Near-Optimal Last-Iterate Convergence for Zero-Sum Games with Bandit Feedback and Opponent Actions

Soumita Hait*
University of Southern California
hait@usc.edu

Ping Li*
National University of Singapore
pingli@nus.edu.sg

Haipeng Luo*
University of Southern California
haipengl@usc.edu

Mengxiao Zhang*
University of Iowa
mengxiao-zhang@uiowa.edu

Abstract

Last-iterate convergence of learning dynamics in games has attracted significant recent attention. In two-player zero-sum games with bandit feedback, where only the loss of the selected action pair is observed, Fiegel et al. [2025] show a separation between average-iterate and last-iterate convergence in duality gap: while the optimal $t^{-\frac{1}{2}}$ rate after t rounds is achievable for the former via standard no-regret algorithms, the latter cannot converge faster than $t^{-\frac{1}{3}}$ in expectation or $t^{-\frac{1}{4}}$ with high probability. However, in many practical settings (such as preference learning), the players observe not only their loss but also the opponent's action. This raises a natural question: can such additional information enable faster last-iterate convergence? We answer this question affirmatively, showing that $t^{-\frac{1}{2}}$ last-iterate convergence is achievable with high probability in this setting, via an efficient algorithm that updates its strategy infrequently by solving an estimated log-barrier-regularized game. We identify fundamental obstacles preventing standard analysis for multi-armed bandits (the single-player case) from generalizing to games, and develop a novel analysis to overcome them. Experiments confirm that our algorithm indeed converges faster than naive baselines and prior methods that do not exploit opponent-action feedback. Finally, we note that our results also improve those for dueling bandits, a special case with skew-symmetric game matrices.

1 Introduction

Understanding learning dynamics in games is an important area in game theory and machine learning. In particular, last-iterate convergence to Nash equilibria of two-player zero-sum games has received increasing attention since it does not require averaging iterates (which could be cumbersome for neural nets) and provides strong day-to-day behavior guarantees. There has been a long line of research on this topic, much of which focuses on an idealized setting where each player obtains gradient feedback of their strategies. In this case, it has been shown that t^{-1} last-iterate convergence in duality gap is achievable after playing the games for t rounds; see e.g. Diakonikolas [2020], Yoon and Ryu [2021], Cai and Zheng [2023a,b], Cai et al. [2024, 2025].

However, gradient feedback might be too much to ask for in many applications. Often, players can see a (noisy) loss/reward of their selected action, but not the gradient of their mixed strategy. This is referred to as the bandit feedback setting and can be viewed as a direct generalization of stochastic multi-armed bandits (that is, single-player games) to multi-player games. Under this more

*Authors are listed in alphabetical order.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

challenging model, even for the weaker notion of average-iterate convergence, the best rate one can hope for is $t^{-\frac{1}{2}}$ due to standard statistical barriers, and any bandit algorithms with standard $\sqrt{t}$-regret achieve such average-iterate convergence [Freund and Schapire, 1999].

Achieving last-iterate convergence with bandit feedback, however, is even more challenging due to the severer tension between exploration and exploitation. Earlier results establish slow rates such as $t^{-\frac{1}{8}}$ [Cai et al., 2023, Chen et al., 2023, 2024], $t^{-\frac{1}{6}}$ [Cai et al., 2023], and $t^{-\frac{1}{5}}$ [Cai et al., 2025]. More recently, Fiegel et al. [2025, 2026] show that with bandit feedback, there is indeed a strict separation between average-iterate convergence, for which the optimal rate is $t^{-\frac{1}{2}}$, and last-iterate convergence, for which the optimal in-expectation rate is $t^{-\frac{1}{3}}$ and the optimal high-probability rate is $t^{-\frac{1}{4}}$.

While bandit feedback is more realistic than gradient feedback, in many situations, in addition to the losses, the players also observe their opponents' actions. For example, in pricing, companies can naturally observe their competitors' price; in online advertising auctions, bidders can observe the winning bid; and in security games, a defender may observe which targets an attacker chose to probe. There are also games where a central controller decides the actions for all players, in which case the action information is also directly available — a salient example is dueling bandits [Yue et al., 2012] or more generally learning from human preference feedback, often modeled as learning over a skew-symmetric preference game matrix where the algorithm decides which pair of actions to recommend to the user; see e.g., Munos et al. [2024], Swamy et al. [2024]. Such additional feedback intuitively facilitates estimation and mitigates the exploration burden, leading to a central question that our work focuses on: *does observing the opponent's actions enable faster last-iterate convergence in two-player zero-sum games?*

Contribution We answer this question affirmatively and develop an algorithm with $t^{-\frac{1}{2}}$ high-probability last-iterate convergence, which, up to dimension dependence, is optimal and closes the gap between average-iterate and last-iterate convergence. Our algorithm is computationally efficient and only updates its strategy $\log t$ times after t rounds, where each update solves an estimated log-barrier-regularized game.

Our algorithm shares similarity with the FALCON algorithm of Simchi-Levi and Xu [2022] for contextual bandits and can be viewed as a generalization of their algorithm to two-player zero-sum games (but specialized to the non-contextual setting). However, we highlight that our analysis significantly deviates from theirs and is the main technical contribution of our work. Indeed, we provide extensive discussion in Section 3 on why even with the extra information of opponent actions, achieving $t^{-\frac{1}{2}}$ last-iterate convergence is still highly challenging and existing analysis in the bandit literature does not generalize to games. This is true even when the game has a pure-strategy Nash equilibrium. At a high level, the key obstacle is that estimating the duality gap of a strategy requires playing other (potentially bad) strategies. Nevertheless, we overcome this obstacle via a novel analysis that relies on proving *aggregate multiplicative stability* between two consecutive strategies and might be of independent interest even for the single-player case. We also highlight that since dueling bandits are a special case of our setting, our results also strengthen this line of work by providing last-iterate convergence guarantees rather than only regret bounds.

Finally, we conduct experiments on the security game benchmark of Krever et al. [2025], showing that our algorithm consistently outperforms both existing methods that ignore opponent-action feedback and naive baselines that exploit it.

Related work There is a vast literature on learning in games. We refer the readers to the related work section of Cai et al. [2025] and also a sequence of recent advances [Tsuchiya, 2026, Liu et al., 2026, Abbadi et al., 2026, Mai, 2026, Ha, 2026b, Soleymani and Piliouras, 2026, Ha, 2026a] for prior work under the gradient feedback model and only focus on related work under the bandit feedback model here.

Regret and average/random-iterate convergence The seminal result of Freund and Schapire [1999] shows that average-iterate convergence rate in two-player zero-sum games is controlled by the social regret of the players. In particular, standard multi-armed bandit algorithms [Auer et al., 2002a, Zimmert and Seldin, 2021] with optimal $\sqrt{t}$-regret ensure $t^{-\frac{1}{2}}$ average-iterate convergence. To go beyond the $t^{-\frac{1}{2}}$ barrier, recent work either considers weaker regret notions or develop instance-dependent regret bounds [Maiti et al., 2025, Ito et al., 2025, 2026]. Sessa et al. [2019], O'Donoghue

et al. [2021], Ito et al. [2026] also study the effect of opponent-action information on regret, and we refer to Ito et al. [2026] for more discussion related to the line of research on Markov games where the opponent-action information also plays an important role.

On the other hand, social dynamic regret, which is equivalent to the cumulative duality gap, controls the random-iterate convergence of the learning dynamic. The algorithms of Xie et al. [2020], Bai and Jin [2020] based on the well-known optimism principle enjoy $\sqrt{t}$ social dynamic regret (with opponent-action information) and thus $t^{-\frac{1}{2}}$ random-iterate convergence. Note that our $t^{-\frac{1}{2}}$ last-iterate convergence trivially implies the same for random-iterate convergence, but not the other way around. In fact, it is known that such optimistic algorithms provably fail to achieve last-iterate convergence even for multi-armed bandits [Liu et al., 2024].

Last-iterate convergence As mentioned, a line of recent work studies last-iterate convergence without opponent-action feedback, improving the results from $t^{-\frac{1}{8}}$ down to the optimal in-expectation rate $t^{-\frac{1}{3}}$ and the optimal high-probability rate $t^{-\frac{1}{4}}$ [Cai et al., 2023, Chen et al., 2023, 2024, Cai et al., 2025, Fiegel et al., 2025, 2026]. Maiti et al. [2026] achieve similar $t^{-\frac{1}{4}}$ rate but require seeing the loss of the mixed strategies instead of the loss of the selected actions. Another recent work by Ito et al. [2025] establishes $t^{-\frac{1}{2}}$ last-iterate convergence for games with a unique pure strategy Nash equilibrium, but the bound involves game-dependent constants that can be arbitrarily large. To our knowledge, no prior work has studied the benefits of opponent-action information for last-iterate convergence.

Connection to bandits When one player has only one feasible action, our problem becomes a single-player game and recovers the classic stochastic multi-armed bandit problem [Lai and Robbins, 1985]. Similarly, most work in this area focuses on developing regret guarantees. To our knowledge, Liu et al. [2024] are the first to explicitly study last-iterate convergence, although an earlier result by Simchi-Levi and Xu [2022] also achieves last-iterate convergence implicitly for the more general contextual bandit problem. We will discuss in depth why these approaches do not generalize to game in Section 3. Similar to Ito et al. [2025], Zhan et al. [2025] develop last-iterate convergence rate (for multi-armed bandits) that involves potentially large instance-dependent constants.

2 Preliminaries

General notations For a positive integer k, let $[k]$ denote the set $\{1, 2, \ldots, k\}$. Define $\mathbb{R}^d_+$ to be the positive orthant of the d-dimensional Euclidean space, and $\Delta_d \triangleq \{x \in \mathbb{R}^d_+, \sum_{i=1}^d x_i = 1\}$ to be the $(d-1)$-dimensional simplex. For a vector $v \in \mathbb{R}^d$, we denote by v_i its i-th entry. Similarly, for a matrix $M \in \mathbb{R}^{m \times n}$, we denote by M_{ij} its (i,j)-th entry. Let e_i be the one-hot vector in an appropriate dimension with the i-th entry being 1 and all other entries being 0. Let $\mathbf{1}$ denote the all-one vector in an appropriate dimension.

Two-player zero-sum games We study two-player zero-sum games specified by a game matrix $A \in [-1, 1]^{d \times d}$, where A_{ij} denotes the expected loss incurred by the row player and, equivalently, the expected reward obtained by the column player, when the row player selects action i and the column player selects action j.[2] For a mixed strategy pair $(x, y) \in \Delta_d \times \Delta_d$, the expected loss of the row player is given by $x^\top A y$ (which is also the expected reward of the column player), and the duality gap of this strategy pair is defined as

$$\mathrm{DGap}(x, y) \triangleq \max_{y' \in \Delta_d} x^\top A y' - \min_{x' \in \Delta_d} x'^\top A y = \max_{i,j \in [d]} x^\top A e_j - e_i^\top A y,$$

which is always non-negative by definition. It is well-known that a strategy pair (x, y) has 0 duality gap if and only if it is a Nash equilibrium of the game (that is, both players are best-responding to each other: $x^\top A y' \le x^\top A y \le x'^\top A y$ for all $x', y' \in \Delta_d$). Therefore, duality gap is a natural measure for convergence rate to Nash equilibria.

While Nash equilibria always exist, a pure-strategy Nash equilibrium (PSNE) of the form (e_i, e_j) for some $i, j \in [d]$ might not. In Section 3, we will discuss why our problem is challenging even

[2] For the ease of presentation, we assume that both players have the same number of actions, but our results directly generalize to the case where they have different number of actions.

when the game has a unique PSNE, but we emphasize that our results hold generally, without any assumption on the games (other than boundedness).

Learning via repeated play with bandit feedback and opponent actions We consider a standard repeated play setting in which both players have no prior knowledge of the game matrix A. In each round $t = 1, 2, \dots$, the row player selects a mixed strategy $x_t \in \Delta_d$ and samples an action $i_t \sim x_t$; similarly, the column player selects a mixed strategy $y_t \in \Delta_d$ and samples an action $j_t \sim y_t$. Following that, both players observe bandit feedback, that is, a noisy loss/reward signal $r_t \in [-1, 1]$ satisfying $\mathbb{E}[r_t] = A_{i_t j_t}$. In addition, row player observes the opponent's action j_t and similarly the column player observes i_t.

An algorithm satisfies last-iterate convergence if it ensures $\text{DGap}(x_t, y_t) \to 0$ as t increases, which is stronger than other heavily-studied measures such as average-iterate convergence: $\text{DGap}\left(\frac{1}{t}\sum_{\tau=1}^{t} x_\tau, \frac{1}{t}\sum_{\tau=1}^{t} y_\tau\right) \to 0$, random-iterate convergence: $\frac{1}{t}\sum_{\tau=1}^{t} \text{DGap}(x_\tau, y_\tau) \to 0$, and best-iterate convergence: $\min_{\tau \le t} \text{DGap}(x_\tau, y_\tau) \to 0$. As discussed, the focus of our work is to show how the additional opponent-action information helps speed up last-iterate convergence beyond the $t^{-\frac{1}{3}}$ (in-expectation) or $t^{-\frac{1}{4}}$ (high-probability) rate that are optimal in the absence of opponent-action feedback, a question that has not been explored in the literature to our knowledge.

We also recall the real-life examples in Section 1 that illustrate the ubiquitousness of opponent-action information in practice. In particular, we point out again that dueling bandits are a special case of our setting where the game matrix A (known as the preference matrix in this case) is skew-symmetric ($A = -A^\top$), and the algorithm needs to select a pair of actions (i_t, j_t) to solicit preference feedback from the user. In this case, the information (i_t, j_t) is obviously known to the algorithm itself.

Finally, since we will use multi-armed bandits as an illustrative special case in many subsequent discussions, we remind the reader that when the game matrix A is d by 1, our setting recovers the standard stochastic d-armed bandit problem. Since the game matrix is just a vector in this case, we will denote it by $\ell \in [-1, 1]^d$ instead. It is clear that the duality gap of a strategy $x \in \Delta_d$ in this case reduces to the (instantaneous) regret of this strategy compared to the best action, which we denote as $\text{Reg}(x) = \langle x, \ell \rangle - \min_{i \in [d]} \ell_i$.

3 The Challenges: Why Existing Approaches Fail

Before introducing our solution, this section provides an in-depth discussion on the difficulty of the problem and why existing approaches in the bandit literature do not generalize to games, thereby highlighting both the need for a fundamentally different approach and the novelty of our work.

First and foremost, we recall the difficulty of achieving last-iterate convergence compared to other goals such as achieving low regret, even in the special case of multi-armed bandits. Indeed, while it is well-known that a key challenge for any learning problems with partial-information feedback is the exploitation-exploration trade-off, ensuring last-iterate convergence creates even stronger tension between them. This is because last-iterate convergence requires that in *every* round, a reasonably good strategy is played, while achieving low regret only requires that in *most* rounds, a good strategy is played. For example, while optimism-based approaches, such as the classic Upper Confidence Bound (UCB) algorithm [Auer et al., 2002b], ensures optimal regret in many bandit problems, the fact that it occasionally selects bad actions for exploration renders them incompatible with last-iterate convergence, as formally proven in Liu et al. [2024].

Naively, to play a good strategy in every round while ensuring sufficient exploration, one can mix the empirically best strategy with a uniform exploration strategy. However, it is well-known that such a naive exploitation-exploration trade-off fails to achieve $t^{-\frac{1}{2}}$ convergence even for bandits. Indeed, in Appendix C.1, we prove that for our problem, this naive strategy only achieves $\text{DGap}(x_t, y_t) = \mathcal{O}(d^{\frac{1}{2}} t^{-\frac{1}{4}})$ with high probability. We will use this strategy as one of the baselines in our experiments in Section 5.

So how do we achieve $t^{-\frac{1}{2}}$ convergence even in bandits? To our knowledge, there are essentially two types of approaches in the literature: hard-elimination and soft-elimination. We review both approaches and discuss why they do not generalize to games.

Algorithm 1: FALCON for Stochastic Multi-Armed Bandits

Input: Learning rates $\{\gamma_s\}_{s\geq 1}$.
Initialize loss estimator $\widehat{\ell}_1$ to be the all-zero vector.
for *epoch* $s = 1, 2, \ldots$ **do**
  Compute sampling strategy: $x_s = \arg\min_{x\in\Delta_d} \langle x, \widehat{\ell}_s\rangle + \gamma_s \sum_{i=1}^d \log \frac{1}{x_i}$.
  Initialize counters: $n_{s,i} = 0$ for all $i \in [d]$.
  for *round* $t = 2^{s-1}, \ldots, 2^s - 1$ **do**
    Sample $i_t \sim x_s$ and observe realized loss r_t such that $\mathbb{E}[r_t] = \ell_{i_t}$.
    Increment counter $n_{s,i_t} \leftarrow n_{s,i_t} + 1$.
  Compute loss estimate $\widehat{\ell}_{s+1,i} = \frac{\mathbb{1}\{n_{s,i}\neq 0\}}{n_{s,i}} \sum_{\tau=2^{s-1}}^{2^s-1} \mathbb{1}\{i_\tau = i\} \cdot r_\tau$ for all $i \in [d]$.

Hard-elimination Liu et al. [2024] study different variants of the hard-elimination approach from the literature, but they all share the same idea: eliminate actions that, based on current observations, cannot be the best action (with high probability), and (almost) uniformly explore the remaining actions. Such uniform exploration among plausibly good actions makes sure that in each round t, every action i that is not eliminated yet has regret $\mathrm{Reg}(e_i)$ less than $\widetilde{\mathcal{O}}(\sqrt{d/t})$, meaning that it is always "safe" to play any of them.

When trying to generalize hard-elimination to games, the first challenge one immediately meets is that we are no longer searching for the best strategy over a finite space; indeed, while in bandits the best strategy can always be in the form of e_i for some i, it is entirely possible that a game has no PSNE of the form (e_i, e_j) for some i and j. Therefore, the search space in game becomes $\Delta_d \times \Delta_d$ instead of $[d] \times [d]$.

However, we emphasize that *even when a PSNE exists, hard-elimination over the space* $[d]\times[d]$ *is still a bad idea.* This is due to the following key obstacle: estimating $\mathrm{DGap}(e_i, e_j) = \max_{h,k\in[d]} A_{ih} - A_{kj}$ for some pure strategy (e_i, e_j) cannot be done by simply playing (e_i, e_j) itself since this only yields samples of A_{ij}; instead, by definition, it requires playing (e_i, e_h) and (e_k, e_j) where e_h and e_k are the best response to e_i and e_j respectively that realize the maximum in the duality gap definition. Importantly, *these strategies* (e_i, e_h) *and* (e_k, e_j) *themselves can be bad strategies with a large duality gap!* To see this, simply consider a 3 by 3 example $A = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & -\varepsilon \\ 0 & \varepsilon & 0 \end{bmatrix}$ for some small and unknown $\varepsilon > 0$. The strategies (e_2, e_3) and (e_3, e_2) have a large duality gap of $1 + \varepsilon$, but if we eliminate them early, then there is no way to get samples from A_{23} or A_{32} to determine the value of ε and thus no way to estimate the duality gap of (e_3, e_3) (which is 2ε) and distinguish it from the true PSNE (e_1, e_1). This is in sharp contrast to the bandit case where playing an action naturally yields samples that helps estimate the quality of this action itself.

Soft-elimination We use the term soft-elimination to refer to approaches that never completely eliminate any actions, but instead adjust the probability of selecting an action based on its estimated quality. While there is no explicit study of this approach for achieving $t^{-\frac{1}{2}}$ last-iterate convergence in multi-armed bandits, the idea is in fact implicitly described in several works on contextual bandits, starting from Agarwal et al. [2014]. To facilitate discussions, we describe an algorithm based on this idea in Algorithm 1, which is most closely related to the FALCON algorithm of Simchi-Levi and Xu [2022] for contextual bandits and can be seen as a variant of it specified to the non-contextual setting.

Specifically, Algorithm 1 proceeds in epochs with doubling length. Within each epoch s, a sampling strategy $x_s \in \Delta_d$ for this epoch is computed via $\mathrm{argmin}_{x\in\Delta_d} \langle x, \widehat{\ell}_s\rangle + \gamma_s \sum_{i=1}^d \log \frac{1}{x_i}$, where $\widehat{\ell}_s$ is a loss estimator computed by simply averaging data from the last epoch, $\gamma_s > 0$ is a learning rate, and the part $\sum_{i=1}^d \log \frac{1}{x_i}$ is often known as the log-barrier regularizer, included to encourage exploration (since it explodes when x concentrates on a subset of actions). In other words, we solve an estimated log-barrier-regularized bandit problem in each epoch. With this strategy x_s, we simply sample i_t from x_s for all rounds $t = 2^{s-1}, \ldots, 2^s - 1$ within this epoch, and finally use the observed data to compute the next loss estimate $\widehat{\ell}_{s+1}$.

Other than specifying the FALCON algorithm to the non-contextual setting and instantiating its regression oracle with the simple empirical mean estimators, the only slight difference between Algorithm 1 and FALCON is how x_s is defined, but it is known that our definition shares the same properties as theirs; see e.g. Foster and Rakhlin [2020], Foster et al. [2020], Zhang and Luo [2024]. Our definition is also closely related to the well-known Follow-the-Regularized-Leader (FTRL) framework, and using log-barrier regularizer in this framework for bandits has been extensively studied since the early work such as Foster et al. [2016], Wei and Luo [2018]. However, an important distinction is that these FTRL algorithms all use importance-weighted estimators instead of the empirical mean estimators that Algorithm 1 uses.

The key idea in the analysis for such soft-elimination methods is to show that, with an appropriate choice of γ_s (that is of order $1/\sqrt{d2^s}$), for each action i, its regret $\mathrm{Reg}(e_i)$ is close to the estimated version $\widehat{\mathrm{Reg}}_s(e_i)$, where $\widehat{\mathrm{Reg}}_s(x) = \langle x, \widehat{\ell}_s\rangle - \min_{i\in[d]} \widehat{\ell}_{s,i}$, in a multiplicative sense: $\mathrm{Reg}(e_i) \le 2\widehat{\mathrm{Reg}}_s(e_i) + \widetilde{\mathcal{O}}(d\gamma_s)$ and similarly $\widehat{\mathrm{Reg}}_s(e_i) \le 2\mathrm{Reg}(e_i) + \widetilde{\mathcal{O}}(d\gamma_s)$. This is enabled by the following key low-regret-low-variance property of the algorithm that underscores the essence of a line of previous work on contextual bandits; see e.g., Agarwal et al. [2014, Algorithm 1], Simchi-Levi and Xu [2022, Observation 2], and Zhang and Luo [2024, Lemma 6].

Lemma 3.1 (Low-Regret-Low-Variance). *For every epoch $s \ge 1$, Algorithm 1 satisfies the following:*

$$\widehat{\mathrm{Reg}}_s(x_s) \le d\gamma_s, \tag{1}$$

$$\frac{1}{x_{s,i}} \le d + \frac{\widehat{\mathrm{Reg}}_s(e_i)}{\gamma_s}, \ \forall\, i \in [d]. \tag{2}$$

Note that $x_{s,i}$ is the probability of getting a sample for action i in each round of epoch s and thus $1/x_{s,i}$ controls the variance of the loss estimate for this action. Eq. (2) therefore states that the variance for action i is controlled by its estimated regret: the better the action appears to be, the lower the estimation variance is, which is exactly the key to show the multiplicative closeness between $\mathrm{Reg}(e_i)$ and $\widehat{\mathrm{Reg}}_s(e_i)$. Such a low-variance property holds thanks to the exploration induced by the special log-barrier regularization. On the other hand, due to such closeness, the low-estimated-regret property of Eq. (1) thus immediately implies that the true per-round regret of the algorithm is also small, hence achieving $\widetilde{\mathcal{O}}(d\gamma_s) = \widetilde{\mathcal{O}}(\sqrt{d/t})$ last-iterate convergence [Simchi-Levi and Xu, 2022, Lemma 9].

To understand the difficulty of generalizing such soft-elimination methods to games, let us consider again the simpler case with a PSNE. The most obvious generalization would be to try to ensure multiplicative closeness between $\mathrm{DGap}(e_i, e_j)$ and its estimated version for any action pair (e_i, e_j). This is impossible due to the exactly same obstacle mentioned before: estimating $\mathrm{DGap}(e_i, e_j)$ requires playing other potentially bad strategies. Indeed, Saha and Krishnamurthy [2022] already use the same aforementioned 3 by 3 example to show that if such multiplicative closeness holds, then the algorithm cannot have $\sqrt{t}$ regret, let alone $t^{-\frac{1}{2}}$ last-iterate convergence. In Appendix C.2, we provide more discussion on why a low-duality-gap-low-variance property similar to Lemma 3.1, which our final algorithm *does* ensure, fails to show $t^{-\frac{1}{2}}$ last-iterate convergence.

We point out that Saha and Schapire [2025] make an attempt to get around this obstacle in the contextual dueling bandit setting by imposing a strong "Idempotent Best-Response" assumption. When specified to our setting, this would mean that their results only hold for a special subset of games with a PSNE. Also, their algorithm requires performing least squares regression over this class of games, which is unlikely to be computationally efficient.

4 Our Algorithm and Analysis

Given all the obstacles discussed above, it might seem that all hope is lost. Nevertheless, we propose a solution that algorithmically follows the soft-elimination approach but analytically requires a different and novel analysis. Specifically, our algorithm, called Phased Minimax Optimization with Log-Barrier Regularization (PMO-LB), is shown in Algorithm 2. It follows Algorithm 1 with the same epoch schedule and empirical mean estimator $\widehat{A}_{s,ij}$ for each (i, j) entry of the unknown matrix A, and generalizes it in how the sampling strategy is computed: (x_s, y_s) is the Nash equilibrium of the estimated log-barrier-regularized game specified by Φ_s in Eq. (3). Equivalently,

Algorithm 2: Phased Minimax Optimization with Log-Barrier Regularization (PMO-LB)

Input: Learning rate $\{\gamma_s\}_{s\geq 1}$.
Initialize game estimator $\widehat{A}_1$ to be the all-zero d by d matrix.
for $s = 1, 2, \ldots$ **do**
Define an estimated log-barrier-regularized game with the following payoff function:

$$\Phi_s(x, y) \triangleq x^\top \widehat{A}_s y + \gamma_s \sum_{i=1}^{d} \log \frac{1}{x_i} - \gamma_s \sum_{j=1}^{d} \log \frac{1}{y_j}. \tag{3}$$

For row player: compute $x_s = \operatorname{argmin}_{x\in\Delta_d} \max_{y\in\Delta_d} \Phi_s(x, y)$.
For column player: compute $y_s = \operatorname{argmax}_{y\in\Delta_d} \min_{x\in\Delta_d} \Phi_s(x, y)$.
Initialize counters: $n_{s,ij} = 0$ for all $i, j \in [d]$.
for $t = 2^{s-1}, \ldots, 2^s - 1$ **do**
Row player samples $i_t \sim x_s$ and column player samples $j_t \sim y_s$.
Both players observe (i_t, j_t) and r_t with $\mathbb{E}[r_t] = A_{i_t j_t}$.
Increment counter $n_{s,i_t j_t} \leftarrow n_{s,i_t j_t} + 1$.
Compute game estimator $\widehat{A}_{s+1,ij} = \frac{\mathbb{1}\{n_{s,ij}\neq 0\}}{n_{s,ij}} \sum_{\tau=2^{s-1}}^{2^s-1} r_\tau \cdot \mathbb{1}\{i_\tau = i, j_\tau = j\}$ for all $i, j \in [d]$.

$x_s = \operatorname{argmin}_{x\in\Delta_d} \max_{y\in\Delta_d} \Phi_s(x, y)$ and $y_s = \operatorname{argmax}_{y\in\Delta_d} \min_{x\in\Delta_d} \Phi_s(x, y)$. Such a (unique) Nash equilibrium must exist since Φ_s is strongly convex in x and strongly concave in y, and can be computed efficiently using standard methods from the saddle point optimization literature. When the game matrix is d by 1, it is clear that our algorithm reduces to Algorithm 1. Before analyzing our algorithm, we make the following important remarks.

Uncoupled/coupled dynamics While we present both players' algorithm together in Algorithm 2, it is clear that these two players can compute their own strategy *independently without any extra coordination/coupling/shared randomness*. Indeed, thanks to the opponent-action information, each player can construct the estimator $\widehat{A}_s$ and compute their strategy on their own. Whether this is considered as an uncoupled or coupled learning dynamic depends on how one define them; for example, Fiegel et al. [2025] define uncoupled learning dynamics as those that do not have the opponent-action information.

Connection to existing algorithms Other than being a generalization of Algorithm 1 (which itself is a variant of FALCON [Simchi-Levi and Xu, 2022]), our algorithm is also connected to several others in the literature. For example, Saha and Schapire [2025] also generalize the FALCON algorithm to contextual dueling bandits. Their algorithm shares similarities with ours but importantly computes a *joint* distribution over action pairs and thus requires coupling when sampling actions. Moreover, as mentioned in Section 3, their results hold only under a strong assumption, which is inevitable since their analysis still follows the same low-regret-low-variance approach discussed in Section 3, proven to fail generally according to Saha and Krishnamurthy [2022]. On the other hand, our algorithm also shares similarity with Cardoso et al. [2019, Algorithm 3], with the following differences: first, they use importance-weighted estimators instead of empirical mean estimators; second, they use entropy regularizer instead of the log-barrier regularizer (the latter is critical for our analysis as it will become clear); finally, their algorithm is designed for a quite different problem with time-varying game matrices, and the metric they study is a notion called NE regret instead of last-iterate convergence. Finally, we note that the algorithm of Fiegel et al. [2026] for the setting without opponent-action information also uses log-barrier regularization but in the OMD framework and does not involve minimax optimization. Their algorithm is closer to that of Cai et al. [2023] with the difference being the form of the regularizer.

4.1 Main Results and Analysis

Our main result is the following theorem.

Theorem 4.1. *For any fixed $\delta \in (0,1)$, Algorithm 2 with $\gamma_s = 80\sqrt{d} \cdot 2^{-s/2}\sqrt{\log(8d^2s^2/\delta)}$ for all $s \geq 1$ guarantees that with probability at least $1-\delta$, $\mathrm{DGap}(x_t, y_t) = \mathcal{O}\left(d^{3/2}\sqrt{\frac{\log(d\log t/\delta)}{t}}\right)$ for all $t \geq 1$, where (x_t, y_t) denotes the strategy pair played in round t.*

As mentioned, the best existing results prior to our work are those that do not utilize the opponent-action information: $t^{-\frac{1}{3}}$ (in-expectation) from Fiegel et al. [2025] and $t^{-\frac{1}{4}}$ (high-probability) from Fiegel et al. [2026]. Our work is the first to show that with opponent-action feedback, $t^{-\frac{1}{2}}$ last-iterate convergence is achievable, eliminating the gap between last-iterate and average-iterate convergence, at least in terms of the dependence on t. As for the dependence on d, the standard $\Omega(\sqrt{d/t})$ lower bound for average-iterate convergence follows from the bandit literature [Auer et al., 2002a], and our $d^{3/2}$ dependence is suboptimal with respect to that.

Despite our algorithm being a direct generalization of Algorithm 1, as mentioned, our analysis is (and has to be) different to overcome the obstacles discussed in Section 3. To illustrate our core ideas, we discuss the single-player case again as a warm-up, which can be seen as an alternative analysis of Algorithm 1. (All omitted proof can be found in the appendix).

Warm-up for the single-player case Our starting point is the following property that is strictly stronger than the low-regret-low-variance property of Lemma 3.1.

Lemma 4.2. *For every epoch $s \geq 1$, Algorithm 1 satisfies*

$$\langle x_s, \widehat{\ell}_s\rangle - \widehat{\ell}_{s,i} \leq d\gamma_s - \frac{\gamma_s}{x_{s,i}}, \ \forall\, i \in [d]. \tag{4}$$

This is obtained by simply applying the first-order optimality condition to the optimization objective that defines x_s. It is stronger than Lemma 3.1 since it trivially implies Eq. (1) by dropping the negative term $-\frac{\gamma_s}{x_{s,i}}$, and it also implies Eq. (2) by rearranging terms and realizing $\widehat{\ell}_{s,i} - \langle x_s, \widehat{\ell}_s\rangle \leq \widehat{\ell}_{s,i} - \min_{j\in[d]}\widehat{\ell}_{s,j} = \widehat{\mathrm{Reg}}_s(e_i)$. This new lemma is useful since it asserts that the algorithm's estimated regret compared against a particular action is much smaller when this action is selected with a low probability, which in turn compensates the large estimation variance for this action. Indeed, consider bounding the true regret of x_s as follows (with $i^\star \in \operatorname{argmin}_{i\in[d]}\ell_i$ being an optimal action):

$$\mathrm{Reg}(x_s) = \langle x_s, \widehat{\ell}_s\rangle - \widehat{\ell}_{s,i^\star} + \langle x_s - e_{i^\star}, \ell - \widehat{\ell}_s\rangle \leq d\gamma_s - \frac{\gamma_s}{x_{s,i^\star}} + \frac{\beta_s}{\sqrt{x_{s-1,i^\star}}} + \sum_{i=1}^{d}\frac{\beta_s x_{s,i}}{\sqrt{x_{s-1,i}}},$$

where the inequality uses Lemma 4.2 and standard concentration to bound $|\widehat{\ell}_{s,i} - \ell_i|$ for each i by $\frac{\beta_s}{\sqrt{x_{s-1,i}}}$ with $\beta_s = \widetilde{\mathcal{O}}(1/\sqrt{2^s})$, since $\widehat{\ell}_{s,i}$ is constructed using samples collected from strategy x_{s-1}. At this point, it is clear that if x_s and x_{s-1} are close enough in a multiplicative sense, then the negative term from Eq. (4) can be used to compensate for the potentially large estimation error $\frac{\beta_s}{\sqrt{x_{s-1,i^\star}}}$. To this end, we introduce the *aggregate multiplicative stability* property which requires that $\sum_{i=1}^{d}\left(\frac{x_{s,i}}{x_{s-1,i}} + \frac{x_{s-1,i}}{x_{s,i}}\right) \leq \alpha$ holds for some $\alpha = \mathcal{O}(\mathrm{poly}(d))$. This implies $\frac{x_{s,i}}{x_{s-1,i}} \leq \alpha$ for all $i \in [d]$. Using these bounds and the Cauchy-Schwarz inequality, we have,

$$\mathrm{Reg}(x_s) \leq d\gamma_s - \frac{\gamma_s}{\alpha x_{s-1,i^\star}} + \frac{\beta_s}{\sqrt{x_{s-1,i^\star}}} + \beta_s\sqrt{\alpha} \leq d\gamma_s + \frac{\alpha\beta_s^2}{4\gamma_s} + \beta_s\sqrt{\alpha},$$

where the last step uses AM-GM inequality. This concludes the $t^{-\frac{1}{2}}$ last-iterate convergence after setting γ_s appropriately. It thus remains to prove aggregate multiplicative stability, which, thanks to the log-barrier regularization, indeed holds with $\alpha = \mathcal{O}(d)$; see Lemma A.5 in the appendix. Although it is known that log-barrier regularization ensures multiplicative stability in the FTRL framework, as first shown by Wei and Luo [2018] and later utilized in many other works, our proof is different and needs to resolve several new challenges, including handling the time-varying learning rate γ_s and understanding how $\widehat{\ell}_s$ deviates from $\widehat{\ell}_{s-1}$, the latter of which itself requires using multiplicative stability from previous epochs.

Formally, we show a $\widetilde{\mathcal{O}}(d/\sqrt{t})$ last-iterate convergence rate for Algorithm 1 using this new analysis (see Proposition A.1), slightly worse than the $\widetilde{\mathcal{O}}(\sqrt{d/t})$ rate obtained from the analysis discussed in Section 3. However, the benefit is that this new analysis immediately generalizes to games (unlike the one in Section 3 that provably cannot).

Generalization to games The generalization starts with the following analogue of Lemma 4.2.

Lemma 4.3. *For every epoch $s \geq 1$, Algorithm 2 satisfies:*

$$x_s^\top \widehat{A}_s e_j - e_i^\top \widehat{A}_s y_s \leq 2d\gamma_s - \frac{\gamma_s}{x_{s,i}} - \frac{\gamma_s}{y_{s,j}}, \ \forall i, j \in [d]. \tag{5}$$

This lemma shows that the lower the sampling probability of an action, the smaller the increase in payoff one can obtain by deviating unilaterally to this action. We note that Eq. (2) of Saha and Schapire [2025] is close to our Eq. (5), but theirs involves a joint distribution over all action pairs and is thus fundamentally different from ours. Next, we generalize the aggregate multiplicative stability from single player to two players as follows.

Lemma 4.4 (Aggregate Multiplicative stability)**.** *With probability at least $1-\delta$, Algorithm 2 with γ_s set to the value stated in Theorem 4.1 ensures for all $s \geq 2$, $\sum_{i=1}^d \left(\frac{x_{s,i}}{x_{s-1,i}} + \frac{x_{s-1,i}}{x_{s,i}}\right) + \sum_{j=1}^d \left(\frac{y_{s,j}}{y_{s-1,j}} + \frac{y_{s-1,j}}{y_{s,j}}\right) \leq 8d$.*

The proof of this lemma is the most technical part of our analysis. It not only needs to address all the issues mentioned earlier in the single-player case, but also needs to take into account the fact that x_s is the solution of an optimization defined in terms of y_s (and vice versa) and both of them are changing over time. With these two key lemmas, the proof of our main result Theorem 4.1 can be proven following similar ideas outlined earlier for the single-player case; see Appendix B for details.

5 Experiments

Experimental setup We evaluate the performance of PMO-LB (Algorithm 2) against a set of baselines on four security games from the benchmark of Krever et al. [2025]. We run the algorithms for 10^7 rounds for each of the two smaller games of size 39×35 and 61×21, and 10^8 rounds for each of the two larger games of size 221×210 and 520×495. The five baselines we compared against are: Cai et al. [2023, Algorithm 1], Cai et al. [2025, Algorithm 2], Fiegel et al. [2025, Algorithm 1], Fiegel et al. [2026, Algorithm 1], and the naive baseline (Algorithm 3) analyzed in Appendix C.1. Note that only our Algorithm 2 and Algorithm 3 utilize access to the opponent's actions, whereas the remaining algorithms do not. For more details, see Appendix D.

Results and discussion Figure 1 compares the duality gap as a function of the number of rounds across all algorithms, in log-log scale. We also include in the legend the slope (that is, the exponent of the convergence rate) of each plot as predicted by the theory and as fitted based on the actual results starting from $t = 10^6$ for Figure 1a and Figure 1b, and from $t = 10^7$ for Figure 1c and Figure 1d. For most algorithms, the predicted slope indeed aligns with the empirical one. More importantly, the figure shows that our PMO-LB algorithm consistently outperforms all other baselines.

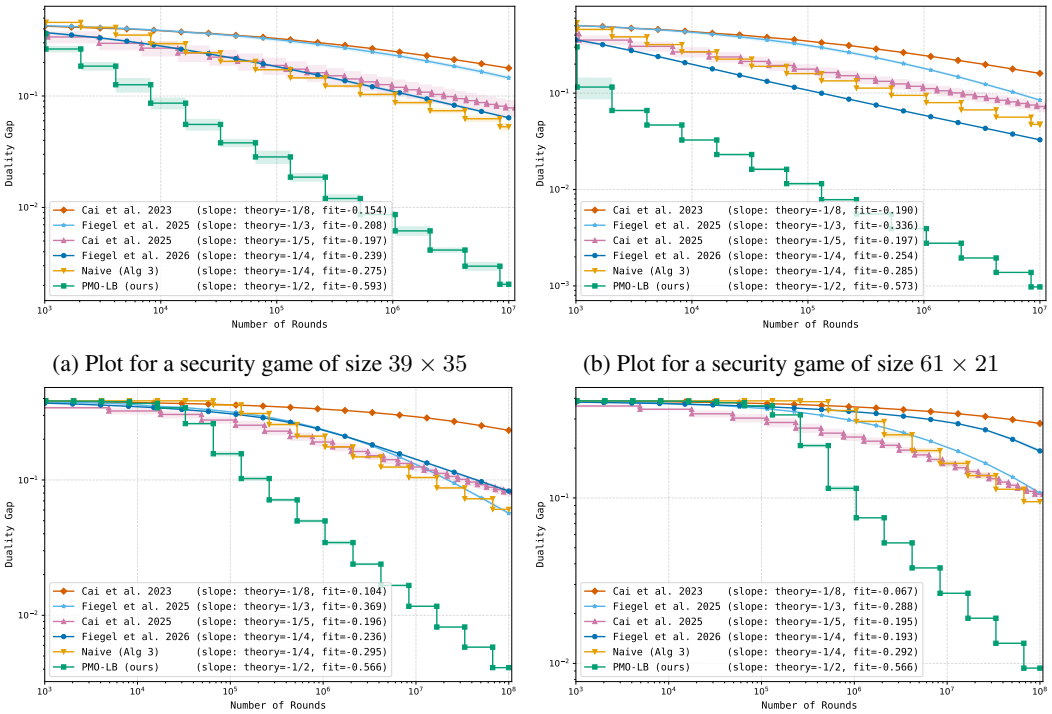
Acknowledgment HL and SH are supported by NSF award CCF-2551666.

References

Omar Abbadi, Rida Laraki, and Panayotis Mertikopoulos. Constant regret in general games via higher-order optimism. *arXiv preprint arXiv:2609.04113*, 2026.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International conference on machine learning*, pages 1638–1646. PMLR, 2014.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002a.



(a) Plot for a security game of size 39×35

(b) Plot for a security game of size 61×21

(c) Plot for a security game of size 221×210

(d) Plot for a security game of size 520×495

Figure 1: Plots of duality gap vs. number of rounds for PMO-LB and five baseline algorithms

Peter Auer, Nicolo Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002b.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pages 551–560. PMLR, 2020.

Yang Cai and Weiqiang Zheng. Doubly optimal no-regret learning in monotone games. In *International Conference on Machine Learning*, pages 3507–3524. PMLR, 2023a.

Yang Cai and Weiqiang Zheng. Accelerated Single-Call Methods for Constrained Min-Max Optimization. In *Proceedings of the 11th International Conference on Learning Representations*, 2023b.

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

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*, pages 5312–5347, 2024.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. From average-iterate to last-iterate convergence in games: A reduction and its applications. *Neural Information Processing Systems*, 2025.

Adrian Rivera Cardoso, Jacob Abernethy, He Wang, and Huan Xu. Competing against equilibria in zero-sum games with evolving payoffs. *International conference on machine learning*, 2019.

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. *Advances in Neural Information Processing Systems*, 36:75826–75883, 2023.

Zaiwei Chen, Kaiqing Zhang, Eric Mazumdar, Asuman Ozdaglar, and Adam Wierman. Last-iterate convergence of payoff-based independent learning in zero-sum stochastic games. *arXiv preprint arXiv:2409.01447*, 2024.

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.

Côme Fiegel, Pierre Menard, Tadashi Kozuno, Michal Valko, and Vianney Perchet. The harder path: Last iterate convergence for uncoupled learning in zero-sum games with bandit feedback. *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Come Fiegel, Pierre Menard, Tadashi Kozuno, Michal Valko, and Vianney Perchet. Optimal last-iterate convergence in matrix games with bandit feedback using the log-barrier. *arXiv preprint arXiv:2604.15242*, 2026.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International conference on machine learning*, pages 3199–3210. PMLR, 2020.

Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. *Advances in Neural Information Processing Systems*, 29, 2016.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33:11478–11489, 2020.

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

Junsoo Ha. A horizon-independent regret bound for optimistic hedge in general-sum games. *arXiv preprint arXiv:2609.22839*, 2026a.

Junsoo Ha. A logarithmic regret bound for optimistic hedge in general-sum games. *arXiv preprint arXiv:2609.19677*, 2026b.

Shinji Ito, Haipeng Luo, Taira Tsuchiya, and Yue Wu. Instance-dependent regret bounds for learning two-player zero-sum games with bandit feedback. *Conference on Learning Theory*, 2025.

Shinji Ito, Haipeng Luo, Arnab Maiti, Taira Tsuchiya, and Yue Wu. Adversarial learning in games with bandit feedback: Logarithmic pure-strategy maximin regret. *arXiv preprint arXiv:2602.06348*, 2026.

Noah Krever, Jakub Černỳ, Moïse Blanchard, and Christian Kroer. Guard: Constructing realistic two-player matrix and security games for benchmarking game-theoretic algorithms. *arXiv preprint arXiv:2505.14547*, 2025.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Junyan Liu, Yunfan Li, Ruosong Wang, and Lin F Yang. Uniform last-iterate guarantee for bandits and reinforcement learning. *Advances in Neural Information Processing Systems*, 37:133600–133660, 2024.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. Constant individual regret in general games. *arXiv preprint arXiv:2608.31166*, 2026.

Tung Mai. Constant swap regret in general-sum games via two-scale higher-order optimism. *arXiv preprint arXiv:2609.16751*, 2026.

Arnab Maiti, Kevin Jamieson, and Lillian J Ratliff. On the limitations and possibilities of nash regret minimization in zero-sum matrix games under noisy feedback. In *International Symposium on Algorithmic Game Theory*, pages 3–20. Springer, 2025.

Arnab Maiti, Claire Jie Zhang, Kevin Jamieson, Jamie Heather Morgenstern, Ioannis Panageas, and Lillian J Ratliff. Efficient uncoupled learning dynamics with $\widetilde{O}(t^{-1/4})$ last-iterate convergence in bilinear saddle-point problems over convex sets under bandit feedback. *Twenty-Ninth Annual Conference on Artificial Intelligence and Statistics*, 2026.

Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, et al. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2024.

Brendan O'Donoghue, Tor Lattimore, and Ian Osband. Matrix games with bandit feedback. In *Uncertainty in Artificial Intelligence*, pages 279–289. PMLR, 2021.

Aadirupa Saha and Akshay Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *International Conference on Algorithmic Learning Theory*, pages 968–994. PMLR, 2022.

Aadirupa Saha and Robert E Schapire. Efficient and near-optimal algorithm for contextual dueling bandits with offline regression oracles. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, and Andreas Krause. No-regret learning in unknown games with correlated payoffs. *Advances in Neural Information Processing Systems*, 32, 2019.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.

Ashkan Soleymani and Georgios Piliouras. Multiplicative optimism for constant regret in games. *arXiv preprint arXiv:2609.21976*, 2026.

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.

Taira Tsuchiya. Sublogarithmic swap regret in multiplayer general-sum games via hybrid regularization. *arXiv preprint arXiv:2608.04149*, 2026.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR, 2020.

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)*, pages 12098–12109. PMLR, 2021.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Jingxin Zhan, Yuze Han, and Zhihua Zhang. Last iterate analyses of ftrl in stochasitc bandits. *arXiv preprint arXiv:2510.22819*, 2025.

Mengxiao Zhang and Haipeng Luo. Contextual multinomial logit bandits with general value functions. *Advances in Neural Information Processing Systems*, 37:34123–34160, 2024.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.

A Omitted Details for Multi-Armed Bandits

In this section, we analyze the single-agent setting from Section 4 and formally state the convergence guarantee for FALCON under our approach.

Proposition A.1 (Last-iterate convergence of FALCON)**.** *For any fixed $\delta \in (0,1)$, Algorithm 1 with $\gamma_s = 40 \cdot 2^{-s/2}\sqrt{\log(8ds^2/\delta)}$ guarantees that with probability at least $1-\delta$, $\mathrm{Reg}(x_t) = \mathcal{O}\left(d\sqrt{\frac{\log(d\log t/\delta)}{t}}\right)$ for all $t \geq 1$, where x_t denotes the strategy played in round t.*

We begin by showing a high-probability bound on the estimation error of each arm's loss under the epoch-wise sampling scheme of Algorithm 1.

Definition A.2. *Define $\mathcal{E}_1$ to be the event: $\mathcal{E}_1 \triangleq \left\{\forall s \geq 2, \forall i \in [d], |\widehat{\ell}_{s,i} - \ell_i| \leq \frac{\beta_s}{\sqrt{x_{s-1,i}}}\right\}$ where $\beta_s = \sqrt{\frac{16\log(8ds^2/\delta)}{2^{s-2}}}$.*

Lemma A.3 (Estimation error)**.** *Algorithm 1 guarantees that $\mathcal{E}_1$ holds with probability at least $1-\delta$.*

First, we state a standard concentration inequality that will be used in the proof of the estimation error bound.

Lemma A.4 (Multiplicative Chernoff bound)**.** *Let $X = \sum_{i=1}^n X_i$, where X_i is a Bernoulli random variable with mean p_i and all X_i are independent. Let $\mu = \mathbb{E}[X] = \sum_{i=1}^n p_i$. Then*

$$\mathsf{Pr}\left\{X \leq (1-\delta)\mu\right\} \leq e^{-\mu\delta^2/2}, \quad \forall\, \delta \in (0,1).$$

Proof of Lemma A.3. Fix $s \geq 2$. Conditional on the history prior to epoch $s-1$, the sampling distribution x_{s-1} is fixed. The length of epoch $s-1$ is 2^{s-2}, and hence the expected number of times arm $i \in [d]$ is selected in epoch $s-1$ is $2^{s-2}x_{s-1,i}$. We first show that for any $\xi_s \in (0,1)$, with probability at least $1-\xi_s$, the inequality $|\widehat{\ell}_{s,i} - \ell_i| \leq \sqrt{\frac{16\log(4d/\xi_s)}{2^{s-2}x_{s-1,i}}}$ holds simultaneously for all $i \in [d]$.

For any arm $i \in [d]$, if $2^{s-2}x_{s-1,i} \leq 16\log(4d/\xi_s)$, then $\sqrt{\frac{16\log(4d/\xi_s)}{2^{s-2}x_{s-1,i}}} \geq 1$. Since both $\widehat{\ell}_{s,i}$ and ℓ_i lie in $[0,1]$, the bound holds trivially.

Now suppose $2^{s-2}x_{s-1,i} > 16\log(4d/\xi_s)$. Applying the multiplicative Chernoff bound in Lemma A.4, we obtain

$$\mathsf{Pr}\left\{n_{s-1,i} < \frac{2^{s-2}x_{s-1,i}}{2}\right\} \leq e^{-\frac{2^{s-2}x_{s-1,i}}{8}} \leq e^{-2\log(4d/\xi_s)}.$$

Conditioned on $n_{s-1,i} = n \geq 2^{s-2}x_{s-1,i}/2$, the estimate $\widehat{\ell}_{s,i}$ is an average of n independent random variables in $[0,1]$ with mean ℓ_i. Thus, by Hoeffding's bound,

$$\begin{aligned}\mathsf{Pr}\left\{|\widehat{\ell}_{s,i} - \ell_i| > \sqrt{\frac{16\log(4d/\xi_s)}{2^{s-2}x_{s-1,i}}} \;\middle|\; n_{s-1,i} = n\right\} &\leq 2\exp\left(-2n\frac{16\log(4d/\xi_s)}{2^{s-2}x_{s-1,i}}\right)\\ &\leq 2e^{-16\log(4d/\xi_s)}.\end{aligned}$$

Combining the above bounds, for each fixed arm i,

$$\mathsf{Pr}\left\{|\widehat{\ell}_{s,i} - \ell_i| > \sqrt{\frac{16\log(4d/\xi_s)}{2^{s-2}x_{s-1,i}}}\right\} \leq e^{-2\log(4d/\xi_s)} + 2e^{-16\log(4d/\xi_s)} \leq \frac{\xi_s}{d}. \tag{6}$$

A union bound over all d arms gives us

$$\mathsf{Pr}\left\{\forall i \in [d], |\widehat{\ell}_{s,i} - \ell_i| \leq \sqrt{\frac{16\log(4d/\xi_s)}{2^{s-2}x_{s-1,i}}}\right\} \geq 1-\xi_s. \tag{7}$$

Setting $\xi_s = \frac{\delta}{2s^2}$ in Eq. (7) gives that $|\widehat{\ell}_{s,i} - \ell_i| \le \frac{\beta_s}{\sqrt{x_{s-1,i}}}$ holds for all $i \in [d]$ with probability at least $1 - \xi_s$.

Finally, applying a union bound over $s \ge 2$ yields a total failure probability of at most $\sum_{s=2}^{\infty} \frac{\delta}{2s^2} \le \delta$. Therefore, the event $\mathcal{E}_1$ defined in Definition A.2 holds with probability at least $1 - \delta$. □

Next, we prove Lemma 4.2, which is a stronger version of Lemma 3.1.

Proof of Lemma 4.2. The sampling strategy x_s is obtained by solving the convex optimization problem $F(x) \triangleq \min_{x \in \Delta_d} \langle x, \widehat{\ell}_s \rangle + \gamma_s \sum_{i=1}^{d} \log \frac{1}{x_i}$. With a slight abuse of notation, for $v \in \mathbb{R}^d$, we denote by $\frac{1}{v} \in \mathbb{R}^d$ the vector whose i-th coordinate is $\frac{1}{v_i}$. By first-order optimality, for any $x \in \Delta_d$,

$$\langle x - x_s, \nabla F(x_s) \rangle \ge 0 \implies \left\langle x - x_s, \widehat{\ell}_s - \frac{\gamma_s}{x_s} \right\rangle \ge 0.$$

Choosing $x = e_i$ gives us the desired inequality. □

We next state an aggregate multiplicative stability bound for successive strategies.

Lemma A.5 (Aggregate multiplicative stability)**.** *With probability at least $1 - \delta$, Algorithm 1 with γ_s set to the value stated in Proposition A.1 ensures for all $s \ge 2$, $Q_s \triangleq \sum_{i=1}^{d} \left(\frac{x_{s,i}}{x_{s-1,i}} + \frac{x_{s-1,i}}{x_{s,i}} \right) \le 4d$.*

First, we prove the more general stability result below, and Lemma A.5 follows from it by choosing γ_s appropriately, as shown in the proof of Proposition A.1.

Lemma A.6. *Suppose that $\mathcal{E}_1$ holds and define $\rho_s = \gamma_s / \gamma_{s+1}$. Further suppose that there exists $\alpha \ge 2d$ such that the sequence $\{\gamma_s\}_{s \ge 1}$ in Algorithm 1 satisfies (1) $Q_2 \le \alpha$, (2) for all $s \ge 2$, $2(\rho_s - 1)d + \frac{\rho_s^2 (\beta_s \sqrt{\alpha} + \beta_{s+1})^2}{\gamma_s^2} + 2d \le \alpha$, (3) $\rho_s \in [1, \sqrt{2}]$. Then Algorithm 1 guarantees that, for all $s \ge 2$, $Q_s \le \alpha$.*

Proof. We prove using induction on s. The base case $s = 2$ follows from the assumption that $Q_2 \le \alpha$. Assume that for some $s \ge 2$, $Q_s \le \alpha$. We show that this also holds for $s + 1$.

First, for all $s \ge 2$, we have

$$\sum_{i=1}^{d} \frac{(x_{s,i} - x_{s-1,i})^2}{x_{s,i} x_{s-1,i}} = Q_s - 2d, \tag{8}$$

where the left-hand side is always non-negative, implying that $Q_s \ge 2d$, and thus $\alpha \ge 2d$.

Under event $\mathcal{E}_1$, for every $i \in [d]$,

$$\begin{aligned} |\widehat{\ell}_{s,i} - \widehat{\ell}_{s+1,i}| &\le |\widehat{\ell}_{s,i} - \ell_i| + |\widehat{\ell}_{s+1,i} - \ell_i| \\ &\le \frac{\beta_s}{\sqrt{x_{s-1,i}}} + \frac{\beta_{s+1}}{\sqrt{x_{s,i}}}. \end{aligned} \tag{9}$$

By first-order optimality at epochs s and $s + 1$, for any $x \in \Delta_d$,

$$\left\langle \widehat{\ell}_s - \frac{\gamma_s}{x_s}, x - x_s \right\rangle \ge 0, \tag{10}$$

$$\left\langle \widehat{\ell}_{s+1} - \frac{\gamma_{s+1}}{x_{s+1}}, x - x_{s+1} \right\rangle \ge 0. \tag{11}$$

Setting $x = x_{s+1}$ in Eq. (10) and $x = x_s$ in Eq. (11), and then summing them and using Eq. (8), gives

$$Q_{s+1} - 2d \le \left\langle x_{s+1} - x_s, \frac{\widehat{\ell}_s}{\gamma_s} - \frac{\widehat{\ell}_{s+1}}{\gamma_{s+1}} \right\rangle.$$

We have $\rho_s = \gamma_s / \gamma_{s+1}$. Thus, we get

$$Q_{s+1} \le \frac{\rho_s}{\gamma_s} \left\langle x_{s+1} - x_s, \widehat{\ell}_s - \widehat{\ell}_{s+1} \right\rangle + \frac{1 - \rho_s}{\gamma_s} \left\langle x_{s+1} - x_s, \widehat{\ell}_s \right\rangle + 2d. \tag{12}$$

From Eq. (10) with $x = x_{s+1}$,

$$\left\langle x_{s+1} - x_s, \frac{\widehat{\ell}_s}{\gamma_s} \right\rangle \geq \left\langle x_{s+1} - x_s, \frac{1}{x_s} \right\rangle.$$

Since $\rho_s \geq 1$, this implies

$$\begin{aligned}\frac{1-\rho_s}{\gamma_s}\left\langle x_{s+1} - x_s, \widehat{\ell}_s \right\rangle &\leq (1-\rho_s)\left\langle x_{s+1} - x_s, \frac{1}{x_s} \right\rangle \\ &= (\rho_s - 1)\left(d - \left\langle \frac{1}{x_s}, x_{s+1} \right\rangle\right) \leq (\rho_s - 1)d.\end{aligned}$$

For the first term, using Eq. (9),

$$\begin{aligned}|\langle x_{s+1} - x_s, \widehat{\ell}_s - \widehat{\ell}_{s+1}\rangle| &\leq \sum_{i=1}^{d} |x_{s+1,i} - x_{s,i}||\widehat{\ell}_{s,i} - \widehat{\ell}_{s+1,i}| \\ &\leq \beta_s \sum_{i=1}^{d} \frac{|x_{s+1,i} - x_{s,i}|}{\sqrt{x_{s-1,i}}} + \beta_{s+1} \sum_{i=1}^{d} \frac{|x_{s+1,i} - x_{s,i}|}{\sqrt{x_{s,i}}}.\end{aligned}$$

The first term can be bounded as follows:

$$\begin{aligned}\beta_s \sum_{i=1}^{d} \frac{|x_{s+1,i} - x_{s,i}|}{\sqrt{x_{s-1,i}}} &= \beta_s \sum_{i=1}^{d} \frac{|x_{s+1,i} - x_{s,i}|}{\sqrt{x_{s+1,i}x_{s,i}}} \sqrt{\frac{x_{s+1,i}x_{s,i}}{x_{s-1,i}}} \\ &\leq \beta_s \sqrt{\sum_{i=1}^{d} \frac{(x_{s+1,i} - x_{s,i})^2}{x_{s+1,i}x_{s,i}}} \sqrt{\sum_{i=1}^{d} \frac{x_{s+1,i}x_{s,i}}{x_{s-1,i}}} \\ &\leq \beta_s \sqrt{\sum_{i=1}^{d} \frac{(x_{s+1,i} - x_{s,i})^2}{x_{s+1,i}x_{s,i}}} \sqrt{\sum_{i=1}^{d} \frac{x_{s,i}}{x_{s-1,i}}} \\ &\leq \beta_s \sqrt{\alpha(Q_{s+1} - 2d)},\end{aligned}$$

where the first inequality uses the Cauchy-Schwarz inequality and the last step follows from Eq. (8) and $Q_s \leq \alpha$.

The second term can be bounded as follows:

$$\begin{aligned}\beta_{s+1} \sum_{i=1}^{d} \frac{|x_{s+1,i} - x_{s,i}|}{\sqrt{x_{s,i}}} &= \beta_{s+1} \sum_{i=1}^{d} \frac{|x_{s+1,i} - x_{s,i}|}{\sqrt{x_{s+1,i}x_{s,i}}} \sqrt{x_{s+1,i}} \\ &\leq \beta_{s+1} \sqrt{\sum_{i=1}^{d} \frac{(x_{s+1,i} - x_{s,i})^2}{x_{s+1,i}x_{s,i}}} \sqrt{\sum_{i=1}^{d} x_{s+1,i}} \\ &= \beta_{s+1} \sqrt{Q_{s+1} - 2d},\end{aligned}$$

where the inequality follows from the Cauchy-Schwarz inequality and the last step uses Eq. (8).

Combining these two bounds, we get,

$$|\langle x_{s+1} - x_s, \widehat{\ell}_s - \widehat{\ell}_{s+1}\rangle| \leq (\beta_s\sqrt{\alpha} + \beta_{s+1})\sqrt{Q_{s+1} - 2d}.$$

Combining the bounds,

$$Q_{s+1} \leq (\rho_s - 1)d + \frac{\rho_s(\beta_s\sqrt{\alpha} + \beta_{s+1})\sqrt{Q_{s+1} - 2d}}{\gamma_s} + 2d.$$

Using $a\sqrt{Q_{s+1} - 2d} \leq (Q_{s+1} - 2d + a^2)/2$ with $a = \frac{\rho_s(\beta_s\sqrt{\alpha} + \beta_{s+1})}{\gamma_s}$, we obtain

$$Q_{s+1} \leq 2(\rho_s - 1)d + \frac{\rho_s^2(\beta_s\sqrt{\alpha} + \beta_{s+1})^2}{\gamma_s^2} + 2d \leq \alpha,$$

where the last inequality follows from the assumed parameter condition.

This completes the induction. □

Now we are ready to prove Lemma A.5.

Proof of Lemma A.5. We show that the choices of γ_s and β_s in Proposition A.1 satisfy the conditions in Lemma A.6 with $\alpha = 4d$. By definition of γ_s, we know that $\rho_s \triangleq \frac{\gamma_s}{\gamma_{s+1}} = \sqrt{2 \cdot \frac{\log(8ds^2/\delta)}{\log(8d(s+1)^2/\delta)}} \leq \sqrt{2}$. For $s \geq 2$,

$$\log \frac{8d(s+1)^2}{\delta} = \log \frac{8ds^2}{\delta} + 2\log\left(1 + \frac{1}{s}\right) \leq 2\log \frac{8ds^2}{\delta},$$

since $2\log\left(1+\frac{1}{s}\right) \leq 2\log\frac{3}{2} \leq \log\frac{8ds^2}{\delta}$. Thus, $1 \leq \rho_s \leq \sqrt{2}$. This also implies $\beta_s \geq \beta_{s+1}$, and hence $\max\{\beta_s, \beta_{s+1}\} = \beta_s$.

Next we verify the first condition in Lemma B.3. From the initialization, we have $x_1 = \mathbf{1}/d$. We know that $\widehat{\ell}_2 \in [-1, 1]^d$ since each observed entry is an empirical average of rewards in $[-1, 1]$ and unobserved entries are set to zero.

Taking $x = x_1$ in the first-order optimality condition for phase 2, we get

$$\begin{aligned} &\left\langle \widehat{\ell}_2 - \frac{\gamma_2}{x_2}, x_1 - x_2 \right\rangle \geq 0 \\ \implies &\left\langle \frac{1}{x_2}, x_1 - x_2 \right\rangle \leq \frac{\langle \widehat{\ell}_2, x_1 - x_2 \rangle}{\gamma_2}. \end{aligned}$$

Since $x_1 = \mathbf{1}/d$, $\left\langle \frac{1}{x_2}, x_1 - x_2 \right\rangle = \sum_{i=1}^d \frac{x_{1,i}}{x_{2,i}} - d = \sum_{i=1}^d \frac{x_{1,i}}{x_{2,i}} + d\sum_{i=1}^d x_{2,i} - 2d = \sum_{i=1}^d \frac{x_{1,i}}{x_{2,i}} + \sum_{i=1}^d \frac{x_{2,i}}{x_{1,i}} - 2d = Q_2 - 2d$.

Moreover, since $\|\widehat{\ell}_2\|_\infty \leq 1$ and $\|x_1 - x_2\|_1 \leq 2$, we have $Q_2 - 2d \leq \frac{2}{\gamma_2}$.

Consequently, $Q_2 \leq \frac{2}{\gamma_2} + 2d \leq 4d$ since $\gamma_2 = 5\sqrt{16d\log(32d/\delta)} \geq 1/d$ for $\delta \in (0, 1)$.

We now verify the second condition in Lemma A.6. Using $\gamma_s = 5\beta_s$, $\rho_s \leq \sqrt{2}$, $\beta_{s+1} \leq \beta_s$, and $\alpha = 4d$, we have

$$2(\rho_s - 1)d + \frac{\rho_s^2(\beta_s\sqrt{\alpha} + \beta_{s+1})^2}{\gamma_s^2} + 2d \leq 2(\sqrt{2} - 1)d + \frac{2(1 + 2\sqrt{d})^2\beta_s^2}{(5\beta_s)^2} + 2d \leq 4d = \alpha.$$

Therefore, the conditions in Lemma A.6 hold with $\alpha = 4d$, implying that for all $s \geq 2$, $Q_s = \sum_{i=1}^d \left(\frac{x_{s,i}}{x_{s-1,i}} + \frac{x_{s-1,i}}{x_{s,i}}\right) \leq 4d$. □

Finally, we prove the last-iterate convergence guarantee.

Proof of Proposition A.1. We prove the claim on the event $\mathcal{E}_1$, which holds with probability at least $1 - \delta$ by Lemma A.3. Since the length of epoch s is 2^{s-1}, equivalently, we prove that $\mathrm{DGap}(x_s, y_s) = \mathcal{O}(d2^{-s/2}\sqrt{\log(ds^2/\delta)})$. Let $i^\star \in \arg\min_{i\in[d]} \ell_i$ denote the optimal action. Then, using Lemma 4.2 and $\mathcal{E}_1$, we have

$$\begin{aligned} \mathrm{Reg}(x_s) &= \langle x_s, \widehat{\ell}_s \rangle - \widehat{\ell}_{s,i^\star} + \langle x_s - e_{i^\star}, \ell - \widehat{\ell}_s \rangle \\ &\leq d\gamma_s - \frac{\gamma_s}{x_{s,i^\star}} + \frac{\beta_s}{\sqrt{x_{s-1,i^\star}}} + \sum_{i=1}^d \frac{\beta_s x_{s,i}}{\sqrt{x_{s-1,i}}} \\ &\leq d\gamma_s - \frac{\gamma_s}{x_{s,i^\star}} + \frac{\beta_s}{\sqrt{x_{s-1,i^\star}}} + \beta_s \sum_{i=1}^d \sqrt{\frac{x_{s,i}}{x_{s-1,i}}}\sqrt{x_{s,i}} \end{aligned}$$

$$\leq d\gamma_s - \frac{\gamma_s}{x_{s,i^\star}} + \frac{\beta_s}{\sqrt{x_{s-1,i^\star}}} + \beta_s \sqrt{\sum_{i=1}^{d} \frac{x_{s,i}}{x_{s-1,i}}} \sqrt{\sum_{i=1}^{d} x_{s,i}},$$

where the last step follows from the Cauchy-Schwarz inequality.

From Lemma A.5, we have a bound on the sum of ratios $\sum_{i=1}^{d} \frac{x_{s,i}}{x_{s-1,i}} \leq 4d$. Consequently, we have coordinate-wise multiplicative stability: $x_{s,i} \leq 4d\, x_{s-1,i}$ for all $i \in [d]$. Thus,

$$\begin{aligned}\mathrm{Reg}(x_s) &\leq d\gamma_s - \frac{\gamma_s}{4dx_{s-1,i^\star}} + \frac{\beta_s}{\sqrt{x_{s-1,i^\star}}} + \beta_s\sqrt{4d} \\ &\leq d\gamma_s + \frac{4d\beta_s^2}{4\gamma_s} + 2\beta_s\sqrt{d},\end{aligned}$$

where the last step uses AM-GM inequality.

We finish the proof by plugging in $\gamma_s = 5\beta_s$ and $\beta_s = \sqrt{\frac{16\log(8ds^2/\delta)}{2^{s-2}}}$. □

B Omitted Proofs for Games

In this section, we show the omitted proofs in Section 4 for the general game case. We start with the lemma showing the concentration between $\widehat{A}_{s,ij}$ and the true game matrix A_{ij}.

Definition B.1. *Define $\mathcal{E}_2$ to be the event:* $\mathcal{E}_2 \triangleq \left\{ s \geq 2, \forall i,j \in [d], \left|\widehat{A}_{s,ij} - A_{ij}\right| \leq \frac{\beta_s}{\sqrt{x_{s-1,i} y_{s-1,j}}} \right\}$ *where* $\beta_s = \sqrt{\frac{16\log(8d^2s^2/\delta)}{2^{s-2}}}$.

Lemma B.2 (Estimation error)**.** *Algorithm 2 guarantees that $\mathcal{E}_2$ holds with probability at least $1-\delta$.*

Proof. Following a similar analysis to Lemma A.3, we obtain a counterpart for Eq. (6) in the game case: given a pair of $(i,j) \in [d] \times [d]$, we know that

$$\mathsf{Pr}\left[|\widehat{A}_{s,ij} - A_{ij}| > \sqrt{\frac{16\log(4d^2/\xi_s)}{2^{s-2}x_{s-1,i}}}\right] \leq e^{-2\log(4d^2/\xi_s)} + 2e^{-16\log(4d^2/\xi_s)} \leq \frac{\xi_s}{d^2}.$$

Picking $\xi_s = \frac{\delta}{2s^2}$ and taking a union bound over all (i,j) pairs and $s = 1, 2, \ldots$ complete the proof. □

We then prove Lemma 4.3, which is the analogue of Lemma 4.2 in the single player case. For clarity, we restate the lemma as follows.

Lemma 4.3. *For every epoch $s \geq 1$, Algorithm 2 satisfies:*

$$x_s^\top \widehat{A}_s e_j - e_i^\top \widehat{A}_s y_s \leq 2d\gamma_s - \frac{\gamma_s}{x_{s,i}} - \frac{\gamma_s}{y_{s,j}},\ \forall i,j \in [d]. \tag{5}$$

Proof. Since $\Phi_s(x,y)$ is strongly convex in x and strongly concave in y, (x_s, y_s) is the unique saddle point of $\Phi_s(x,y)$. Therefore we know that for any $x \in \Delta_d$ and $y \in \Delta_d$, $\Phi_s(x_s, y) \leq \Phi_s(x_s, y_s) \leq \Phi_s(x, y_s)$, meaning that

$$\begin{aligned}x_s &= \operatorname*{argmin}_{x\in\Delta_d} \Phi_s(x, y_s) = \operatorname*{argmin}_{x\in\Delta_d} \left\{ x^\top \widehat{A}_s y_s + \gamma_s \sum_{i=1}^{d} \log\frac{1}{x_i} \right\}, \\ y_s &= \operatorname*{argmin}_{y\in\Delta_d} -\Phi_s(x_s, y) = \operatorname*{argmin}_{y\in\Delta_d} \left\{ -x_s^\top \widehat{A}_s y + \gamma_s \sum_{i=1}^{d} \log\frac{1}{y_i} \right\}.\end{aligned}$$

With an abuse of notation, for $v \in \mathbb{R}^d$, we denote $\frac{1}{v} \in \mathbb{R}^d$ with the i-th coordinate $\frac{1}{v_i}$. According to first-order optimality, we have for any $x \in \Delta_d$ and $y \in \Delta_d$,

$$\left\langle \widehat{A}_s y_s - \frac{\gamma_s}{x_s}, x - x_s \right\rangle \geq 0,$$
$$\left\langle -\widehat{A}_s^\top x_s - \frac{\gamma_s}{y_s}, y - y_s \right\rangle \geq 0.$$

Picking $x = e_i$ and $y = e_j$ and summing up the above two inequalities lead to the conclusion. □

Next, we aim to prove Lemma 4.4, showing the aggregate multiplicative stability of Algorithm 2 between consecutive epochs. Let $Q_s \triangleq \sum_{i=1}^d \left(\frac{x_{s,i}}{x_{s-1,i}} + \frac{x_{s-1,i}}{x_{s,i}} \right)$ and $R_s \triangleq \sum_{j=1}^d \left(\frac{y_{s,j}}{y_{s-1,j}} + \frac{y_{s-1,j}}{y_{s,j}} \right)$ for $s \geq 2$. We show aggregate multiplicative stability by bounding the quantity $S_s \triangleq Q_s + R_s$.

In fact, we will prove the following more general Lemma B.3 and Lemma 4.4 will serve as a corollary by picking the parameter γ_s as stated in Theorem 4.1.

Lemma B.3 (Aggregate multiplicative stability)**.** *Suppose that $\mathcal{E}_2$ holds and define $\rho_s = \gamma_s/\gamma_{s+1}$. Further suppose that there exists $\alpha \geq 4d$ such that the sequence $\{\gamma_s\}_{s \geq 1}$ in Algorithm 2 satisfies (1) $S_2 \leq \alpha$, (2) for all $s \geq 2$, $4(\rho_s - 1)d + \frac{2\rho_s^2(\alpha\beta_s - 2d\beta_s + \sqrt{d}\beta_{s+1})^2}{\gamma_s^2} + 4d \leq \alpha$, (3) $\rho_s \in [1, \sqrt{2}]$. Then Algorithm 2 guarantees that, for all $s \geq 2$, $S_s = \sum_{i=1}^d \left(\frac{x_{s,i}}{x_{s-1,i}} + \frac{x_{s-1,i}}{x_{s,i}} \right) + \sum_{j=1}^d \left(\frac{y_{s,j}}{y_{s-1,j}} + \frac{y_{s-1,j}}{y_{s,j}} \right) \leq \alpha$.*

Proof. We prove the result by induction. The base case $s = 2$ follows from the assumption that $S_2 \leq \alpha$. Suppose that for some $s \geq 2$, $S_s \leq \alpha$. We prove the same conclusion for $s + 1$.

First, for all $s \geq 2$, we have

$$\sum_{i=1}^d \frac{(x_{s,i} - x_{s-1,i})^2}{x_{s,i} x_{s-1,i}} = \sum_{i=1}^d \left(\frac{x_{s,i}}{x_{s-1,i}} + \frac{x_{s-1,i}}{x_{s,i}} \right) - 2d = Q_s - 2d, \tag{13}$$
$$\sum_{j=1}^d \frac{(y_{s,j} - y_{s-1,j})^2}{y_{s,j} y_{s-1,j}} = \sum_{j=1}^d \left(\frac{y_{s,j}}{y_{s-1,j}} + \frac{y_{s-1,j}}{y_{s,j}} \right) - 2d = R_s - 2d, \tag{14}$$

where the left-hand sides of both equations are always non-negative, implying that $Q_s \geq 2d$ and $R_s \geq 2d$. Consequently, $Q_s + R_s = S_s \geq 4d$, and thus $\alpha \geq 4d$.

Moreover,

$$S_s \leq \alpha \text{ and } R_s \geq 2d \implies Q_s \leq \alpha - 2d \text{ and } \sum_{i=1}^d \frac{x_{s,i}}{x_{s-1,i}} \leq \alpha - 2d, \tag{15}$$
$$S_s \leq \alpha \text{ and } Q_s \geq 2d \implies R_s \leq \alpha - 2d \text{ and } \sum_{j=1}^d \frac{y_{s,j}}{y_{s-1,j}} \leq \alpha - 2d. \tag{16}$$

By $\mathcal{E}_2$, for every $i, j \in [d]$,

$$\begin{aligned} \left| \widehat{A}_{s,ij} - \widehat{A}_{s+1,ij} \right| &\leq \left| \widehat{A}_{s,ij} - A_{ij} \right| + \left| \widehat{A}_{s+1,ij} - A_{ij} \right| \\ &\leq \frac{\beta_s}{\sqrt{x_{s-1,i} y_{s-1,j}}} + \frac{\beta_{s+1}}{\sqrt{x_{s,i} y_{s,j}}}. \end{aligned} \tag{17}$$

Applying first-order optimality at phases s and $s + 1$, we know that for any $x \in \Delta_d$,

$$\left\langle \widehat{A}_s y_s - \frac{\gamma_s}{x_s}, x - x_s \right\rangle \geq 0, \tag{18}$$
$$\left\langle \widehat{A}_{s+1} y_{s+1} - \frac{\gamma_{s+1}}{x_{s+1}}, x - x_{s+1} \right\rangle \geq 0. \tag{19}$$

Picking $x = x_{s+1}$ in Eq. (18) and $x = x_s$ in Eq. (19), and then summing the two inequalities and using Eq. (13), gives

$$Q_{s+1} - 2d \le \left\langle x_{s+1} - x_s, \frac{\widehat{A}_s y_s}{\gamma_s} - \frac{\widehat{A}_{s+1} y_{s+1}}{\gamma_{s+1}} \right\rangle.$$

Let $\rho_s \triangleq \gamma_s/\gamma_{s+1}$. A further decomposition shows that

$$\begin{aligned} Q_{s+1} \le \frac{\rho_s}{\gamma_s} \left\langle x_{s+1} - x_s, \widehat{A}_{s+1}(y_s - y_{s+1}) \right\rangle + \frac{1-\rho_s}{\gamma_s} \left\langle x_{s+1} - x_s, \widehat{A}_s y_s \right\rangle \\ + \frac{\rho_s}{\gamma_s} \left\langle x_{s+1} - x_s, (\widehat{A}_s - \widehat{A}_{s+1}) y_s \right\rangle + 2d. \end{aligned} \tag{20}$$

Similarly, first-order optimality of the y-updates gives

$$\begin{aligned} R_{s+1} \le \frac{\rho_s}{\gamma_s} \left\langle y_{s+1} - y_s, \widehat{A}_{s+1}^\top (x_{s+1} - x_s) \right\rangle + \frac{\rho_s - 1}{\gamma_s} \left\langle y_{s+1} - y_s, \widehat{A}_s^\top x_s \right\rangle \\ + \frac{\rho_s}{\gamma_s} \left\langle y_{s+1} - y_s, (\widehat{A}_{s+1} - \widehat{A}_s)^\top x_s \right\rangle + 2d. \end{aligned} \tag{21}$$

We now add Eq. (20) and Eq. (21). The first terms cancel because

$$\left\langle x_{s+1} - x_s, \widehat{A}_{s+1}(y_s - y_{s+1}) \right\rangle + \left\langle y_{s+1} - y_s, \widehat{A}_{s+1}^\top (x_{s+1} - x_s) \right\rangle = 0.$$

We next control the two terms caused by the change of γ_s. By Eq. (18) with $x = x_{s+1}$,

$$\left\langle x_{s+1} - x_s, \frac{\widehat{A}_s y_s}{\gamma_s} \right\rangle \ge \left\langle x_{s+1} - x_s, \frac{1}{x_s} \right\rangle.$$

Since $\rho_s \ge 1$, this implies

$$\begin{aligned} \frac{1-\rho_s}{\gamma_s} \left\langle x_{s+1} - x_s, \widehat{A}_s y_s \right\rangle &\le (1-\rho_s) \left\langle x_{s+1} - x_s, \frac{1}{x_s} \right\rangle \\ &= (\rho_s - 1) \left(d - \left\langle \frac{1}{x_s}, x_{s+1} \right\rangle \right) \le (\rho_s - 1) d. \end{aligned}$$

Similarly, first-order optimality of the y-update at phase s gives

$$\left\langle -\widehat{A}_s^\top x_s - \frac{\gamma_s}{y_s}, y - y_s \right\rangle \ge 0, \qquad \forall y \in \Delta_d.$$

Taking $y = y_{s+1}$ yields

$$\left\langle y_{s+1} - y_s, \frac{\widehat{A}_s^\top x_s}{\gamma_s} \right\rangle \le - \left\langle y_{s+1} - y_s, \frac{1}{y_s} \right\rangle.$$

Therefore,

$$\begin{aligned} \frac{\rho_s - 1}{\gamma_s} \left\langle y_{s+1} - y_s, \widehat{A}_s^\top x_s \right\rangle &\le -(\rho_s - 1) \left\langle y_{s+1} - y_s, \frac{1}{y_s} \right\rangle \\ &= (\rho_s - 1) \left(d - \left\langle \frac{1}{y_s}, y_{s+1} \right\rangle \right) \le (\rho_s - 1) d. \end{aligned}$$

It remains to bound the estimator-difference terms. By Eq. (17),

$$\begin{aligned} \left| \left\langle x_{s+1} - x_s, (\widehat{A}_s - \widehat{A}_{s+1}) y_s \right\rangle \right| &\le \sum_{i=1}^d |x_{s+1,i} - x_{s,i}| \sum_{j=1}^d \left| \widehat{A}_{s,ij} - \widehat{A}_{s+1,ij} \right| y_{s,j} \\ &\le \beta_s \sum_{i=1}^d \frac{|x_{s+1,i} - x_{s,i}|}{\sqrt{x_{s-1,i}}} \sum_{j=1}^d \frac{y_{s,j}}{\sqrt{y_{s-1,j}}} \end{aligned}$$

$$+ \beta_{s+1} \sum_{i=1}^{d} \frac{|x_{s+1,i} - x_{s,i}|}{\sqrt{x_{s,i}}} \sum_{j=1}^{d} \sqrt{y_{s,j}}.$$

The first term can be bounded as follows:

$$\begin{aligned}
\beta_s \sum_{i=1}^{d} \frac{|x_{s+1,i} - x_{s,i}|}{\sqrt{x_{s-1,i}}} \sum_{j=1}^{d} \frac{y_{s,j}}{\sqrt{y_{s-1,j}}} &= \beta_s \sum_{i=1}^{d} \frac{|x_{s+1,i} - x_{s,i}|}{\sqrt{x_{s+1,i} x_{s,i}}} \sqrt{\frac{x_{s+1,i} x_{s,i}}{x_{s-1,i}}} \sum_{j=1}^{d} \sqrt{\frac{y_{s,j}}{y_{s-1,j}}} \sqrt{y_{s,j}} \\
&\le \beta_s \sqrt{Q_{s+1} - 2d} \sqrt{\sum_{i=1}^{d} \frac{x_{s+1,i} x_{s,i}}{x_{s-1,i}}} \sqrt{\sum_{j=1}^{d} \frac{y_{s,j}}{y_{s-1,j}}} \sqrt{\sum_{j=1}^{d} y_{s,j}} \\
&\le \beta_s \sqrt{Q_{s+1} - 2d} \sqrt{\sum_{i=1}^{d} \frac{x_{s,i}}{x_{s-1,i}}} \sqrt{\sum_{j=1}^{d} \frac{y_{s,j}}{y_{s-1,j}}} \\
&\le \beta_s (\alpha - 2d) \sqrt{Q_{s+1} - 2d},
\end{aligned}$$

where the first inequality uses the Cauchy-Schwarz inequality and Eq. (13), and the last step follows from Eq. (15) and Eq. (16).

The second term can be bounded as follows:

$$\begin{aligned}
\beta_{s+1} \sum_{i=1}^{d} \frac{|x_{s+1,i} - x_{s,i}|}{\sqrt{x_{s,i}}} \sum_{j=1}^{d} \sqrt{y_{s,j}} &= \beta_{s+1} \sum_{i=1}^{d} \frac{|x_{s+1,i} - x_{s,i}|}{\sqrt{x_{s+1,i} x_{s,i}}} \sqrt{x_{s+1,i}} \sum_{j=1}^{d} \sqrt{y_{s,j}} \\
&\le \beta_{s+1} \sqrt{Q_{s+1} - 2d} \sqrt{\sum_{i=1}^{d} x_{s+1,i}} \sqrt{d \sum_{j=1}^{d} y_{s,j}} \\
&= \beta_{s+1} \sqrt{d(Q_{s+1} - 2d)},
\end{aligned}$$

where the inequality follows from the Cauchy-Schwarz inequality and Eq. (13).

Combining these two bounds, we get,

$$\left| \left\langle x_{s+1} - x_s, (\widehat{A}_s - \widehat{A}_{s+1}) y_s \right\rangle \right| \le (\alpha \beta_s - 2d\beta_s + \sqrt{d}\beta_{s+1}) \sqrt{Q_{s+1} - 2d}.$$

Similarly, using Eq. (17) and Eq. (14),

$$\left| \left\langle y_{s+1} - y_s, (\widehat{A}_{s+1} - \widehat{A}_s)^\top x_s \right\rangle \right| \le (\alpha \beta_s - 2d\beta_s + \sqrt{d}\beta_{s+1}) \sqrt{R_{s+1} - 2d}.$$

Combining the above bounds gives

$$S_{s+1} = Q_{s+1} + R_{s+1} \le 2(\rho_s - 1)d + \frac{\rho_s (\alpha \beta_s - 2d\beta_s + \sqrt{d}\beta_{s+1})}{\gamma_s} \left(\sqrt{Q_{s+1} - 2d} + \sqrt{R_{s+1} - 2d} \right) + 4d.$$

Since

$$\sqrt{Q_{s+1} - 2d} + \sqrt{R_{s+1} - 2d} \le \sqrt{2S_{s+1} - 8d},$$

we have

$$S_{s+1} \le 2(\rho_s - 1)d + \frac{\rho_s (\alpha \beta_s - 2d\beta_s + \sqrt{d}\beta_{s+1})}{\gamma_s} \sqrt{2S_{s+1} - 8d} + 4d.$$

Using $a\sqrt{2S_{s+1} - 8d} \le (S_{s+1} - 4d)/2 + a^2$ with

$$a = \frac{\rho_s (\alpha \beta_s - 2d\beta_s + \sqrt{d}\beta_{s+1})}{\gamma_s},$$

we obtain

$$S_{s+1} \le 4(\rho_s - 1)d + \frac{2\rho_s^2 (\alpha \beta_s - 2d\beta_s + \sqrt{d}\beta_{s+1})^2}{\gamma_s^2} + 4d \le \alpha,$$

where the last inequality follows from the assumed parameter condition.

Thus the induction closes. □

Lemma 4.4 is then proven by verifying the γ_s choice used in Theorem 4.1 satisfies the condition of Lemma B.3.

Proof of Lemma 4.4. We verify that the choice of γ_s and β_s satisfy the condition in Lemma B.3 with $\alpha = 8d$. By definition of γ_s, we know that $\rho_s \triangleq \frac{\gamma_s}{\gamma_{s+1}} = \frac{\beta_s}{\beta_{s+1}} = \sqrt{2 \cdot \frac{\log(8d^2 s^2/\delta)}{\log(8d^2(s+1)^2/\delta)}} \le \sqrt{2}$. Moreover, for $s \ge 2$,

$$\log \frac{8d^2(s+1)^2}{\delta} = \log \frac{8d^2 s^2}{\delta} + 2\log\left(1 + \frac{1}{s}\right) \le 2\log \frac{8d^2 s^2}{\delta},$$

where the inequality follows from $2\log\left(1+\frac{1}{s}\right) \le 2\log\frac{3}{2} \le \log\frac{8d^2s^2}{\delta}$. Therefore, we have $1 \le \rho_s \le \sqrt{2}$. In addition, it also means that $\beta_s \ge \beta_{s+1}$.

Now we check the first condition in Lemma B.3. From the initialization, we have $x_1 = y_1 = \mathbf{1}/d$. Moreover, $\widehat{A}_2 \in [-1,1]^{d\times d}$, since each observed entry is an empirical average of rewards in $[-1,1]$, while unobserved entries are set to zero.

Taking $x = x_1$ in the first-order optimality condition for phase 2, we get

$$\left\langle \widehat{A}_2 y_2 - \frac{\gamma_2}{x_2}, x_1 - x_2 \right\rangle \ge 0$$
$$\implies \left\langle \frac{1}{x_2}, x_1 - x_2 \right\rangle \le \frac{\langle \widehat{A}_2 y_2, x_1 - x_2\rangle}{\gamma_2}.$$

Since $x_1 = \mathbf{1}/d$, $\left\langle \frac{1}{x_2}, x_1 - x_2 \right\rangle = \sum_{i=1}^d \frac{x_{1,i}}{x_{2,i}} - d = \sum_{i=1}^d \frac{x_{1,i}}{x_{2,i}} + d\sum_{i=1}^d x_{2,i} - 2d = \sum_{i=1}^d \frac{x_{1,i}}{x_{2,i}} + \sum_{i=1}^d \frac{x_{2,i}}{x_{1,i}} - 2d = Q_2 - 2d$.

Moreover, $\widehat{A}_2 \in [-1,1]^{d\times d}$ implies $\|\widehat{A}_2 y_2\|_\infty \le 1$, and $\|x_1 - x_2\|_1 \le 2$. Hence, $Q_2 - 2d \le \frac{2}{\gamma_2}$.

Similarly, taking $y = y_1$ in the first-order condition for phase 2 and using the fact that $y_1 = \mathbf{1}/d$, we get $R_2 - 2d \le \frac{2}{\gamma_2}$.

Consequently, $S_2 = Q_2 + R_2 \le \frac{4}{\gamma_2} + 4d \le 8d$ since $\gamma_2 = 12\sqrt{16d\log(32d^2/\delta)} \ge 1/d$ for $\delta \in (0,1)$.

Next we check the second condition in Lemma B.3. Using $\gamma_s = 10\sqrt{d}\beta_s$, $\rho_s \le \sqrt{2}$, $\beta_{s+1} \le \beta_s$, and $\alpha = 8d$, we have

$$4(\rho_s - 1)d + \frac{2\rho_s^2(\alpha\beta_s - 2d\beta_s + \sqrt{d}\beta_{s+1})^2}{\gamma_s^2} + 4d \le 4(\sqrt{2}-1)d + \frac{4(1+6\sqrt{d})^2\beta_s^2 d}{100 d\beta_s^2} + 4d$$
$$\le 8d = \alpha.$$

Therefore the conditions in Lemma B.3 hold with $\alpha = 8d$. Hence, for all $s \ge 2$, $S_s = \sum_{i=1}^d \left(\frac{x_{s,i}}{x_{s-1,i}} + \frac{x_{s-1,i}}{x_{s,i}}\right) + \sum_{j=1}^d \left(\frac{y_{s,j}}{y_{s-1,j}} + \frac{y_{s-1,j}}{y_{s,j}}\right) \le 8d$. □

With all the above lemmas, we are ready to prove our main result Theorem 4.1.

Proof of Theorem 4.1. We prove the claim on the event $\mathcal{E}_2$, which holds with probability at least $1-\delta$ by Lemma B.2. Since the length of epoch s is 2^{s-1}, equivalently, we prove that $\mathrm{DGap}(x_s, y_s) = \mathcal{O}(d^{3/2}2^{-s/2}\sqrt{\log(d^2s^2/\delta)})$. To proceed with the proof, fix any epoch $s \ge 2$ and any pair $i, j \in [d]$. We first decompose

$$x_s^\top A e_j - e_i^\top A y_s = x_s^\top \widehat{A}_s e_j - e_i^\top \widehat{A}_s y_s + x_s^\top (A - \widehat{A}_s) e_j + e_i^\top (\widehat{A}_s - A) y_s. \tag{22}$$

By Lemma 4.3, we know that

$$x_s^\top \widehat{A}_s e_j - e_i^\top \widehat{A}_s y_s \le 2\gamma_s d - \frac{\gamma_s}{x_{s,i}} - \frac{\gamma_s}{y_{s,j}}.$$

It remains to control the last two estimation-error terms in Eq. (22). For the first one, using $\mathcal{E}_2$ gives

$$x_s^\top (A - \widehat{A}_s) e_j \le \sum_{k=1}^{d} x_{s,k} \left| A_{kj} - \widehat{A}_{s,kj} \right| \le \beta_s \sum_{k=1}^{d} \frac{x_{s,k}}{\sqrt{x_{s-1,k} y_{s-1,j}}}.$$

From Eq. (15), Eq. (16) and Lemma 4.4, we have sum of ratio bounds $\sum_{k=1}^{d} \frac{x_{s,k}}{x_{s-1,k}} \le 6d$ and $\sum_{k=1}^{d} \frac{y_{s,k}}{y_{s-1,k}} \le 6d$. Consequently, we have coordinate-wise multiplicative stability: $x_{s,i} \le 6d\, x_{s-1,i}$ and $y_{s,j} \le 6d\, y_{s-1,j}$ for all $s \ge 2, i, j \in [d]$. Hence

$$\begin{aligned}
x_s^\top (A - \widehat{A}_s) e_j &\le \beta_s \sum_{k=1}^{d} \frac{x_{s,k}}{\sqrt{x_{s-1,k} y_{s-1,j}}} \\
&= \frac{\beta_s}{\sqrt{y_{s-1,j}}} \sum_{k=1}^{d} \sqrt{x_{s,k}} \sqrt{\frac{x_{s,k}}{x_{s-1,k}}} \\
&\le \frac{\beta_s}{\sqrt{y_{s-1,j}}} \sqrt{\sum_{k=1}^{d} x_{s,k}} \sqrt{\sum_{k=1}^{d} \frac{x_{s,k}}{x_{s-1,k}}} \\
&\le \frac{6d\beta_s}{\sqrt{y_{s,j}}},
\end{aligned}$$

where the second-last step uses the Cauchy-Schwarz inequality.

Similarly, for the other term, we have

$$e_i^\top (\widehat{A}_s - A) y_s \le \sum_{\ell=1}^{d} y_{s,\ell} \left| \widehat{A}_{s,i\ell} - A_{i\ell} \right| \le \beta_s \sum_{\ell=1}^{d} \frac{y_{s,\ell}}{\sqrt{x_{s-1,i} y_{s-1,\ell}}} \le \frac{6d\beta_s}{\sqrt{x_{s,i}}}.$$

Combining the three bounds, for every $i, j \in [d]$,

$$x_s^\top A e_j - e_i^\top A y_s \le 2\gamma_s d - \frac{\gamma_s}{x_{s,i}} - \frac{\gamma_s}{y_{s,j}} + \frac{6d\beta_s}{\sqrt{x_{s,i}}} + \frac{6d\beta_s}{\sqrt{y_{s,j}}} \le 2\gamma_s d + \frac{18 d^2 \beta_s^2}{\gamma_s},$$

where the last inequality is due to AM-GM inequality. We finish the proof by plugging in $\gamma_s = 10\sqrt{d}\beta_s$ and $\beta_s = \sqrt{\frac{16 \log(8d^2 s^2/\delta)}{2^{s-2}}}$ and noticing that $\mathrm{DGap}(x_s, y_s) = \max_{i,j \in [d]} \left\{ x_s^\top A e_j - e_i^\top A y_s \right\}$. □

C Additional Omitted Details for Section 3

C.1 Last-Iterate Convergence for Empirical NE with Uniform Exploration

In this section, we present one of our baseline algorithms that mixes the empirically best strategy with uniform exploration. Algorithm 3 also operates in an epoch-wise manner.

Proposition C.1. *For any fixed $\delta \in (0,1)$, Algorithm 3 guarantees that with probability at least $1-\delta$, $\mathrm{DGap}(x_t, y_t) = \mathcal{O}\left(\sqrt{d \log(d \log t/\delta)} t^{-\frac{1}{4}}\right)$ for all $t \ge 1$, where (x_t, y_t) denotes the strategy pair played in round t.*

Proof. We prove the claim on the event $\mathcal{E}_2$, which holds with probability at least $1-\delta$ by Lemma B.2. Since epoch s has length 2^{s-1}, it suffices to show that $\mathrm{DGap}(x_s, y_s) = \mathcal{O}(2^{-s/4}\sqrt{d \log(d^2 s^2/\delta)})$. Let $(\widehat{x}_s, \widehat{y}_s)$ denote a Nash Equilibrium of $\widehat{A}_s$. Fix any epoch $s \ge 2$. We begin by controlling the duality gap with respect to the empirical estimate $\widehat{A}_s$. Note that $\widehat{A}_{s,ij} \in [-1, 1]$ holds for all $i, j \in [d]$. From the sampling strategy of Algorithm 3, we have $\|x_s - \widehat{x}_s\|_1 \le 2\alpha_s$ and $\|y_s - \widehat{y}_s\|_1 \le 2\alpha_s$. Therefore, for all pairs $i, j \in [d]$,

$$x_s^\top \widehat{A}_s e_j \le \widehat{x}_s^\top \widehat{A}_s e_j + 2\alpha_s \le \max_{y \in \Delta_d} \widehat{x}_s^\top \widehat{A}_s y + 2\alpha_s = \widehat{x}_s^\top \widehat{A}_s \widehat{y}_s + 2\alpha_s,$$

Algorithm 3: Empirical NE with uniform exploration

Initialize game estimator $\widehat{A}_1$ to be the all-zero d by d matrix.
for $s = 1, 2, \ldots$ **do**
  Compute $\alpha_s = d^{1/2} 2^{-(s-1)/4}$.
  For row player: compute $\widehat{x}_s \in \operatorname{argmin}_{x \in \Delta_d} \max_{y \in \Delta_d} x^\top \widehat{A}_s y$ and $x_s = (1-\alpha_s)\widehat{x}_s + \alpha_s \frac{\mathbf{1}}{d}$.
  For column player: compute $\widehat{y}_s \in \operatorname{argmax}_{y \in \Delta_d} \min_{x \in \Delta_d} x^\top \widehat{A}_s y$ and $y_s = (1-\alpha_s)\widehat{y}_s + \alpha_s \frac{\mathbf{1}}{d}$.
  Initialize counters: $n_{s,ij} = 0$ for all $i, j \in [d]$.
  for $t = 2^{s-1}, \ldots, 2^s - 1$ **do**
    Row player samples $i_t \sim x_s$ and column player samples $j_t \sim y_s$.
    Both players observe (i_t, j_t) and r_t with $\mathbb{E}[r_t] = A_{i_t j_t}$.
    Increment counter $n_{s,i_t j_t} \leftarrow n_{s,i_t j_t} + 1$.
  Compute game estimator $\widehat{A}_{s+1,ij} = \frac{\mathbb{1}\{n_{s,ij} \neq 0\}}{n_{s,ij}} \sum_{\tau=2^{s-1}}^{2^s-1} r_\tau \cdot \mathbb{1}\{i_\tau = i, j_\tau = j\}$ for all $i, j \in [d]$.

$$e_i^\top \widehat{A}_s y_s \geq e_i^\top \widehat{A}_s \widehat{y}_s - 2\alpha_s \geq \min_{x \in \Delta_d} x^\top \widehat{A}_s \widehat{y}_s - 2\alpha_s = \widehat{x}_s^\top \widehat{A}_s \widehat{y}_s - 2\alpha_s.$$

Thus, for all $i, j \in [d]$, we have $x_s^\top \widehat{A}_s e_j - e_i^\top \widehat{A}_s y_s \leq 4\alpha_s$.

It remains to bound the last two estimation error terms in Eq. (22). For the first term, Lemma B.2 implies

$$x_s^\top (A - \widehat{A}_s) e_j \leq \sum_{k=1}^d x_{s,k} \left| A_{kj} - \widehat{A}_{s,kj} \right| \leq \beta_s \sum_{k=1}^d \frac{x_{s,k}}{\sqrt{x_{s-1,k} y_{s-1,j}}} \leq \frac{\beta_s d}{\alpha_{s-1}} = \frac{\beta_s d}{2\alpha_s},$$

where the last inequality follows from the fact that $x_{s-1,k} \geq \frac{\alpha_{s-1}}{d}$ and $y_{s-1,j} \geq \frac{\alpha_{s-1}}{d}$ for all $k, j \in [d]$, due to mixing with the uniform distribution.

Similarly, for the other term

$$e_i^\top (\widehat{A}_s - A) y_s \leq \sum_{\ell=1}^d y_{s,\ell} \left| \widehat{A}_{s,i\ell} - A_{i\ell} \right| \leq \beta_s \sum_{\ell=1}^d \frac{y_{s,\ell}}{\sqrt{x_{s-1,i} y_{s-1,\ell}}} \leq \frac{\beta_s d}{\alpha_{s-1}} = \frac{\beta_s d}{2\alpha_s}.$$

Combining the three bounds, for every $i, j \in [d]$, we get

$$x_s^\top A e_j - e_i^\top A y_s \leq 4\alpha_s + \frac{\beta_s d}{\alpha_s}.$$

We have $\beta_s = \sqrt{\frac{16 \log(8d^2 s^2/\delta)}{2^{s-2}}}$. Setting α_s optimally as $\sqrt{\beta_s d}$ yields $\mathrm{DGap}(x_s, y_s) = \max_{i,j \in [d]} \left\{ x_s^\top A e_j - e_i^\top A y_s \right\} = \mathcal{O}(2^{-s/4} \sqrt{d \log(d^2 s^2/\delta)})$. □

C.2 Discussion on Low-Regret-Low-Variance Property in Games

In this subsection, we explain why a low-duality-gap-low-variance property, analogous to Lemma 3.1, is insufficient to establish last-iterate convergence in the game setting. We begin by presenting the following lemma, the game counterpart to Lemma 3.1, which follows directly from Lemma 4.3.

Lemma C.2. *Define* $\widehat{\mathrm{DGap}}_s(x, y) = \max_{j \in [d]} x^\top \widehat{A}_s e_j - \min_{i \in [d]} e_i^\top \widehat{A}_s y$ *for all* $s \geq 1$. *For every epoch* $s \geq 1$, *Algorithm 2 satisfies:*

$$\widehat{\mathrm{DGap}}_s(x_s, y_s) \leq 2d\gamma_s, \tag{23}$$

$$\frac{1}{x_{s,i}} + \frac{1}{y_{s,j}} \leq 2d + \frac{\widehat{\mathrm{DGap}}_s(e_i, e_j)}{\gamma_s}, \ \forall i, j \in [d]. \tag{24}$$

Proof. By Lemma 4.3, for all $i, j \in [d]$, we have:

$$x_s^\top \widehat{A}_s e_j - e_i^\top \widehat{A}_s y_s \le 2d\gamma_s - \frac{\gamma_s}{x_{s,i}} - \frac{\gamma_s}{y_{s,j}} \le 2d\gamma_s.$$

Furthermore, rearranging the terms in Eq. (5) yields:

$$\frac{1}{x_{s,i}} + \frac{1}{y_{s,j}} \le 2d - \frac{1}{\gamma_s}\left(x_s^\top \widehat{A}_s e_j - e_i^\top \widehat{A}_s y_s\right) \le 2d + \frac{\widehat{\mathrm{DGap}}_s(e_i, e_j)}{\gamma_s}.$$

This completes the proof. □

Unlike the single-agent case, however, we cannot directly bound the true duality gap $\mathrm{DGap}(x_s, y_s)$ even if the following multiplicative closeness analogous to the single-player case holds due to Eq. (24):

$$\widehat{\mathrm{DGap}}_s(e_i, e_j) \le 2\mathrm{DGap}(e_i, e_j) + \widetilde{\mathcal{O}}(d\gamma_s),$$
$$\mathrm{DGap}(e_i, e_j) \le 2\widehat{\mathrm{DGap}}_s(e_i, e_j) + \widetilde{\mathcal{O}}(d\gamma_s).$$

This is because of the following crucial distinction: In the single-agent case, expectations translate linearly (i.e., $\mathbb{E}_{i\sim x_s}[\widehat{\mathrm{Reg}}_s(e_i)] = \widehat{\mathrm{Reg}}_s(x_s)$ and $\mathbb{E}_{i\sim x_s}[\mathrm{Reg}(e_i)] = \mathrm{Reg}(x_s)$), which allows multiplicative closeness for pure actions to neatly imply closeness for mixed strategies. In the game setting, this linear translation breaks down. Instead, we are restricted to Jensen's inequality ($\mathbb{E}_{i\sim x_s, j\sim y_s}[\mathrm{DGap}(e_i, e_j)] \ge \mathrm{DGap}(x_s, y_s)$), which fundamentally prevents the application of a similar argument.

D Additional Details for Section 5

D.1 Implementation Details

We generate the four game matrices (Figure 1a, Figure 1b, Figure 1c, Figure 1d) using the GUARD Green Security Game environment [Krever et al., 2025]. All instances share the same Lobéké region, the centroid-based target-value method, and a single mobile defender facing stationary attackers; the resulting zero-sum payoff matrices are affinely rescaled to $[0, 1]$. The remaining parameters are listed in Table 1.

Table 1: Parameters of the four game instances.

	Grid	Clusters	Attackers	Horizon	Size
Figure 1a	7×7	6	4	5	39×35
Figure 1b	7×7	6	5	5	61×21
Figure 1c	6×6	9	4	6	221×210
Figure 1d	6×6	11	4	7	520×495

Hyperparameters of all algorithms are chosen according to theoretical requirements where available, and otherwise tuned to ensure stable performance. Results are averaged over 10 independent runs for each game. All experiments were conducted on a Linux server running Ubuntu 24.04.3 LTS. The server was equipped with two Intel Xeon Gold 6530 processors, providing 64 physical CPU cores and 128 logical threads, and 1.0 TiB of RAM. The code was implemented in Python 3.13.12.

D.2 Details on the Computational Aspects of PMO-LB (Algorithm 2)

PMO-LB performs only $\mathcal{O}(\log T)$ strategy updates because its epochs have geometrically increasing lengths. At each update, it computes the unique saddle point of a strongly convex-strongly concave log-barrier-regularized empirical game.

In our experiments, we approximately solve this saddle-point problem using an optimistic multiplicative-weights solver, warm-started from the solution obtained in the preceding epoch. Each inner iteration requires computing the two matrix-vector products $\widehat{A}_s y$ and $\widehat{A}_s^\top x$, and therefore

costs $\mathcal{O}(mn)$ time for an $m \times n$ game. We terminate the solver when the centered log-barrier KKT stationarity residual is at most 10^{-10}, subject to a safeguard of 20,000 inner iterations. If the safeguard is reached before the target tolerance is attained, we return the iterate with the smallest residual encountered.

Let K_s denote the number of inner iterations performed in epoch s. The computational cost of the corresponding strategy update is $\mathcal{O}(K_s mn)$. Since PMO-LB invokes the saddle-point solver only $\mathcal{O}(\log T)$ times up to round T, its total inner-optimization cost is

$$\mathcal{O}\left(mn \sum_{s \le \lceil \log_2 T \rceil + 1} K_s\right).$$

Thus, for the experimental horizons $T = 10^7$ and $T = 10^8$, PMO-LB solves only approximately 23 and 27 regularized minimax problems, respectively. We retain K_s explicitly because our implementation uses a numerical stopping criterion, and we do not require a separate iteration-complexity guarantee for the inner OMWU solver in the statistical analysis.

Theorem 4.1 is stated for exact saddle-point solutions, but the proof is robust to approximate first-order solutions. Suppose that the pair returned in epoch s, denoted by (x_s, y_s), has centered KKT stationarity residual at most ε_s. Then the exact first-order inequalities used in Lemma 4.3 hold up to an additive error:

$$x_s^\top \widehat{A}_s e_j - e_i^\top \widehat{A}_s y_s \le 2d\gamma_s - \frac{\gamma_s}{x_{s,i}} - \frac{\gamma_s}{y_{s,j}} + 4\varepsilon_s, \qquad \forall i, j \in [d].$$

Accordingly, the multiplicative stability argument in Lemma 4.4 acquires additional perturbation terms of order $\mathcal{O}\left(\frac{\varepsilon_s}{\gamma_s}\right)$. Hence, exact equilibrium computation is not essential. It suffices to solve the regularized game to accuracy $\varepsilon_s \le c\, d\gamma_s$ for a sufficiently small universal constant $c > 0$, so that the additional perturbation can be absorbed into the $\mathcal{O}(d)$ aggregate-stability bound. Moreover, the approximate first-order conditions contribute an additive $\mathcal{O}(\varepsilon_s)$ term to the final duality-gap bound, and therefore any $\varepsilon_s = \mathcal{O}(d\gamma_s)$ preserves the convergence rate in Theorem 4.1.

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 claims made in the abstract and introduction accurately reflect the paper's contributions and scope. See Section 3, Section 4, Section 5 for details.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: See discussions in Section 2 and Section 4.1.

   Guidelines:

   - The answer [N/A] 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: See Section 2 and Section 4 for details.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: See Section 5 and the supplementary material.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: See the supplemental material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 5 for experimental details.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: See Section 5 for details.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: See Section 5 for details.

   Guidelines:

   - The answer [N/A] 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: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] 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: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: See Section 5 for details.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] 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: [N/A]

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

    Guidelines:

    - The answer [N/A] 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: [N/A]

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

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

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

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.