai+24a; SPF25]. For d-dimensional two-player zero-sum games, while a recent result [Cai+24a] shows that OMWU's last-iterate convergence can be arbitrarily slow, it is well-known that OMWU achieves an $O(\frac{\log d}{T})$ average-iterate convergence rate [RS13]. Now let us denote the algorithm after A2L reduction as A2L-OMWU. By Theorem 2 and Lemma 1, we get $O(\log d \cdot T^{-1})$ last-iterate convergence rate result of A2L-OMWU (Theorem 3). As a corollary of our fast last-iterate convergence rates, A2L-OMWU also guarantees that each player's individual dynamic regret is only $O(\log d \log T)$. Missing proofs in this section are in Appendix C.

Theorem 3. *Let $\{\overline{x}^t\}$ be the iterates of A2L-OMWU learning dynamics in an n-player zero-sum polymatrix game with step size $\eta \le \frac{1}{2(n-1)}$. Then for any $T \ge 1$, $\overline{x}^T$ is an $(\frac{\sum_{i=1}^{n}\log d_i}{\eta T})$-approximate Nash equilibrium.*

Corollary 1. *In the same setup as Theorem 3, for any player $i \in [n]$, $\mathrm{DReg}_i^T = O(\frac{\sum_{i=1}^{n}\log d_i}{\eta}\log T)$.*

To our knowledge, our results give the fastest last-iterate convergence rates for learning in zero-sum games. Compared to the accelerated optimistic gradient (AOG) algorithm [CZ23b] which has $O(\frac{\mathrm{poly}(d)}{T})$ last-iterate convergence rates, our results achieves the same optimal $\frac{1}{T}$ dependence while improve exponentially on the dependence on the dimension d. Compared to the entropy regularized extragradient algorithm [CWC21; CWC24], which achieves $O(\frac{\log d \log T}{T})$ last-iterate convergence rate, our results do not require regularization and shave the $\log T$ factor. Moreover, our results and those of [CZ23b] are stronger *anytime convergence results* that holds for every iterate $T \ge 1$, while results in [CWC21; CWC24] hold only for the final iterations and require knowing the total number

of iterations T in advance for tuning the regularization parameter (as mentioned in Section 1.1, a trivial solution exists for such weaker guarantees), thus with no dynamic regret guarantee.

Robustness against Adversaries In the case where other players are adversarial and may not follow the A2L-OMWU algorithm, a simple modification can ensure sublinear regret for the player. Specifically, the player can track their cumulative utilities and regret up to time T, and if the regret exceeds $\Omega(\log T)$—which cannot happen if all players follow A2L-OMWU —they can switch to a standard no-regret algorithm that guarantees $O(\sqrt{T})$ regret in the worst case.

5 Learning in Zero-Sum Games with Bandit Feedback

In this section, we show the application of our reduction for uncoupled learning dynamics in zero-sum games with bandit feedback. Unlike the gradient feedback setting, where each player observes the utility of every action, the bandit feedback setting is more realistic and much harder. Specifically, we consider the following uncoupled interaction protocol: in each iteration t

- Each player i maintains a mixed strategy $x_i^t \in \Delta^{d_i}$ and plays an action $a_i^t \sim x_i^t$;
- Each player i then receives a number $u_i(a^t)$ as feedback.

Importantly, a player only observes the utility of one action but not the whole utility vector. Also, a player does not observe the actions/strategies of other players.

Algorithm Design Ideally, we would like to have a bandit learning dynamics whose last iterates are the average iterates of another bandit learning dynamics. This would give a $O(T^{-\frac{1}{2}})$ last-iterate convergence rate since bandit algorithms have $O(\sqrt{T})$ regret. However, the lack of utility vector feedback prevents us from directly applying A2L reduction since we can no longer use the linearity to recover the original utility when other players use the averaged strategies. To overcome this challenge, we design A2L-OMWU-Bandit (Algorithm 2), where in addition to A2L, we add a new component for estimating the utility vector from bandit feedback. With a sufficiently accurate estimation of the utility vector, we can then use A2L as in the gradient feedback setting.

Algorithm 2: A2L-OMWU-Bandit: OMWU with bandit feedback and A2L reduction (for player i)

1 **Parameters:** Step size $\eta > 0$, $\{B_t \geq t^4\}$, $\{\varepsilon_t = t^{-1}\}$
2 **Initialization:** $x_i^1 \leftarrow \text{Uniform}(d_i))$; $\widehat{U}_i^0 = 0$.
3 **for** $t = 1, 2, \ldots,$ **do**
4 $\quad \overline{x}_i^t \leftarrow \frac{1}{t}\sum_{k=1}^t x_i^k$
5 $\quad \overline{x}_{i,\varepsilon}^t \leftarrow (1-\varepsilon_t)\overline{x}_i^t + \varepsilon_t \text{Uniform}(d_i)$
6 $\quad$ Sample actions from $\overline{x}_{i,\varepsilon}^t$ for B_t rounds and compute $\widehat{U}_{i,\varepsilon}^t$ as an estimate of $\overline{u}_{i,\varepsilon}^t$ based on (1)
7 $\quad$ Computed estimated utility $\hat{u}_{i,\varepsilon}^t \leftarrow t \cdot \widehat{U}_{i,\varepsilon}^t - (t-1)\cdot \widehat{U}_{i,\varepsilon}^{t-1}$
8 $\quad$ Update using OMWU: $x_i^{t+1} \leftarrow \text{argmax}_{x\in\Delta^{d_i}} \left\{\langle x, \sum_{k\leq t} \hat{u}_{i,\varepsilon}^k + \hat{u}_{i,\varepsilon}^t\rangle - \frac{1}{\eta}\phi(x)\right\}$.

Estimation To estimate the utility vector when each player only observes the reward after taking an action, we let the each player use their current strategy to interact with each other for a sequence of B_t rounds and then use the bandit feedback to calculate an estimated utility (Line 6). Let us denote the sampled actions over the B_t rounds as $\{a^1, \ldots, a^{B_t}\}$ and the corresponding bandit feedback as $\{r^1, \ldots, r^{B_T}\}$. We use the following simple estimator for the utility vector:

$$\widehat{U}_{i,\varepsilon}^t[a] = \frac{\sum_{k=1}^{B_t} r^k \cdot \mathbb{I}[a^k = a]}{\sum_{k=1}^{B_t} \mathbb{I}[a^k = a]}, \forall a. \tag{1}$$

To encourage exploration, we also mix the strategy with ε_t amount of uniform distribution (Line 5) to ensure that every action receives sufficient samples in each epoch.

Last-Iterate Convergence Rate In Algorithm 2, each iterate $\{\overline{x}_\varepsilon^t\}$ is repeated B_t times within epoch t. We will show that this sequence $\{\overline{x}_\varepsilon^t\}$ enjoys a $\widetilde{O}(t^{-1})$ last-iterate convergence rate. Let $\{\overline{y}^k\}$

denote the actual iterates executed by the players in every round k, which remains the same in each epoch. When $B_t = d \cdot t^4$, we get the following $\tilde{O}(d^{\frac{1}{5}} k^{-\frac{1}{5}})$ last-iterate rate of $\{\overline{y}^k\}$. If the maximal number of actions d is unknown to the players in a fully uncoupled setting, we can choose $B_t = t^4$ and get an $\tilde{O}(dk^{-\frac{1}{5}})$ convergence rate.

Theorem 4. *Consider an n-player zero-sum polymatrix game and denote $d := \max_{i\in[n]} d_i$. Let $\{\overline{y}^k\}$ be the iterates generated by all players running* `A2L-OMWU-Bandit` *with $\eta \le \frac{1}{6n}$, then for any $\delta \in (0,1)$, with probability at least $1 - O(nd^3\delta)$, for any $k \ge 1$, $\overline{y}^k$ is an β_k-approximate Nash equilibrium with*

$$\beta_k = O\left(\frac{n\log^2(\frac{dk}{\delta})}{\eta} \cdot d^{\frac{1}{5}} k^{-\frac{1}{5}}\right) \text{ when } B_t = d\cdot t^4 \quad \text{or} \quad \beta_k = O\left(\frac{n\log^2(\frac{dk}{\delta})}{\eta} \cdot dk^{-\frac{1}{5}}\right) \text{ when } B_t = t^4.$$

To our knowledge, $\tilde{O}(T^{-\frac{1}{5}})$ is the fastest last-iterate convergence rate of uncoupled learning dynamics in zero-sum polymatrix games, improving the previous $\tilde{O}(T^{-\frac{1}{8}})$ (high-probability) and $\tilde{O}(T^{-\frac{1}{6}})$ (expectation) rates for two-player zero-sum games [Cai+23].

Technical highlight The main challenge in analyzing the last-iterate convergence rate of Algorithm 2 lies in handling the estimation error. Let $\Delta_i^t = \widehat{U}_{i,\varepsilon}^t - \overline{u}_{i,\varepsilon}^t$ denote the estimation error from Line 6. A naive analysis leads to a summation of *first-order* error terms, $O(\sum_{t=1}^T \|t\Delta_i^t\|)$, which results in a suboptimal $\tilde{O}(T^{-1/7})$ rate. To improve this, we introduce a refined analysis that carefully balances the error terms with the negative terms in the RVU inequality (Theorem 1), resulting in a summation of *second-order* error terms, $O(\sum_{t=1}^T \|t\Delta_i^t\|^2)$ (see Lemma 3). This sharper analysis ultimately yields an improved $\tilde{O}(T^{-1/5})$ convergence rate. The full proof is provided in Appendix D.

Discussion Our results also reveal an intriguing distinction between the gradient and bandit feedback settings. In the gradient feedback case, while our reduction improves the dependence on the dimension d, it achieves the same $O(T^{-1})$ convergence rate as existing algorithms such as the accelerated optimistic gradient (AOG) algorithm [CZ23b]. This might suggest that applying a similar utility estimation procedure to AOG would also yield an $\tilde{O}(T^{-1/5})$ rate under bandit feedback. However, the analyses of these two types of algorithms diverge significantly in the presence of estimation error:

For AOG, establishing last-iterate convergence requires proving the (approximate) monotonicity of a carefully designed potential function at *every iteration*. This sensitivity to estimation error leads to error propagation and ultimately a slower rate: we did not manage to get a rate faster than $\tilde{O}(T^{-1/8})$ [Cai+23]. In contrast, our reduction—by transforming last-iterate convergence into average-iterate convergence—only requires analysis on the regret, which is simpler, less sensitive, and gives the improved $\tilde{O}(T^{-1/5})$ rate. This suggests an advantage of our `A2L` reduction over existing algorithms in the bandit feedback setting. Nevertheless, it remains an interesting open question whether one can design uncoupled learning dynamics with fast convergence rates in the bandit feedback setting using existing algorithms like AOG without relying on the `A2L` reduction.

Robustness against Adversaries When facing possibly adversarial opponents who may not follow Algorithm 2, the player could still guarantee sublinear regret by running Algorithm 2 with an additional regret estimation procedure. Whenever the player detects that her regret is $\Omega(T^{\frac{4}{5}})$—which can not happen when all the players employ Algorithm 2—she can switch to a standard bandit algorithm with optimal regret. This modification guarantees $\widetilde{O}(T^{\frac{4}{5}})$ regret in the worst-case. We present the regret estimation procedure and the proof in Appendix E.

6 Conclusion

In this paper, we give a simple black-box reduction, `A2L`, that transforms the average iterates of an uncoupled learning dynamics to the last iterates of a new uncoupled learning dynamics. By `A2L` reduction, we present improved last-iterate convergence rates of uncoupled learning dynamics in zero-sum polymatrix games: (1) $O(\log d \cdot T^{-1})$ rate under gradient feedback, and (2) $\tilde{O}(d^{\frac{1}{5}} T^{-\frac{1}{5}})$ rate under bandit feedback. It is an interesting future direction to explore further applications of the `A2L` reduction for learning in games. Other directions include designing uncoupled learning dynamics for more general games like Markov games and time-varying games with faster last-iterate convergence rates under bandit feedback.

Acknowledgement

We thank the anonymous NeurIPS reviewers for constructive comments. We also thank Arnab Maiti and Nathan Monette for comments on typos and minor errors in the previous version of the paper, which helped us improve the paper. YC is supported by the NSF Awards CCF-1942583 (CAREER) and CCF-2342642. HL is supported by NSF award IIS-1943607. WZ is supported by the NSF Awards CCF-1942583 (CAREER), CCF-2342642, and a Research Fellowship from the Center for Algorithms, Data, and Market Design at Yale (CADMY).

References

[Abe+23] Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, Kentaro Toyoshima, and Atsushi Iwasaki. "Last-Iterate Convergence with Full and Noisy Feedback in Two-Player Zero-Sum Games". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 7999–8028.

[Abe+24] Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. "Adaptively Perturbed Mirror Descent for Learning in Games". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 31–80.

[Abe+25] Kenshi Abe, Mitsuki Sakamoto, Kaito Ariu, and Atsushi Iwasaki. "Boosting Perturbed Gradient Ascent for Last-Iterate Convergence in Games". In: *The Thirteenth International Conference on Learning Representations*. 2025.

[AHK12] Sanjeev Arora, Elad Hazan, and Satyen Kale. "The multiplicative weights update method: a meta-algorithm and applications". In: *Theory of computing* 8.1 (2012), pp. 121–164.

[Ana+22a] Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. "Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games". In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*. 2022.

[Ana+22b] Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Tuomas Sandholm. "Uncoupled Learning Dynamics with $O(\log T)$ Swap Regret in Multiplayer Games". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[ASI22] Kenshi Abe, Mitsuki Sakamoto, and Atsushi Iwasaki. "Mutation-driven follow the regularized leader for last-iterate convergence in zero-sum games". In: *Uncertainty in Artificial Intelligence*. PMLR. 2022, pp. 1–10.

[Ba+25] Wenjia Ba, Tianyi Lin, Jiawei Zhang, and Zhengyuan Zhou. "Doubly optimal no-regret online learning in strongly monotone games with bandit feedback". In: *Operations Research* (2025).

[BDX11] Benjamin Birnbaum, Nikhil R Devanur, and Lin Xiao. "Distributed algorithms via gradient descent for Fisher markets". In: *Proceedings of the 12th ACM conference on Electronic commerce*. 2011, pp. 127–136.

[Bow+15] Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. "Heads-up limit hold'em poker is solved". In: *Science* 347.6218 (2015), pp. 145–149.

[BP18] James P Bailey and Georgios Piliouras. "Multiplicative weights update in zero-sum games". In: *Proceedings of the 2018 ACM Conference on Economics and Computation*. 2018, pp. 321–338.

[BS18] Noam Brown and Tuomas Sandholm. "Superhuman AI for heads-up no-limit poker: Libratus beats top professionals". In: *Science* 359.6374 (2018), pp. 418–424.

[BS19] Noam Brown and Tuomas Sandholm. "Superhuman AI for multiplayer poker". In: *Science* 365.6456 (2019), pp. 885–890.

[Cai+16] Yang Cai, Ozan Candogan, Constantinos Daskalakis, and Christos Papadimitriou. "Zero-sum polymatrix games: A generalization of minmax". In: *Mathematics of Operations Research* 41.2 (2016), pp. 648–655.

[Cai+23] Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. "Uncoupled and convergent learning in two-player zero-sum markov games with bandit feedback". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 36364–36406.

[Cai+24a] Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng. "Fast Last-Iterate Convergence of Learning in Games Requires Forgetful Algorithms". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.

[Cai+24b] Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. "Near-optimal policy optimization for correlated equilibrium in general-sum markov games". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 3889–3897.

[CCT25] Yun Kuen Cheung, Richard Cole, and Yixin Tao. "Proportional Response Dynamics in Gross Substitutes Markets". In: *Proceedings of the 26th ACM Conference on Economics and Computation*. 2025, pp. 389–389.

[CD11] Yang Cai and Constantinos Daskalakis. "On minmax theorems for multiplayer games". In: *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete algorithms (SODA)*. 2011.

[Che+23] Zaiwei Chen, Kaiqing Zhang, Eric Mazumdar, Asuman Ozdaglar, and Adam Wierman. "A finite-sample analysis of payoff-based independent learning in zero-sum stochastic games". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 75826–75883.

[Che+24] Zaiwei Chen, Kaiqing Zhang, Eric Mazumdar, Asuman Ozdaglar, and Adam Wierman. "Last-Iterate Convergence of Payoff-Based Independent Learning in Zero-Sum Stochastic Games". In: *arXiv preprint arXiv:2409.01447* (2024).

[COZ22] Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. "Finite-Time Last-Iterate Convergence for Learning in Multi-Player Games". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[COZ24] Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. "Accelerated algorithms for constrained nonconvex-nonconcave min-max optimization and comonotone inclusion". In: *Proceedings of the 41st International Conference on Machine Learning*. 2024, pp. 5312–5347.

[CP20] Xi Chen and Binghui Peng. "Hedging in games: Faster convergence of external and swap regrets". In: *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), pp. 18990–18999.

[Cut19] Ashok Cutkosky. "Anytime online-to-batch, optimism and acceleration". In: *International conference on machine learning*. PMLR. 2019, pp. 1446–1454.

[CWC21] Shicong Cen, Yuting Wei, and Yuejie Chi. "Fast policy extragradient methods for competitive games with entropy regularization". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 27952–27964.

[CWC24] Shicong Cen, Yuting Wei, and Yuejie Chi. "Fast policy extragradient methods for competitive games with entropy regularization". In: *Journal of machine learning Research* 25.4 (2024), pp. 1–48.

[CZ23a] Yang Cai and Weiqiang Zheng. "Accelerated Single-Call Methods for Constrained Min-Max Optimization". en. In: *Proceedings of the 11th International Conference on Learning Representations*. Feb. 2023.

[CZ23b] Yang Cai and Weiqiang Zheng. "Doubly optimal no-regret learning in monotone games". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 3507–3524.

[Das+18] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. "Training GANs with Optimism." In: *International Conference on Learning Representations (ICLR)*. 2018.

[DDK11] Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. "Near-optimal no-regret algorithms for zero-sum games". In: *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*. SIAM. 2011, pp. 235–254.

[DFG21] Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. "Near-optimal no-regret learning in general games". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2021).

[Dia20] Jelena Diakonikolas. "Halpern Iteration for Near-Optimal and Parameter-Free Monotone Inclusion and Strong Solutions to Variational Inequalities". In: *Conference on Learning Theory (COLT)*. 2020.

[DP18] Constantinos Daskalakis and Ioannis Panageas. "The limit points of (optimistic) gradient descent in min-max optimization". In: *the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS)*. 2018.

[DP19] Constantinos Daskalakis and Ioannis Panageas. "Last-Iterate Convergence: Zero-Sum Games and Constrained Min-Max Optimization". In: *10th Innovations in Theoretical Computer Science Conference (ITCS)*. 2019.

[DWY24] Jing Dong, Baoxiang Wang, and Yaoliang Yu. "Uncoupled and Convergent Learning in Monotone Games under Bandit Feedback". In: *arXiv preprint arXiv:2408.08395* (2024).

[FAI+22] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. "Human-level play in the game of Diplomacy by combining language models with strategic reasoning". In: *Science* 378.6624 (2022), pp. 1067–1074.

[FS99] Yoav Freund and Robert E Schapire. "Adaptive game playing using multiplicative weights". In: *Games and Economic Behavior* 29.1-2 (1999), pp. 79–103.

[Gol+20] Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. "Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems". In: *Conference on Learning Theory (COLT)*. 2020.

[GPD20] Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. "Tight last-iterate convergence rates for no-regret learning in multi-player games". In: *Advances in neural information processing systems (NeurIPS)* (2020).

[GTG22] Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. "Last-Iterate Convergence of Optimistic Gradient Method for Monotone Variational Inequalities". In: *Advances in Neural Information Processing Systems*. 2022.

[GVM21] Angeliki Giannou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. "On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 22655–22666.

[Hal67] Benjamin Halpern. "Fixed points of nonexpanding maps". In: *Bulletin of the American Mathematical Society* 73.6 (1967), pp. 957–961.

[HAM21] Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. "Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium". In: *Conference on Learning Theory*. PMLR. 2021, pp. 2388–2422.

[How72] Joseph T Howson Jr. "Equilibria of polymatrix games". In: *Management Science* 18.5-part-1 (1972), pp. 312–318.

[Hsi+22] Yu-Guan Hsieh, Kimon Antonakopoulos, Volkan Cevher, and Panayotis Mertikopoulos. "No-regret learning in games with noisy feedback: Faster rates and adaptivity via learning rate separation". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 6544–6556.

[Jan68] Elena Janovskaja. "Equilibrium points in polymatrix games". In: *Lithuanian Mathematical Journal* 8.2 (1968), pp. 381–384.

[JLZ24] Michael Jordan, Tianyi Lin, and Zhengyuan Zhou. "Adaptive, doubly optimal no-regret learning in strongly monotone and exp-concave games with gradient feedback". In: *Operations Research* (2024).

[Kor76] G. M. Korpelevich. "The extragradient method for finding saddle points and other problems". In: *Matecon* 12 (1976), pp. 747–756.

[Liu+23] Mingyang Liu, Asuman E Ozdaglar, Tiancheng Yu, and Kaiqing Zhang. "The Power of Regularization in Solving Extensive-Form Games". In: *The Eleventh International Conference on Learning Representations*. 2023.

[Liu+24] Yixin Liu, Argyris Oikonomou, Weiqiang Zheng, Yang Cai, and Arman Cohan. "COMAL: A Convergent Meta-Algorithm for Aligning LLMs with General Preferences". In: *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*. 2024.

[Liu+25] Kaizhao Liu, Qi Long, Zhekun Shi, Weijie J Su, and Jiancong Xiao. "Statistical impossibility and possibility of aligning llms with human preferences: From condorcet paradox to nash equilibrium". In: *arXiv preprint arXiv:2503.10990* (2025).

[Mao+24] Weichao Mao, Haoran Qiu, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, and Tamer Başar. "$\widetilde{O}(T^{-1})$ Convergence to (coarse) correlated equilibria in full-information general-sum Markov games". In: *6th Annual Learning for Dynamics & Control Conference*. PMLR. 2024, pp. 361–374.

[MPP18] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. "Cycles in adversarial regularized learning". In: *Proceedings of the twenty-ninth annual ACM-SIAM symposium on discrete algorithms*. SIAM. 2018, pp. 2703–2717.

[Mun+23] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. *Nash Learning from Human Feedback*. arXiv:2312.00886 [cs, stat]. Dec. 2023.

[Nas50] John F Nash Jr. "Equilibrium points in n-person games". In: *Proceedings of the national academy of sciences* 36.1 (1950), pp. 48–49.

[OSZ21] Asuman Ozdaglar, Muhammed O Sayin, and Kaiqing Zhang. "Independent learning in stochastic games". In: *International Congress of Mathematicians*. 2021.

[OX21] Yuyuan Ouyang and Yangyang Xu. "Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems". In: *Mathematical Programming* 185.1 (2021), pp. 1–35.

[Per+22] Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al. "Mastering the game of Stratego with model-free multiagent reinforcement learning". In: *Science* 378.6623 (2022), pp. 990–996.

[Pop80] Leonid Denisovich Popov. "A modification of the Arrow-Hurwicz method for search of saddle points". In: *Mathematical notes of the Academy of Sciences of the USSR* 28.5 (1980). Publisher: Springer, pp. 845–848.

[PZO23] Chanwoo Park, Kaiqing Zhang, and Asuman Ozdaglar. "Multi-player zero-sum markov games with networked separable interactions". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 37354–37369.

[RS13] Sasha Rakhlin and Karthik Sridharan. "Optimization, learning, and games with predictable sequences". In: *Advances in Neural Information Processing Systems* (2013).

[Sil+17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. "Mastering the game of go without human knowledge". In: *nature* 550.7676 (2017), pp. 354–359.

[SPF25] Ashkan Soleymani, Georgios Piliouras, and Gabriele Farina. "Faster Rates for No-Regret Learning in General Games via Cautious Optimism". In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. 2025.

[Swa+24] Gokul Swamy, Christoph Dann, Rahul Kidambi, Steven Wu, and Alekh Agarwal. "A Minimaximalist Approach to Reinforcement Learning from Human Feedback". In: *Forty-first International Conference on Machine Learning*. 2024.

[Syr+15] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. "Fast convergence of regularized learning in games". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2015).

[Tam14] Oskari Tammelin. *Solving Large Imperfect Information Games Using CFR+*. arXiv:1407.5042 [cs]. July 2014.

[Wei+21] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. "Linear Last-iterate Convergence in Constrained Saddle-point Optimization". In: *International Conference on Learning Representations (ICLR)*. 2021.

[Wu+24] Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. "Self-play preference optimization for language model alignment". In: *arXiv preprint arXiv:2405.00675* (2024).

[Wu+25] Jiduan Wu, Anas Barakat, Ilyas Fatkhullin, and Niao He. "Learning zero-sum linear quadratic games with improved sample complexity and last-iterate convergence". In: *SIAM Journal on Control and Optimization* 63.5 (2025), pp. 3244–3271.

[WZ07] Fang Wu and Li Zhang. "Proportional response dynamics leads to market equilibrium". In: *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. 2007, pp. 354–363.

[Ye+24] Chenlu Ye, Wei Xiong, Yuheng Zhang, Nan Jiang, and Tong Zhang. "A theoretical analysis of nash learning from human feedback under general kl-regularized preference". In: *arXiv preprint arXiv:2402.07314* (2024).

[YR21] TaeHo Yoon and Ernest K Ryu. "Accelerated Algorithms for Smooth Convex-Concave Minimax Problems with $\mathcal{O}(1/k^2)$ Rate on Squared Gradient Norm". In: *International Conference on Machine Learning (ICML)*. PMLR. 2021, pp. 12098–12109.

[ZDR22] Sihan Zeng, Thinh Doan, and Justin Romberg. "Regularized gradient descent ascent for two-player zero-sum Markov games". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 34546–34558.

[Zin03] Martin Zinkevich. "Online convex programming and generalized infinitesimal gradient ascent". In: *Proceedings of the 20th international conference on machine learning (ICML)*. 2003.

Contents

A Proof of Lemma 1

Proof. By the zero-sum property and the linearity of the utility, we have

$$\begin{aligned}
&\max_{x\in\times\Delta^{d_i}} \sum_{i=1}^{n} u_i(x_i, \overline{x}^T_{-i}) - u_i(\overline{x}^T) \\
&= \sum_{i=1}^{n} \max_{x_i\in\Delta^{d_i}} u_i(x_i, \overline{x}^T_{-i}) && (\textstyle\sum_{i=1}^n u_i(\overline{x}^T) = 0) \\
&= \frac{1}{T}\sum_{i=1}^{n} \max_{x_i\in\Delta^{d_i}} \sum_{t=1}^{T} u_i(x_i, x^t_{-i}) && (\text{Linearity of } u_i \text{ and } \overline{x}^T_{-i} = \tfrac{1}{T}\textstyle\sum_{t=1}^T x^t_{-i}) \\
&= \frac{1}{T}\sum_{i=1}^{n} \left(\max_{x_i\in\Delta^{d_i}} \sum_{t=1}^{T} u_i(x_i, x^t_{-i}) - \sum_{t=1}^{T} u_i(x^t) \right) && (\textstyle\sum_{i=1}^n u_i(x^t) = 0 \text{ for all } t) \\
&= \frac{1}{T}\sum_{i=1}^{n} \mathrm{Reg}_i(T).
\end{aligned}$$

This completes the proof. □

B `A2L` Reduction for Proportional Response Dynamics in Fisher Markets

The Fisher market is a classic model of allocating divisible goods, a special case of the Arrow-Debreu market. In a Fisher market, there are m agents and n divisible goods. Each agent $i \in [n]$ has a budget $B_i > 0$ and each good $j \in [m]$ has a unit supply. Each agent i also has a utility function $u_i : \mathbb{R}^m_{\geq 0} \to \mathbb{R}$ such that $u_i(x_i)$ is her utility of receiving the bundle $x_i \in \mathbb{R}^m_{\geq 0}$ where $x_{ij} \geq 0$ is the amount of good j. Here we do not assume u_i is linear.

A price vector $p \in \mathbb{R}^m_{\geq 0}$ specifies a price $p_j > 0$ for each good j. A *Competitive Equilibrium (CE)* of a Fisher market is a pair of price vector and personalized allocations $(p, \{x_i\})$ that satisfies the following properties:

1. *Budget Feasible:* $\langle p, x_i\rangle \leq B_i$ for each agent i.
2. *Utility Maximizing:* $u_i(x_i) = \max_{y_i\in\mathbb{R}^m_{\geq 0}:\langle p,y_i\rangle\leq e_i} u_i(y_i)$ for each agent i.
3. *Market Clears:* $\sum_i x_{ij} \leq 1$ for each good j and $\sum_i x_{ij} = 1$ if $p_j > 0$.

The *proportional response dynamics (PRD)* is a distributed, independent dynamics leading to a CE in Fisher market. PRD works as follows:

- *Budget Spending:* each agent i decides the allocation of her budget b^t_i, where b^t_{ij} is her spending on good j. It must hold that $\sum_j b^t_{ij} = e_i$.
- *Goods allocation:* the amount of good j allocated to agent i is proportional to their spending: $x^t_{ij} = \frac{b^t_{ij}}{\sum_i b^t_{ij}}$. Moreover, the price of each good j is defined as the sum of agents' spending $p^t_j = \sum_{i'} b^t_{i'j}$.
- *Update:* Based on the allocation x^t_i, each agent i then updates their spending of budget b^{t+1}_i in the next round using the following rule:

$$b^t_{ij} = B_i \frac{x^t_{ij}\nabla_j u_i(x^t_i)}{\sum_{j'} x^t_{ij'}\nabla_{j'} u_i(x^t_i)}.$$

Although it has been shown that the price vector $\{p^t\}$ converges to a Market equilibrium price vector for Fisher markets in many settings including linear utilities [WZ07; BDX11], PRD for Fisher markets with Gross Substitutes utilities has a $O(1/T)$ convergence rate only in terms of the average price $\{\frac{1}{t}\sum_{k=1}^t p^k\}$, but not the last iterate price $\{p^t\}$ [CCT25].

A2L-PRD for Last-Iterate Convergence We note that in PRD, the only feedback needed for their update is her allocation x_i^t, where good j's allocation $x_{ij}^t = \frac{b_{ij}^t}{\sum_{i'} b_{i'j}^t}$ is proportional according to every agent's spending. Here, the key observation is that the feedback in PRD, x_i^t, has a nice proportional structure with the spending b_i^t despite each player having nonlinear utilities. We can then apply our A2L reduction. We just let each agent spend their averaged budget allocation $\bar{b}_i^t = \frac{1}{t}\sum_{k=1}^t b_i^k$ and observe the induced allocation $\bar{x}_i^t$. Due to the proportional structure, each agent can recover the original iterate $\{b_i^t\}_{i\in[n]}$'s allocation and then update according to the original PRD [CCT25]. In this way, we get a modified A2L-PRD whose last-iterate price vectors is equivalent to the averaged price vectors in the original PRD. The A2L-PRD enjoys $O(1/T)$ last-iterate convergence rate in the price vectors.

C Last-Iterate Convergence with Gradient Feedback

C.1 Proof of Theorem 3

Proof. Let $\{x^t\}$ be the iterates of OMWU dynamics with η. Recall that we denote $u_i^t := u_i(\cdot, x_{-i}^t)$. By Theorem 1, we have

$$\begin{aligned}\sum_{i=1}^n \mathrm{Reg}_i(T) &\le \frac{\sum_{i=1}^n \log d_i}{\eta} + \sum_{i=1}^n\sum_{t=1}^T \left(\eta\left\|u_i^t - u_i^{t-1}\right\|_\infty^2 - \frac{1}{4\eta}\left\|x_i^t - x_i^{t-1}\right\|_1^2\right)\\ &\le \frac{\sum_{i=1}^n \log d_i}{\eta},\end{aligned}$$

where we use Lemma 2 and $\eta \le \frac{1}{2(n-1)}$ in the last inequalty. Invoking Lemma 1, the average-iterate has total gap bounded by $\textsc{TGap}(\overline{x}^T) \le \frac{\sum_{i=1}^n \log d_i}{\eta T}$ for all T. By the A2L reduction guarantee in Theorem 2, we conclude the proof. □

C.2 Proof of Corollary 1

Proof. By Theorem 3, we know $\overline{x}^t$ is an $\frac{\sum_{i=1}^n \log d_i}{\eta t}$-approximate Nash equilibrium for any $t \ge 1$. Thus, the dynamic regret of any player $i \in [n]$ is bounded by

$$\mathrm{DReg}_i^T = \sum_{t=1}^T \left(\max_{x_i\in\Delta^{d_i}} u_i(x_i, \overline{x}_{-i}^t) - u_i(\overline{x}^t)\right) \le \sum_{t=1}^T \frac{\sum_{i=1}^n \log d_i}{\eta t} = O\left(\frac{\sum_{i=1}^n \log d_i}{\eta}\log T\right).$$

This completes the proof. □

D Last-Iterate Convergence with Bandit Feedback

D.1 Analysis of Regret with Estimation Error

We first analyze the sequence $\{x^t\}$ generated in the subroutine (Line 8) of Algorithm 2, which is updated according to OMWU with the estimated utility sequence $\{\hat{u}_\varepsilon^t\}$.

We define $\Delta_i^t := \widehat{U}_{i,\varepsilon}^t - \overline{u}_{i,\varepsilon}^t$ as the error of estimation in Line 6. The error of estimation for u_i^t, denoted as $\delta_i^t := \hat{u}_{i,\varepsilon}^t - u_i^t$, can be bounded as follows.

Proposition 1. *For any $t \ge 1$ and i, $\|\delta_i^t\|_\infty \le \|t\Delta_i^t\| + \|(t-1)\Delta_i^{t-1}\| + 2\varepsilon_t$. This further implies $\|\delta_i^t\|_\infty^2 \le 3\|t\Delta_i^t\|^2 + 3\|(t-1)\Delta_i^{t-1}\|^2 + 12\varepsilon_t^2$.*

Proof. Let us define $u_{i,\mathrm{unif}} := u_i(\cdot, \mathcal{U}_{-i})$ the utility vector of i when other players use the uniform strategy. Since the utility $u_i(x_i, x_{-i})$ is linear in x_{-i}, we have

$$\overline{u}_{i,\varepsilon}^t = u_i(\cdot, x_{-i,\varepsilon}^t) = (1-\varepsilon_t)u_{-i}^t + \varepsilon_t u_{i,\mathrm{unif}}, \forall t.$$

Using the above equality and the definition of Δ_i^t and δ_i^t, as well as $\varepsilon_t = t^{-1}$, we have

$$\begin{aligned}
&t\Delta_i^t - (t-1)\Delta_i^{t-1} \\
&= t \cdot \widehat{U}_{i,\varepsilon}^t - (t-1) \cdot \widehat{U}_{i,\varepsilon}^{t-1} - t \cdot \overline{u}_{i,\varepsilon}^t + (t-1) \cdot \overline{u}_{i,\varepsilon}^{t-1} \\
&= \hat{u}_{i,\varepsilon}^t - t \cdot \left((1-\varepsilon_t)\overline{u}_i^t + \varepsilon_t u_{i,\mathrm{unif}}\right) + (t-1) \cdot \left((1-\varepsilon_{t-1})\overline{u}_i^{t-1} + \varepsilon_{t-1} u_{i,\mathrm{unif}}\right) \\
&= \hat{u}_{i,\varepsilon}^t - (1-\varepsilon_t)\sum_{k=1}^{t} u_i^k + (1-\varepsilon_{t-1})\sum_{k=1}^{t-1} u_i^k \\
&= \hat{u}_{i,\varepsilon}^t - u_i^t + \varepsilon_t u_i^t + (\varepsilon_t - \varepsilon_{t-1})\sum_{k=1}^{t-1} u_i^k \\
&= \delta_i^t + \varepsilon_t u_i^t + (\varepsilon_t - \varepsilon_{t-1})\sum_{k=1}^{t-1} u_i^k. \qquad (2)
\end{aligned}$$

Since $\|u_i^k\|_\infty \le 1$ for all k, the above equality implies

$$\begin{aligned}
\left\|\delta_i^t\right\|_\infty &\le \left\|t\Delta_i^t\right\| + \left\|(t-1)\Delta_i^{t-1}\right\| + \varepsilon_t + (t-1)(\varepsilon_t - \varepsilon_{t-1}) \\
&= \left\|t\Delta_i^t\right\| + \left\|(t-1)\Delta_i^{t-1}\right\| + 2\varepsilon_t.
\end{aligned}$$

We then apply the basic inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ to bound $\|\delta_i^t\|_\infty^2$. This completes the proof. □

In the following, we show that the regret against the original utility sequence $\{u^t\}$ can be bounded by RVU terms with an additional summation of second-order error terms. We remark that a naive analysis would lead to a summation of first-order terms, which gives a suboptimal convergence rate.

Lemma 3. *Let $\{x_i^t\}$ be the iterates of* OMWU *with step size $\eta > 0$ and utilities $\{\hat{u}_i^t\}$, we have*

$$\begin{aligned}
\max_{x_i \in \Delta^{d_i}} \sum_{t=1}^{T} \left\langle u_i^t, x_i^t - x_i \right\rangle \le \frac{\log d_i}{\eta} + 4\eta \sum_{t=1}^{T} \left\|u_i^t - u_i^{t-1}\right\|_\infty^2 - \frac{1}{8\eta}\sum_{t=1}^{T}\left\|x_i^t - x_i^{t-1}\right\|_1^2 \\
+ 2\left\|T\Delta_i^T\right\|_\infty + 26\eta\sum_{t=1}^{T-1}\left\|t\Delta_i^t\right\|_\infty^2 + 2\sum_{t=1}^{T}\varepsilon_t + 16\pi^2\eta.
\end{aligned}$$

Proof. **Analysis** By RVU property of OMWU (Theorem 1), we can bound the regret for the estimated utilities $\{\hat{u}_{i,\varepsilon}^t\}$: $\forall x \in \Delta^{d_i}$,

$$\sum_{t=1}^{T}\left\langle \hat{u}_{i,\varepsilon}^t, x_i - x_i^t\right\rangle \le \frac{\log d_i}{\eta} + \eta\sum_{t=1}^{T}\left\|\hat{u}_{i,\varepsilon}^t - \hat{u}_{i,\varepsilon}^{t-1}\right\|_\infty^2 - \frac{1}{4\eta}\sum_{t=1}^{T}\left\|x_i^t - x_i^{t-1}\right\|_1^2$$

Then by definition of $\hat{u}_{i,\varepsilon}^t = u_i^t + \delta_i^t$, we can bound the regret for the true utilities $\{u_i^t\}$ by

$$\begin{aligned}
\sum_{t=1}^{T}\left\langle u_i^t, x_i - x_i^t\right\rangle \le \frac{\log d_i}{\eta} + 4\eta\sum_{t=1}^{T}\left\|u_i^t - u_i^{t-1}\right\|_\infty^2 - \frac{1}{4\eta}\sum_{t=1}^{T}\left\|x_i^t - x_i^{t-1}\right\|_1^2 \\
+ \underbrace{\sum_{t=1}^{T}\left\langle \delta_i^t, x_i^t - x_i\right\rangle}_{\mathtt{I}} + \underbrace{4\eta\sum_{t=1}^{T}\left(\left\|\delta_i^t\right\|_\infty^2 + \left\|\delta_i^{t-1}\right\|_\infty^2\right)}_{\mathtt{II}}. \qquad (3)
\end{aligned}$$

We note that the first three terms are standard terms in the RVU bound [Syr+15]. We focus on the error terms I and II. The term II is a summation of second-order terms in $\|\delta_i^t\|_\infty^2$ and we can bound it using Proposition 1 and $\sum_{t=1}^{\infty}\varepsilon_t^2 = \sum_{t=1}^{\infty}\frac{1}{t^2} = \frac{\pi^2}{6}$:

$$\mathtt{II} = 4\eta\sum_{t=1}^{T}\left(\left\|\delta_i^t\right\|_\infty^2 + \left\|\delta_i^{t-1}\right\|_\infty^2\right) \le 24\eta\sum_{t=1}^{T}\left\|t\Delta_i^t\right\|_\infty^2 + 96\eta\sum_{t=1}^{T}\varepsilon_t^2 \le 24\eta\sum_{t=1}^{T}\left\|t\Delta_i^t\right\|_\infty^2 + 16\pi^2\eta. \qquad (4)$$

However, a naive analysis of term $\mathtt{I}$ leads to a summation of first-order terms: by Cauchy-Schwarz, we have $\mathtt{I} \le \sum_{t=1}^{T} \|\delta_i^t\|_\infty \|x_i^t - x_i\|_1 \le 2\sum_{t=1}^{T} \|\delta_i^t\|_\infty \le O(\sum_{t=1}^{T} \|t\Delta_i^t\|_\infty) + O(\sum_{t=1}^{T} \varepsilon_t)$. If we tune the parameters optimally, the existence of the first-order error term $O(\sum_{t=1}^{T} \|t\Delta_i^t\|_\infty)$ would result in an $O(T^{-\frac{1}{7}})$ rate, worse than the claimed $O(T^{-\frac{1}{5}})$ rate.

Improved Analysis of Term $\mathtt{I}$ In the following, we give an improved analysis that bounds term $\mathtt{I}$ by second-order error terms. The key is to rewrite term $\mathtt{I}$ and use the negative term $-\sum_{t=1}^{T} \|x_i^t - x_i^{t-1}\|_1^2$.

By (2), we have

$$\delta_i^t = \hat{u}_{i,\varepsilon}^t - u_i^t = t\Delta_i^t - (t-1)\Delta_i^{t-1} - \varepsilon_t u_i^t - (\varepsilon_t - \varepsilon_{t-1})\sum_{k=1}^{t-1} u_i^k.$$

We note that $\|\varepsilon_t u_i^t - (\varepsilon_t - \varepsilon_{t-1})\sum_{k=1}^{t-1} u_i^k\|_\infty \le 2\varepsilon_t$ as the utility is bounded in $[0,1]$. This implies

$$\begin{aligned}
\mathtt{I} &= \sum_{t=1}^{T} \left\langle \delta_i^t, x_i^t - x_i \right\rangle \\
&\le \sum_{t=1}^{T} \left\langle t\Delta_i^t - (t-1)\Delta_i^{t-1}, x_i^t - x_i \right\rangle + \sum_{t=1}^{T} \left\| \varepsilon_t u_i^t - (\varepsilon_t - \varepsilon_{t-1})\sum_{k=1}^{t-1} u_i^k \right\|_\infty \cdot \left\| x_i^t - x_i \right\|_1. \\
&\le \left\langle T\Delta_i^T, x_i^T - x_i \right\rangle + \sum_{t=1}^{T-1} \left\langle t\Delta_i^t, x_i^t - x_i^{t+1} \right\rangle + 4\sum_{t=1}^{T} \varepsilon_t \\
&\le 2\left\| T\Delta_i^T \right\|_\infty + \sum_{t=1}^{T-1} \left\| t\Delta_i^t \right\|_\infty \left\| x_i^t - x_i^{t+1} \right\|_1 + 4\sum_{t=1}^{T} \varepsilon_t.
\end{aligned}$$

As a result, we can combine $-\frac{1}{8\eta}\sum_{t=1}^{T} \|x_i^t - x_i^{t-1}\|_1^2$ and using basic inequality $\langle a, b\rangle - \|b\|^2 \le \frac{1}{4}\|a\|^2$ to get

$$\begin{aligned}
\mathtt{I} - \frac{1}{8\eta}\sum_{t=1}^{T} \left\| x_i^t - x_i^{t-1} \right\|_1^2 &\le 2\left\| T\Delta_i^T \right\|_\infty + \sum_{t=1}^{T-1} \left\| t\Delta_i^t \right\|_\infty \left\| x_i^t - x_i^{t+1} \right\|_1 - \frac{1}{8\eta}\sum_{t=1}^{T} \left\| x_i^t - x_i^{t-1} \right\|_1^2 + 4\sum_{t=1}^{T} \varepsilon_t \\
&\le 2\left\| T\Delta_i^T \right\|_\infty + 2\eta\sum_{t=1}^{T-1} \left\| t\Delta_i^t \right\|_\infty^2 + 4\sum_{t=1}^{T} \varepsilon_t. \qquad (5)
\end{aligned}$$

Combining (3), (4), and (5), we get

$$\begin{aligned}
\max_{x_i \in \Delta^{d_i}} \sum_{t=1}^{T} \left\langle u_i^t, x_i^t - x_i \right\rangle \le \frac{\log d_i}{\eta} &+ 4\eta\sum_{t=1}^{T} \left\| u_i^t - u_i^{t-1} \right\|_\infty^2 - \frac{1}{8\eta}\sum_{t=1}^{T} \left\| x_i^t - x_i^{t-1} \right\|_1^2 \\
&+ 2\left\| T\Delta_i^T \right\|_\infty + 26\eta\sum_{t=1}^{T-1} \left\| t\Delta_i^t \right\|_\infty^2 + 4\sum_{t=1}^{T} \varepsilon_t + 16\pi^2\eta.
\end{aligned}$$

This completes the proof. □

D.2 Proof of Theorem 4

Recall that $d = \max_{i\in[n]} d_i$. By Lemma 3, we can bound the social regret of all players with respect to the iterates $\{x^t\}$ and utilities $\{u^t\}$ by

$$\sum_{i=1}^{n} \mathrm{Reg}_i(T, \{x_i^t\}, \{u_i^t\}) \le \frac{n\log d}{\eta} + \underbrace{\sum_{i=1}^{n}\sum_{t=1}^{T}\left(4\eta\left\|u_i^t - u_i^{t-1}\right\|_\infty^2 - \frac{1}{8\eta}\left\|x_i^t - x_i^{t-1}\right\|_1^2\right)}_{\mathtt{I}} + \underbrace{\sum_{i=1}^{n}\left(2\left\|T\Delta_i^T\right\|_\infty + 26\eta\sum_{t=1}^{T-1}\left\|t\Delta_i^t\right\|_\infty^2\right)}_{\mathtt{II}} + 4n\sum_{t=1}^{T}\varepsilon_t + 16n\pi^2\eta. \quad (6)$$

By Lemma 2 and the fact that $\eta \le \frac{1}{6n}$, we have $\mathtt{I} \le 0$.

In the following, we analyze term $\mathtt{II}$ and conclude the last-iterate convergence rate of Algorithm 2 for $\{\varepsilon_t = t^{-1}\}$ and different choices of $\{B_t\}$. All these choices give an $O(k^{-\frac{1}{5}})$ last-iterate convergence rate. The difference is that the choices $\{B_t = dt^4\}$ lead to better dependence on d or δ but require knowledge about d or δ, while the choice of $\{B_t = t^4\}$ is fully uncoupled as each player does not need to know even an upper bound of d.

Case 1: $B_t = d\cdot t^4$ We note that for $T = O(\log(\frac{1}{\delta}))$, it trivially holds that $\sum_{i=1}^n \mathrm{Reg}_i(T, \{x_i^t\}, \{u_i^t\}) \le nT = O(n\log(\frac{1}{\delta}))$. Thus $\bar{x}^T$ is an $O(\frac{n\log(\frac{1}{\delta})}{T})$-approximate Nash equilibrium for all $T = O(\log\frac{1}{\delta})$. In the following, we focus on $T \ge \log(\frac{1}{\delta})$.

By Lemma 5, we have with probability $1 - O(nd^3\delta)$,

$$\max_{i\in[n]}\left\|\Delta_i^t\right\|_\infty \le 2\sqrt{\frac{d\log(\frac{B_t t^2}{\delta})}{B_t\varepsilon_t}}, \forall t \ge \log\left(\frac{1}{\delta}\right).$$

For $t \le d\log(\frac{1}{\delta})$, we can use the trivial bound of $\|\Delta_i^t\|_\infty \le 1$. Then we can bound term $\mathtt{II}$ by

$$\mathtt{II} \le 4nT\sqrt{\frac{d\log(\frac{B_T T^2}{\delta})}{B_T\varepsilon_T}} + 104n\eta\sum_{t=1}^{T-1}\frac{t^2\cdot d\log(\frac{B_t t^2}{\delta})}{B_t\varepsilon_t} + 26\eta\log\left(\frac{1}{\delta}\right). \quad (7)$$

Combining (6) and (7), and $B_t = d\cdot t^4$ and $\varepsilon_t = t^{-1}$, we get

$$\begin{aligned}
&\sum_{i=1}^{n}\mathrm{Reg}_i(T, \{x_i^t\}, \{u_i^t\}) \\
&\le \frac{n\log d}{\eta} + 4nT\sqrt{\frac{d\log(\frac{B_T T^2}{\delta})}{B_T\varepsilon_T}} + 104n\eta\sum_{t=1}^{T-1}\frac{t^2\cdot d\log(\frac{B_t t^2}{\delta})}{B_t\varepsilon_t} + 26\eta\log\left(\frac{1}{\delta}\right) + 4n\sum_{t=1}^{T}\varepsilon_t + 16n\pi^2\eta \\
&= O\left(\frac{n\log^2(\frac{dT}{\delta})}{\eta}\right).
\end{aligned}$$

By Lemma 1, we have the average iterate $\overline{x}^T := \frac{1}{T}\sum_{t=1}^T x^t$ is an $O(\frac{n\log^2(\frac{dT}{\delta})}{\eta T})$-approximate Nash equilibrium. Since $\varepsilon_T = \frac{1}{T}$, then we know $\overline{x}_\varepsilon^T := (1-\varepsilon_T)\overline{x}^T + \varepsilon_T\cdot\otimes_{i=1}^n\mathrm{Uniform}(d_i)$ is also an $O(\frac{n\log^2(\frac{dT}{\delta})}{\eta T})$-approximate Nash equilibrium for any $T \ge 1$.

By the choice of the epoch length $B_t = d\cdot t^4$, we have that the actual sequence $\{\overline{y}^k\}$ generated by `A2L-OMWU-Bandit` satisfies $\overline{y}_\varepsilon^k = \overline{x}_\varepsilon^{t_k}$ where $t_k = \Theta(d^{-\frac{1}{5}}k^{\frac{1}{5}})$. Thus we can conclude that with probability at least $1 - O(nd^3\delta)$, it holds for all $k \ge 1$ that $\overline{y}^k$ is a β_k-approximate Nash equilibrium with

$$\beta_k = O\left(\frac{n\log^2(\frac{kd}{\delta})}{\eta}\cdot d^{\frac{1}{5}}k^{-\frac{1}{5}}\right).$$

This completes the proof.

Case 2: $B_t = t^4$ The proof is very similar to case 1. We note that for $T = O(\log(\frac{1}{\delta}))$, it trivially holds that $\sum_{i=1}^n \mathrm{Reg}_i(T, \{x_i^t\}, \{u_i^t\}) \le nT = O(n\log(\frac{1}{\delta}))$. Thus $\bar{x}^T$ is an $O(\frac{n\log(\frac{1}{\delta})}{T})$-approximate Nash equilibrium for all $T = O(\log\frac{1}{\delta})$. In the following, we focus on $T \ge \log(\frac{1}{\delta})$. By item 2 in Lemma 5, with probability $1 - O(nd^3\delta)$, for all $t \ge d\log(\frac{1}{\delta})$,

$$\max_{i\in[n]} \left\|\Delta_i^t\right\|_\infty \le 2\sqrt{\frac{d\log\left(\frac{B_t t^2}{\delta}\right)}{B_t\varepsilon_t}}.$$

For $t \le \log(\frac{1}{\delta})$, we can use the trivial bound of $\|\Delta_i^t\|_\infty \le 1$. These gives

$$\texttt{II} \le 4nT\sqrt{\frac{d\log\left(\frac{B_T T^2}{\delta}\right)}{B_T\varepsilon_T}} + 104n\eta\sum_{t=1}^{T-1}\frac{t^2\cdot d\log\left(\frac{B_t t^2}{\delta}\right)}{B_t\varepsilon_t} + 26\eta\log(\frac{1}{\delta}). \tag{8}$$

Now can combine (6) and (8) with $B_t = t^4$ and $\varepsilon_t = t^{-1}$ and get

$$\begin{aligned}
&\sum_{i=1}^n \mathrm{Reg}_i(T, \{x_i^t\}, \{u_i^t\}) \\
&\le \frac{n\log d}{\eta} + 4nT\sqrt{\frac{d\log\left(\frac{B_T T^2}{\delta}\right)}{B_T\varepsilon_T}} + 104n\eta\sum_{t=1}^{T-1}\frac{t^2\cdot d\log\left(\frac{B_t t^2}{\delta}\right)}{B_t\varepsilon_t} + 26\eta\log(\frac{1}{\delta}) + 4n\sum_{t=1}^{T}\varepsilon_t + 16n\pi^2\eta \\
&= O\left(\frac{nd\log^2(\frac{dT}{\delta})}{\eta}\right).
\end{aligned}$$

Compared to case 1, this regret bound has an additional d dependence.

Similar to the analysis in the former case, we have $\bar{x}_\varepsilon^T$ is an $O(\frac{nd\log^2(\frac{dT}{\delta})}{\eta})$-approximate Nash equilibrium for all $T \ge 1$. By the choice of $B_t = t^4$, we have that the actual sequence $\{\overline{y}^k\}$ generated by `A2L-OMWU-Bandit` satisfies $\overline{y}_\varepsilon^k = \overline{x}_\varepsilon^{t_k}$ where $t_k = \Theta(k^{\frac{1}{5}})$. Thus we can conclude that with probability at least $1 - O(nd^3\delta)$, it holds for all $k \ge 1$ that $\overline{y}^k$ is a β_k-approximate Nash equilibrium with

$$\beta_k = O\left(\frac{n\log^2(\frac{dk}{\delta})}{\eta}\cdot dk^{-\frac{1}{5}}\right).$$

This completes the proof.

D.3 Estimation

In this subsection, we analyze the estimation error $\Delta_i^t := \widehat{U}_{i,\varepsilon}^t - \overline{u}_{i,\varepsilon}^t$ for any $i \in [n]$. Recall that $d := \max_{i\in[n]} d_i$. Within the epoch of length B_t, with high probability, each action receives at least $\Omega(\frac{B_t\varepsilon_t}{d})$ samples. We have with high probability that $|(\widehat{U}_{i,\varepsilon}^t - \overline{u}_{i,\varepsilon}^t)[a]| \le \tilde{O}(\frac{\sqrt{d}}{\sqrt{B_t\varepsilon_t}})$ for all action a, which implies $\|\Delta_i^t\|_\infty \le \tilde{O}(\sqrt{\frac{d}{B_t\varepsilon_t}})$. The formal guarantees are as follows.

Lemma 4. *In Algorithm 2, with probability* $1 - \frac{d\delta}{t^2} - d\exp\left(-\frac{B_t\varepsilon_t^2}{2d^2}\right)$*, we have*

$$\left\|\Delta_{i,\varepsilon}^t\right\|_\infty = \left\|\widehat{U}_i^t - \overline{u}_i^t\right\|_\infty \le 2\sqrt{\frac{d\log\left(\frac{B_t t^2}{\delta}\right)}{B_t\varepsilon_t}}.$$

Proof. Define $N_a^t := \sum_{k=1}^{B_t}\mathbb{I}[a^k = a]$ to be the number of samples for action a. Since $\Pr[a^k = a] \ge \frac{\varepsilon_t}{d_i} \ge \frac{\varepsilon_t}{d}$, we have $\mathbb{E}[N_a^t] \ge \frac{B_t\varepsilon_t}{d}$. By Hoeffding's inequality, we have

$$\Pr\left[N_a^t \le \frac{B_t\varepsilon_t}{2d}\right] \le \exp\left(-\frac{B_t\varepsilon_t^2}{2d^2}\right).$$

Thus with probability $1 - d\exp\left(-\frac{B_t\varepsilon_t^2}{2d^2}\right)$, every action has been sampled at least $\frac{B_t\varepsilon_t}{2d}$ times, i.e., $N_a^t \geq \frac{B_t\varepsilon_t}{2d}$.

Note that the number of samples N_a^t for action a is a random variable, so we can not directly use Azuma-Hoeffding's inequality to argue $|\widehat{U}_{i,\varepsilon}^t[a] - \overline{u}_{i,\varepsilon}^t[a]| \leq \sqrt{\frac{d\log(1/\delta)}{B_t\varepsilon_t}}$ with probability $1-\delta$. We define a sequence of random variables where $a_{-i}^k \sim \overline{x}_{-i,\varepsilon}^t$ and $r^k = u_i(a, a_{-i}^k)$, and define $\widehat{U}_m^t[a] := \frac{1}{m}\sum_{k=1}^m r^k$. Thus $\widehat{U}_m^t$ is an unbiased estimator for $\overline{u}_{i,\varepsilon}^t[a]$ for all $m \in [1, B_t]$. Then we can use Azuma-Hoeffding's inequality to get that

$$\begin{aligned}
&\Pr\left[|\widehat{U}_{i,\varepsilon}^t[a] - \overline{u}_{i,\varepsilon}^t[a]| \geq \sqrt{\frac{2\log\left(\frac{B_t t^2}{\delta}\right)}{N_a^t}}\right] \\
&\leq \Pr\left[\exists m \in [B_t], |\widehat{U}_m^t[a] - \overline{u}_{i,\varepsilon}^t[a]| \geq \sqrt{\frac{2\log\left(\frac{B_t t^2}{\delta}\right)}{m}}\right] \\
&\leq \sum_{m=1}^{B_t} \Pr\left[|\widehat{U}_m^t[a] - \overline{u}_{i,\varepsilon}^t[a]| \geq \sqrt{\frac{2\log\left(\frac{B_t t^2}{\delta}\right)}{m}}\right] \\
&\leq \sum_{m=1}^{B_t} \frac{\delta}{B_t t^2} = \frac{\delta}{t^2}.
\end{aligned}$$

Recall that with probability $1 - d\exp\left(-\frac{B_t\varepsilon_t^2}{2d^2}\right)$, $N_a^t \geq \frac{B_t\varepsilon_t}{2d}$ for all a. By a union bound over actions, we get with probability $1 - \frac{d\delta}{t^2} - d\exp\left(-\frac{B_t\varepsilon_t^2}{2d^2}\right)$,

$$\left\|\widehat{U}_{i,\varepsilon}^t - \overline{u}_{i,\varepsilon}^t\right\|_\infty \leq 2\sqrt{\frac{d\log\left(\frac{B_t t^2}{\delta}\right)}{B_t\varepsilon_t}}.$$

This completes the proof. □

Using a union bound over all players $i \in [n]$ and epoch $t \geq 1$, we have the following guarantee.

Lemma 5. *Consider a polymatrix game in which each player has at most d actions. Let $\{\Delta_i^t = U_{i,\varepsilon}^t - u_{i,\varepsilon}^t\}_{t\geq 1}$ denote the estimation error vector produced by Algorithm 2 for player i at round t. Suppose every player runs Algorithm 2 with $B_t \geq t^4$ and $\varepsilon_t = t^{-1}$. Then, with probability at least $1 - O(nd^3\delta)$, simultaneously for all players $i \in [n]$ and all rounds $t \geq \log(\frac{1}{\delta})$,*

$$\left\|\Delta_i^t\right\|_\infty \leq 2\sqrt{\frac{d\log\left(\frac{B_t t^2}{\delta}\right)}{B_t\varepsilon_t}}.$$

Proof. By Lemma 4, for each fixed player $i \in [n]$ and round $t \geq 1$, we have

$$\Pr\left(\left\|\Delta_i^t\right\|_\infty > 2\sqrt{\frac{d\log\left(\frac{B_t t^2}{\delta}\right)}{B_t\varepsilon_t}}\right) \leq \frac{d\delta}{t^2} + d\exp\left(-\frac{B_t\varepsilon_t^2}{2d^2}\right).$$

Applying a union bound over all players $i \in [n]$ and all rounds $t \geq \log(\frac{1}{\delta})$ gives that the claim holds with probability at least

$$1 - nd\delta\sum_{t=1}^{\infty}\frac{1}{t^2} - nd\sum_{t=\log(\frac{1}{\delta})}^{\infty}\exp\left(-\frac{B_t\varepsilon_t^2}{2d^2}\right).$$

We have $\sum_{t=1}^{\infty}\frac{1}{t^2} = O(1)$. With $\varepsilon_t = t^{-1}$ and $B_t \geq t^4$, we have $B_t\varepsilon_t^2 \geq t^2$, so the second summation is bounded by

$$nd\sum_{t=\log(\frac{1}{\delta})}^{\infty}\exp\left(-\frac{t^2}{2d^2}\right) \leq nd\delta\sum_{t=1}^{\infty}\exp\left(-\frac{t}{2d^2}\right) \leq O(nd^3\delta).$$

So the overall failure probability is at most $O(nd^3\delta)$. This completes the proof. □

E Robustness against Adversaries in the Bandit Setting

Consider the following utility estimation subroutine equipped with Algorithm 2. Here, we no longer assume the other players employ the same algorithm, and they might behave adversarially. To this end, in each epoch t, we denote the true utility vectors for player i as $\{u_i^{t,j} \in [0,1]^{d_i}\}_{j\in[B_t]}$ and the true accumulated utility within epoch t as

$$U_i^t = \sum_{j=1}^{B_t} u_i^{t,j}$$

In each epoch t, the player plays actions $\{a^{t,j} \sim x_{i,\varepsilon}^t\}$ for $j \in [B_t]$ and receives $\{r^{t,j} = u_i^{t,j}[a^j]\}_{j\in[B_t]}$, it also maintains an importance weighted estimator of the accumulated utility within epoch B_t:

$$\tilde{u}_i^{t,j}[b] = \frac{r^j \cdot \mathbf{1}[b = a^j]}{x_{i,\varepsilon}^t[b]}, \forall b \in [d_i], \quad \widetilde{U}_i^t = \sum_{j=1}^{B_t} \tilde{u}_i^{t,j}.$$

Denote $T_t = \sum_{k=1}^t B_t$, the player maintains estimated regret

$$\widetilde{\mathrm{Reg}}^{T_t} = \max_{x\in\Delta^{d_i}} \sum_{k=1}^t \left\langle \widetilde{U}_i^k, x \right\rangle - \sum_{k=1}^t \left\langle \widetilde{U}_i^k, x_{i,\varepsilon}^k \right\rangle,$$

while the true regret is

$$\mathrm{Reg}^T = \max_{x\in\Delta^{d_i}} \sum_{k=1}^t \langle U_i^k, x\rangle - \sum_{k=1}^t \langle U_i^k, x_{i,\varepsilon}^t\rangle.$$

Then we have the following

Proposition 2. *Fix any $\delta > 0$. With probability at least $1-\delta$, for any $t \geq 1$ and $T_t = \sum_{k=1}^t B_t$, we have*

$$\mathrm{Reg}^{T_t} \in \left[\widetilde{\mathrm{Reg}}^{T_t} - 4d_i\sqrt{\sum_{k=1}^t (k^2 B_k)\log\left(\frac{\pi^2 d_i t^2}{3\delta}\right)}, \widetilde{\mathrm{Reg}}^{T_t} + 4d_i\sqrt{\sum_{k=1}^t (k^2 B_k)\log\left(\frac{\pi^2 d_i t^2}{3\delta}\right)}\right]$$

Proof. The importance weighed estimator is unbiased as $\mathbb{E}[\tilde{u}_i^{t,j}[b]] = u_i^{t,j}[b]$ for all $b \in [d_i]$. Moreover, since $x_{i,\varepsilon}^t[b] \geq \frac{1}{d_i t}$, we have $-1 \leq \tilde{u}_i^{t,j}[b] - u_i^{t,j}[b] \leq d_i t$. By Azuma-Hoeffdind inequality, we have for any $t \geq 1$ and $b \in [d_i]$,

$$\Pr\left[\left|\sum_{k=1}^t (\widetilde{U}_i^k[b] - U_i^k[b])\right| \leq 2d_i\sqrt{\sum_{k=1}^t (k^2 B_k)\log\left(\frac{\pi^2 d_i t^2}{3\delta}\right)}\right] \geq 1 - \frac{6\delta}{\pi^2 d_i t^2}.$$

Using a union bound over $t \geq 1$ and $b \in [d_i]$ gives

$$\Pr\left[\forall t \geq 1, \left\|\sum_{k=1}^t (\widetilde{U}_i^k - U_i^k)\right\|_\infty \leq 2d_i\sqrt{\sum_{k=1}^t (k^2 B_k)\log\left(\frac{\pi^2 d_i t^2}{3\delta}\right)}\right] \geq 1 - \delta.$$

Since the $|\max v - \max v'| \leq \|v - v\|_\infty$, we can bound the error in regret estimation as

$$\widetilde{\mathrm{Reg}}^T - 2\left\|\sum_{k=1}^t (\widetilde{U}_i^k - U_i^k)\right\|_\infty \leq \mathrm{Reg}^T \leq \widetilde{\mathrm{Reg}}^T + 2\left\|\sum_{k=1}^t (\widetilde{U}_i^k - U_i^k)\right\|_\infty.$$

This implies with probability at least $1-\delta$, for all $t \geq 1$ and $T_t = \sum_{k=1}^t B_k$, we have

$$\mathrm{Reg}^{T_t} \in \left[\widetilde{\mathrm{Reg}}^T - 4d_i\sqrt{\sum_{k=1}^t (k^2 B_k)\log\left(\frac{\pi^2 d_i t^2}{3\delta}\right)}, \widetilde{\mathrm{Reg}}^T + 4d_i\sqrt{\sum_{k=1}^t (k^2 B_k)\log\left(\frac{\pi^2 d_i t^2}{3\delta}\right)}\right].$$

This completes the proof. □

Choosing $B_t = t^4$, we have $T_t = \sum_{k=1}^{t} B_k = \theta(t^5)$ and $t = \Theta(T_t^{\frac{1}{5}})$. We have, with probability $1-\delta$, for all T,

$$\mathrm{Reg}^{T_t} = \widetilde{\mathrm{Reg}}^{T_t} \pm \Theta\left(d_i t^{3.5} \log\left(\frac{d_i t}{\delta}\right)\right) = \widetilde{\mathrm{Reg}}^{T_t} \pm \Theta\left(d_i T_t^{\frac{7}{10}} \log\left(\frac{d_i T_t}{\delta}\right)\right).$$

We note that for general $T \in [T_t, T_{t+1}]$ (denote $T_0 = 1$), we have $B_{t+1} = (t+1)^4 = \Theta(T_t^{\frac{4}{5}})$

$$\begin{aligned}
\mathrm{Reg}^{T} &:= \max_{x \in \Delta^{d_i}} \left(\sum_{k=1}^{t} \left\langle U_i^k, x\right\rangle + \sum_{j=1}^{T-T_t} \left\langle u_i^{t+1,j}, x\right\rangle\right) - \left(\sum_{k=1}^{t} \left\langle U_i^k, x_{i,\varepsilon}^k\right\rangle + \sum_{j=1}^{T-T_t} \left\langle u_i^{t+1,j}, x_{i,\varepsilon}^{t+1}\right\rangle\right) \\
&\le \mathrm{Reg}^{T_t} + B_{t+1} \\
&= \widetilde{\mathrm{Reg}}^{T_t} + \max\left\{\Theta\left(T_t^{\frac{4}{5}}\right), \Theta\left(d_i T_t^{\frac{7}{10}} \log\left(\frac{d_i T_t}{\delta}\right)\right)\right\} \\
&= \widetilde{\mathrm{Reg}}^{T_t} + \max\left\{\Theta\left(T^{\frac{4}{5}}\right), \Theta\left(d_i T^{\frac{7}{10}} \log\left(\frac{d_i T}{\delta}\right)\right)\right\}.
\end{aligned}$$

Thus whenever $\widetilde{\mathrm{Reg}}^{T_t} = O(T_t^{\frac{4}{5}})$, we have $\mathrm{Reg}^{T} = O(T^{\frac{4}{5}})$ for all iterations $T \in [T_t, T_{t+1}]$.

Robustness in Adversarial Setting The player could run Algorithm 2 and track its regret using the above estimation $\widetilde{\mathrm{Reg}}^{T_t}$ at the end of each epoch t. Whenever she detects that $\widetilde{\mathrm{Reg}}^{T_t} = \Omega(T_t^{\frac{4}{5}})$, which cannot happen if all the players employ Algorithm 2, she can switch to a standard no-regret bandit algorithm with worst-case optimal regret. This procedure guarantees a worst-case $\widetilde{O}(T^{\frac{4}{5}})$ regret.

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: The main contribution of the paper is a reduction from last-iterate convergence to average-iterate convergence and the new state-of-the-art last-iterate convergence rates of uncoupled learning dynamics in zero-sum polymatrix games with gradient or bandit feedback. See the contribution paragraph in the introduction for details.

   Guidelines:

   - The answer NA means that the 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 discuss the limitations of our reduction in Section 3.

   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: We provide the full set of assumptions in the preliminaries and the statement of each theoretical result. The compltete proofs of the theoretical results are in the main body and the appendix.

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: [NA]

   Justification: This paper does not have experimental results.

   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: [NA]

Justification: This paper does not include experiements requiring code.

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, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: This paper does not include experiments.

   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: [NA]

   Justification: This paper does not include experiments.

   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: [NA]

   Justification: This paper does not include experiments.

   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 is theoretical and is conducted with 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: [NA]

    Justification: This paper contains purely theoretical results and the authors do not see any societal impact that needs to be discussed.

    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: [NA]

    Justification: This paper does not release new assets.

    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.