# NON-ASYMPTOTIC LAST-ITERATE CONVERGENCE FOR SCALE-INVARIANT REGRET-MATCHING ALGORITHMS IN MATRIX GAMES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Regret matching (RM) is a widely used no-regret learning paradigm for computing a Nash Equilibrium (NE) in games, best known for its parameter-free and scale-invariant properties. However, these desirable properties have so far been difficult to reconcile with last-iterate convergence: classical scale-invariant RM variants may exhibit cycling, whereas existing convergent RM methods typically introduce additional step size or smoothing parameters to stabilize the learning dynamics, thereby sacrificing scale invariance. In this work, we take a step toward closing this gap by introducing a scale-invariant RM-type algorithm, termed *Radial IREG-PRM*$^+$, with non-asymptotic last-iterate convergence guarantees in two-player zero-sum matrix games. The key idea is to apply a deterministic radial expansion with exponent $\alpha \in (0, 1/2)$ to the corrected regret vector while leaving the induced strategy unchanged. For arbitrary bounded utility sequences, Radial IREG-PRM$^+$ achieves the optimal $O(\sqrt{T})$ adversarial cumulative regret for every fixed $\alpha$, without prior knowledge of either the horizon or the utility scale. In self-play, a radially weighted RVU-bound analysis further yields an $O(T^{-1+\alpha})$ average-iterate convergence rate, which can be made arbitrarily close to the optimal rate $O(T^{-1})$. More importantly, we establish an explicit, instance-dependent stretched-exponential last-iterate convergence rate of $O(e^{-cT^{1-2\alpha}})$ for some game-dependent constant $c > 0$. We complement this fast instance-dependent guarantee with a uniform worst-case lower bound of $\Omega(T^{-1+\alpha})$ over normalized matrix games, together with a matching $O(T^{-1+\alpha})$ upper bound for the case in which exactly one player never gets non-zero cumulative regrets. Finally, we extend Radial IREG-PRM$^+$ to the single-sample bandit-feedback setting by using regularization technique, and establish an in-expectation last-iterate convergence rate of $O(T^{-1/10+\varepsilon})$ for any $\varepsilon \in (0, 1/10)$. Experiments on four instances of matrix games further validate the competitiveness of Radial IREG-PRM$^+$ over existing baselines.

## 1 INTRODUCTION

No-regret learning provides one of the fundamental algorithmic frameworks for equilibrium computation in games, and has been widely in building superhuman AI agents in large-scale poker Moravčík et al. (2017); Brown & Sandholm (2018; 2019). Within this framework, regret matching (RM) (Hart & Mas-Colell, 2000) is one of the most influential algorithmic paradigms. Regret-matching algorithms update the next mixed strategy by assigning probabilities in proportion to the positive cumulative regret of each action, where the regret measures how much better that action would have performed than the learner's realized play. The appeal of such algorithms lies not only in their simplicity, but also in their parameter-free and scale-invariant nature. In particular, rescaling all utility vectors by an arbitrary positive constant leaves the induced strategy unchanged. Consequently, RM-type methods do not require prior knowledge of the payoff scale or a carefully tuned learning rate. This parameter-free behavior is also practically important in solving large-scale sequential games, where scale invariance has been identified as a key property underlying strong empirical performance (Bowling et al., 2015; Orabona & Pál, 2018; Chakrabarti et al., 2024).

Over the past two decades, RM has evolved into a broad family of practically and theoretically successful algorithms. Regret Matching$^+$ (RM$^+$) (Tammelin et al., 2015) improves the empirical



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

performance of classical RM and becomes a key component of modern game-solving pipelines. Predictive Regret Matching$^+$ (PRM$^+$) (Farina et al., 2021) further incorporates optimism technique Chiang et al. (2012); Rakhlin & Sridharan (2013b) to predict the next utilities for convergence speed up. Both of the above two scale-variant RM variants guarantee an $O(T^{-1/2})$ average-iterate convergence rate. More recently, scale-invariant RM-type algorithms such as IREG-PRM$^+$ (Zhang et al., 2026) and MI-SPRM$^+$ (Meng et al., 2026) achieve fast average-iterate convergence $O(T^{-1})$ while retaining strong empirical performance in both matrix games and extensive-form games.

A major limitation of average-iterate convergence, however, is that it requires averaging over past strategies and provides no guarantee that the strategy played at the current iteration itself converges. Indeed, classical scale-invariant RM variants such as RM$^+$ and PRM$^+$ may exhibit persistent cycling in their last iterates (Lee et al., 2021; Cai et al., 2025a). Recent works have therefore sought to modify regret-matching dynamics so as to obtain last-iterate convergence. For example, mirror-descent-inspired RM methods such as Extra-gradient RM$^+$ (ExRM$^+$) and Smooth Predictive RM$^+$ (SPRM$^+$) admit asymptotic last-iterate guarantees Meng et al. (2025), and their restarting versions can further yield explicit linear convergence rates (Cai et al., 2025a). While effective in achieving convergence, these modifications generally depart from the scale-invariant property of classical regret matching and require careful tuning of the learning rate. In addition, the restarting mechanism needs to verify whether the current strategy satisfies the best-iterate condition at each iteration Meng et al. (2025).

This tension motivates the central question of this work:

> *Can we design a regret-matching-type algorithm that preserves scale invariance while achieving last-iterate convergence, even at a non-asymptotic rate and without restarting?*

We answer this question affirmatively for two-player zero-sum matrix games by introducing *Radial IREG-PRM$^+$*. The algorithm builds on IREG-PRM$^+$ and modifies only the magnitude of its nonnegative corrected regret vector through a deterministic radial expansion, while preserving the strategy induced by normalization. This radial mechanism is designed to gradually strengthen the stability of the learning dynamics without introducing a payoff-dependent learning rate or restarting condition. In summary, our contributions for matrix games are summarized as follows.

1. **Optimal adversarial regret.** For arbitrary bounded utility sequences, Radial IREG-PRM$^+$ preserves scale invariance and achieves the optimal-order adversarial cumulative regret $O(\sqrt{T})$, for every fixed $\alpha \in (0, 1/2)$, without prior knowledge of the horizon or the utility scale.
2. **Faster average-iterate convergence.** We further establish a radially weighted regret bounded by variation in utilities (RVU) for the proposed algorithm, leading to $O(T^{-1+\alpha})$ average-iterate convergence to a Nash equilibrium, which with a small $\alpha$ is close to the optimal rate $O(T^{-1})$.
3. **Explicit last-iterate convergence.** By a Lyapunov function argument, we prove that, for every fixed matrix game, Radial IREG-PRM$^+$ enjoys an anytime instance-dependent stretched-exponential last-iterate convergence rate $O(e^{-ct^{1-2\alpha}})$, for some game-dependent constant $c > 0$.
4. **Uniform worst-case lower bound.** We complement the fast instance-dependent guarantee with a horizon-fixed uniform lower bound $\Omega(T^{-1+\alpha})$ over normalized matrix games. Moreover, in the matrix games where exactly one player's cumulative regret remains identically zero, we prove a matching instance-independent upper bound $O(T^{-1+\alpha})$.
5. **Bandit-feedback extension.** We further extend the radial regret-matching framework to the single-sample bandit-feedback setting. By combining the radial update and vanishing regularization, we obtain an in-expectation last-iterate convergence rate $O(t^{-1/10+\varepsilon})$ for every $\varepsilon \in (0, 1/10)$.
6. **Empirical evaluation.** Experiments on four matrix-game instances demonstrate competitive last-iterate performance of Radial IREG-PRM$^+$ relative to existing baselines.

## 2 Related Works

**Regret Matching Algorithm.** Existing regret matching (RM) algorithms can be categorized into two main groups: **(1)** Scale-invariant regret matching algorithms, including the celebrated methods RM/RM$^+$/PRM$^+$ (Hart & Mas-Colell, 2000; Tammelin et al., 2015; Farina et al., 2021) and the modern variants IREG-PRM$^+$ Zhang et al. (2026), MI-SPRM$^+$ Meng et al. (2026), only have average-iterate convergence guarantee or best-iterate convergence. Even worse, (Lee et al., 2021; Cai et al., 2025a) demonstrate theoretically or empirically that the divergence of last-iterate strategy of

RM$^+$/PRM$^+$. **(2)** Two mirror-descent-inspired regret matching algorithms, namely Extragradient RM$^+$ (ExRM$^+$) and smooth predictive RM$^+$ (SPRM$^+$) that achieve fast average-iterate convergence rate $O(1/T)$ (Farina et al., 2023), have been demonstrated to admit asymptotic last-iterate convergence without explicit rate (Cai et al., 2025a). Their restarting versions yield an anytime linear last-iterate convergence rate, but need to check whether the distance between predicted and corrected strategies is small enough at each iteration (Cai et al., 2025a). Using the adaptive online mirror descent formulation (Liu et al., 2022) of smooth RM$^+$, Meng et al. (2025) propose Smooth Optimistic Gradient Based RM+ (SOGRM+) with asymptotic last-iterate convergence guarantee. Nevertheless, due to the introduction of an additional learning rate to ensure algorithmic stability, the second group of RM algorithms is not parameter-free and has no scale-invariance property, which has been demonstrated as the key to the success of regret matching algorithms in solving extensive-form games (Chakrabarti et al., 2024). To the best of our knowledge, IREG PRM$^+$ in Algorithm 1 is the first scale-invariant regret matching algorithm with non-asymptotic last-iterate convergence guarantee.

**Last-iterate Convergence in matrix games.** The seminal works (Bailey & Piliouras, 2018; Cheung & Piliouras, 2019) rigorously demonstrate that the celebrated online learning algorithms, such as multiplicative weight update (MWU) and its generalized version follow-the-regularized-leader (FTRL), may diverge in certain matrix games even with decreasing step sizes. (Daskalakis et al., 2018) show that ordinary gradient descent (GD) with constant step size will diverge even in the simple game of identity payoff matrix. These negative results motivate the analysis of optimistic and extragradient methods, like optimistic MWU (OMWU) Daskalakis & Panageas (2019) and optimistic gradient descent ascent (OGDA) (Daskalakis & Panageas, 2018; Mertikopoulos et al., 2019), for which asymptotic last-iterate convergence is established in bilinear and constrained zero-sum problems. Subsequent works develop non-asymptotic guarantees, establishing tight $O(T^{-1/2})$ last-iterate convergence rates for extragradient (EG) Golowich et al. (2020b) and optimistic-gradient (OG) methods (Golowich et al., 2020a; Cai et al., 2022) in smooth monotone games, which includes two-player zero-sum games as a special case. (Wei et al., 2021) further show that OMWU achieves (instance-dependent) linear last-iterate convergence under the unique Nash equilibrium (NE) assumption, and OGDA yields a (instance-dependent) sublinear convergence rate without NE uniqueness assumption. In this work, we also provide uniform worst-case lower bound $\Omega(T^{-1+\alpha})$ with $\alpha \in (0, 1/2)$ on the last-iterate convergence for the proposed radial IREG-PRM$^+$ algorithm, followed by a matching upper bound in the cases where exactly one player never gets positive regrets.

## 3 PRELIMINARY

For a positive integer $d$, let $[d] = \{1, \dots, d\}$ and $\Delta_d = \{p \in \mathbb{R}^d_+ : \sum_{j=1}^d p_j = 1\}$ denote the probability simplex over $d$ actions. For a matrix $A \in \mathbb{R}^{m\times n}$, assume that its maximum element satisfies $\max_{i,j} |A_{ij}| \le B$, and its operator norm $L = \|A\|_{\mathrm{op}}$. Let $D = \max\{m, n\}$. For any nonzero $w \in \mathbb{R}^d_+$, define the $\ell_1$ normalization $\mathcal{N}(w) = \frac{w}{\|w\|_1}$, and $\ell_2$ normalization $\phi(w) = \frac{w}{\|w\|_2}$.

**Matrix Games with Gradient Feedback.** Consider a two-player zero-sum matrix game (Neumann, 1928) with payoff matrix $A \in \mathbb{R}^{m\times n}$, where $m$ is the number of actions available to the row player and $n$ is the number of actions available to the column player. For a pure-action pair $(a, b) \in [m]\times[n]$, the entry $A_{ab}$ is the payoff to the row player (and $-A_{ab}$ is the payoff to the column player). The row player chooses a mixed strategy $x \in \Delta_m$ to maximize the payoff, whereas the column player chooses a mixed strategy $y \in \Delta_n$ to minimize it. Consequently, the expected payoff to the row player under the mixed-strategy profile $z = (x, y) \in Z = \Delta_m \times \Delta_n$ is

$$U(x, y) = \mathbb{E}_{a\sim x,\, b\sim y}[A_{ab}] = x^\top A y. \tag{1}$$

A profile $z^\star = (x^\star, y^\star) \in Z^\star \subset Z$ is a *Nash equilibrium* if neither player can improve its utility by deviating unilaterally, i.e., $x^\star \in \arg\max_{x\in\Delta_m} x^\top A y^\star, \quad y^\star \in \arg\min_{y\in\Delta_n} (x^\star)^\top A y$. We measure the proximity of a profile $(x, y)$ to the equilibrium using the exploitability gap

$$\mathrm{Gap}(x, y) = \max_{p\in\Delta_m} p^\top A y - \min_{q\in\Delta_n} x^\top A q. \tag{2}$$

The gap $\mathrm{Gap}(x, y) \ge 0$, with $\mathrm{Gap}(x, y) = 0$ if and only if $(x, y) \in Z^\star$. To express the two players' updates in a common utility-maximization form, set $d_x = m$, $d_y = n$, $z_x = x$, and $z_y = y$, and define the joint utility operator $F(z) = (U_x(z), U_y(z)) = \big(Ay, -A^\top x\big)$. Then, the exploitability gap in Eq. (2) admits the following variational characterization: $\mathrm{Gap}(z) = \max_{u\in Z} \langle F(z), z - u\rangle$.

**Algorithm 1** Radial IREG-PRM$^+$

**Require:** full-support anchor $a_i \in \mathrm{ri}(\Delta_{d_i})$, $\alpha \le \frac{1}{2}$

Initialize $R_i^1 \leftarrow 0$ and $u_i^0 \leftarrow 0$
**for** $t = 1, 2, \ldots$ **do**
  $S_i^t \leftarrow \|R_i^t\|_2$
  **if** $S_i^t > 0$ **then** {**Positive Branch**}
    $\widetilde{z}_i^t \leftarrow \mathcal{N}(R_i^t)$ and $m_i^t \leftarrow u_i^{t-1}$
    Find the unique $\gamma_i^t \in \mathbb{R}$ such that $Q_i^t \leftarrow [R_i^t + m_i^t - \gamma_i^t \mathbf{1}]_+, \|Q_i^t\|_2 = S_i^t$
    $z_i^t \leftarrow \mathcal{N}(Q_i^t)$ and $\delta_i^t \leftarrow \langle m_i^t, z_i^t \rangle - \gamma_i^t$
  **else** {**Zero Branch**}
    $\widetilde{z}_i^t \leftarrow a_i$, $z_i^t \leftarrow a_i$, $m_i^t \leftarrow 0$, $Q_i^t \leftarrow 0$
    Leave $\gamma_i^t$ and $\delta_i^t$ undefined
  **end if**
  Play $z_i^t$ and observe utility vector $u_i^t$
  $e_i^t \leftarrow u_i^t - m_i^t$, $E_i^t \leftarrow e_i^t - \langle e_i^t, z_i^t \rangle \mathbf{1}$
  $H_i^t \leftarrow [Q_i^t + E_i^t]_+$
  $\lambda_t \leftarrow ((t+1)/t)^\alpha$ and $R_i^{t+1} \leftarrow \lambda_t H_i^t$
**end for**

**Algorithm 2** IX–Regularized Radial IREG-PRM$^+$

**Require:** full-support anchor $a_i \in \mathrm{ri}(\Delta_{d_i})$, $\varepsilon > 0$

Initialize $R_i^1 \leftarrow a_i / \|a_i\|_2$, $\widetilde{u}_i^0 \leftarrow 0$
**for** $t = 1, 2, \ldots$ **do**
  $S_i^t \leftarrow \|R_i^t\|_2$
  $\epsilon_t \leftarrow t^{-\frac{1}{5}+2\varepsilon}, \theta_t \leftarrow t^{-\frac{2}{5}+4\varepsilon}, \tau_t \leftarrow t^{-\frac{1}{10}+\varepsilon}$
  $\widetilde{z}_i^t \leftarrow \mathcal{N}(R_i^t)$ and $m_i^t \leftarrow \widetilde{u}_i^{t-1}$
  Find the unique $\gamma_i^t$ such that $Q_i^t \leftarrow [R_i^t + m_i^t - \gamma_i^t \mathbf{1}]_+, \|Q_i^t\|_2 = S_i^t$
  $z_i^t \leftarrow \mathcal{N}(Q_i^t)$ and $p_i^t \leftarrow (1 - \epsilon_t) z_i^t + \epsilon_t a_i$
  Independently sample $I_t \sim p_x^t$, $J_t \sim p_y^t$ and observe $r_t = A_{I_t J_t}$
  Construct $\widehat{u}_x^t, \widehat{u}_y^t$ by Eq. 14
  $\widetilde{u}_i^t \leftarrow \widehat{u}_i^t - \tau_t (z_i^t - a_i)$
  $e_i^t \leftarrow \widetilde{u}_i^t - m_i^t$, $E_i^t \leftarrow e_i^t - \langle e_i^t, z_i^t \rangle \mathbf{1}$
  $H_i^t \leftarrow [Q_i^t + E_i^t]_+$
  $\lambda_t \leftarrow ((t+1)/t)^{7/10}$ and $R_i^{t+1} \leftarrow \lambda_t H_i^t$
**end for**

At round $t$ of online learning (Cesa-Bianchi & Lugosi, 2006), the players select $x^t \in \Delta_m$ and $y^t \in \Delta_n$, and we write $z^t = (x^t, y^t)$. Under *gradient feedback*, after playing $z_i^t$, player $i \in \{x, y\}$ observes its entire current utility vector $u_i^t = U_i(z^t)$; hence $u_x^t = Ay^t$ and $u_y^t = -A^\top x^t$ in self-play.

**Radial IREG-PRM$^+$ Algorithm.** Our Algorithm 1 is a prediction-correction regret-matching-type method that is mainly based on a recently proposed parameter-free RM algorithm, named increasing-regret extra gradient-based predictive regret matching$^+$ (IREG-PRM$^+$) (Zhang et al., 2026), that achieves optimal $O(1/T)$ average-iterate convergence and $O(1/\sqrt{T})$ best-iterate convergence. Concretely, the *radial regret vector* for player $i \in \{x, y\}$ is initialized by $R_i^1 = 0$. At the beginning of round $t$, let $S_i^t = \|R_i^t\|_2$. When $S_i^t = 0$, the algorithm avoids normalizing the zero vector and instead plays the anchor $a_i$, with $m_i^t = 0$ and the *predicted regret vector* $Q_i^t = 0$. If $S_i^t > 0$, the past-utility prediction $m_i^t = \mu_i^{t-1}$ is incorporated through an equal-$\ell_2$-norm positive-cone projection

$$Q_i^t = [R_i^t + m_i^t - \gamma_i^t \mathbf{1}]_+, \qquad \|Q_i^t\|_2 = S_i^t,$$

and the real output strategy is $z_i^t = \mathcal{N}(Q_i^t)$. After observing $u_i^t$, the algorithm computes the prediction residual $e_i^t = u_i^t - m_i^t$ and centers it with respect to the played strategy $z_i^t$: $E_i^t = e_i^t - \langle e_i^t, z_i^t \rangle \mathbf{1}$, which leads to $\langle E_i^t, z_i^t \rangle = 0$. This centered residual is then added to $Q_i^t$, and the resulting vector is projected onto the positive orthant to obtain *corrected regret vector*: $H_i^t = [Q_i^t + E_i^t]_+$. Until now, the implementation of Algorithm 1 is almost the same as IREG-PRM$^+$ (Zhang et al., 2026). Then, IREG-PRM$^+$ further sets $R_i^{t+1} = H_i^t$. To achieve last-iterate convergence for IREG-PRM$^+$, we construct a contraction inequality (see Lemma A.3) for the case $S_i^t > 0$:

$$P(H_i^t, z_i^*) - P(R_i^t, z_i^*) \le -\delta_i^t + \frac{\|E_i^t\|_2^2}{2S_i^t} \lesssim -\big(S_i^t - \frac{L^2}{S_i^t}\big)\|z^t - z^{t-1}\|^2, \tag{3}$$

where $P(r, p) = \|r\|_2 \|p\|_2 - \langle r, p \rangle$ is used as basic potential function. To ensure potential descent, which is essential for establishing an explicit last-iterate convergence rate (e.g. see (Facchinei & Pang, 2003)), the right-hand side of Eq. 3 must be non-positive, i.e. $S_i^t \gtrsim L$. However, IREG-PRM$^+$ only guarantees that $\{S_i^t\}_{t \ge 1}$ is increasing, but provides no information about its magnitude.

To solve this issue, we take the radial expansion $R_i^{t+1} = \lambda_t H_i^t$ with radial schedule $g_t = t^\alpha, \lambda_t = \frac{g_{t+1}}{g_t} = \big(\frac{t+1}{t}\big)^\alpha$, which finally leads to an explicit magnitude $S_i^t = O(g_t)$ in Lemma A.5 and potential

descent inequality in Eq. 11. For later use, define the accumulated contraction kernel

$$\mathcal{H}_\alpha(t,s) := \int_s^t r^{-2\alpha}\,dr = \begin{cases} \dfrac{t^{1-2\alpha}-s^{1-2\alpha}}{1-2\alpha}, & 0<\alpha<\frac{1}{2}, \\ \log\dfrac{t}{s}, & \alpha=\frac{1}{2}, \end{cases} \qquad t\ge s\ge 1. \tag{4}$$

The range $0 < \alpha \le 1/2$ is exactly the range covered by the present proof framework: $\alpha > 0$ guarantees automatic stabilization because $g_t \to \infty$, whereas $\alpha \le 1/2$ guarantees $\sum_t g_t^{-2} = \infty$, which is necessary for the multiplicative Lyapunov contractions to accumulate to zero (Lemma 5.4). We next define *activation time* and discuss its relation to Algorithm 1. This notion plays an essential role in establishing the last-iterate convergence of the Radial IREG-PRM$^+$ algorithm.

**Definition 3.1** (Activation time)**.** *For any player $i \in \{x, y\}$, the activation time is the first time that the radial regret vector $R_i^t$ is nonzero, $\tau_i = \inf\{t \ge 1 : S_i^t > 0\} \in \{2, 3, \ldots\} \cup \{\infty\}$.*

**Lemma 3.2** (Once-active-always-active property)**.** *In Algorithm 1, if $S_i^t = 0$, then $H_i^t \ne 0$ if and only if $u_i^t$ is not a constant vector. If $S_i^t > 0$, then $S_i^{t+1} > 0$ and $\sigma_i^{t+1} = \frac{S_i^{t+1}}{g_{t+1}} \ge \frac{S_i^t}{g_t} = \sigma_i^t$.*

Lemma 3.2 shows that each player $i$ in Algorithm 1 can be in only one of two cases: never active, $\tau_i = \infty$; or activated at some finite time and thereafter remaining in the positive branch, $\tau_i < \infty$. Define $\tau = \max(\tau_x, \tau_y)$, if $\tau < \infty$, then for any $t \ge \tau$, we have $\overline{\sigma} \ge \sigma_i^t \ge \underline{\sigma} \triangleq S_i^\tau / g_\tau$ (see Lemma C.4). Define $C_d = 1 + \sqrt{d}$, $c_d = 1/[2\sqrt{d}(1+\sqrt{d})^2]$ that will be used through the remaining paper.

## 4 Average-Iterate Convergence with Scale Invariance

In this section, we provide average-iterate convergence guarantee and scale-invariance property for Algorithm 1. Define the *instantaneous regret vector* for player $i$ as $r_i^t = u_i^t - \langle u_i^t, z_i^t\rangle \mathbf{1}$.

**Theorem 4.1** (Adversarial regret)**.** *If $\|u_i^t\|_2 \le G_i$ (e.g. set $G_i = \|A\|_{\mathrm{op}}$), then the radial IREG-PRM$^+$ algorithm need not know $G_i, T$, and the cumulative regret of player $i$ is upper bounded as*

$$\mathrm{Reg}_i(T) = \max_j \sum_{t=1}^T r_{i,j}^t \le S_i^{T+1} \le \begin{cases} \frac{2C_{d_i}G_i}{\sqrt{1-2\alpha}}(T+1)^\alpha T^{\frac{1-2\alpha}{2}}, & 0<\alpha<1/2. \\ 2C_{d_i}G_i(T+1)^{1/2}\sqrt{1+\log T}, & \alpha=1/2. \end{cases} \tag{5}$$

*Furthermore, the strategy trajectory is unchanged under $u_i^t \mapsto c u_i^t$ for any $c > 0$.*

The above result shows that, analogous to traditional regret matching-type algorithms like RM$^+$ (Tammelin et al., 2015) and PRM$^+$ (Farina et al., 2021), the radial IREG-PRM$^+$ algorithm has scale-invariance property and sublinear cumulative regret $\widetilde{\mathcal{O}}(\sqrt{T})$. Using a more delicate analysis, we can obtain the following weighted RVU regret bound for some $\beta > 0$ (see Lemma B.4):

$$\sum_{t=1}^T \frac{1}{g_t}\langle u_i^t, p - z_i^t\rangle \le O\Big(\beta + \sum_{t=1}^T \frac{\|u_i^t - m_i^t\|_2^2}{g_t^2} - \sum_{t=2}^T \|z_i^t - z_i^{t-1}\|_2^2\Big)\mathbf{1}(\tau<\infty).$$

Using the above RVU regret bound, we achieve the following $T^{-(1-\alpha)}$ average-iterate convergence rate with $\alpha \in (0, 1/2]$ for uncoupled Algorithm 1 in two-player zero-sum games.

**Theorem 4.2** ($T^{-(1-\alpha)}$ average-iterate convergence)**.** *In any two-player zero-sum matrix game, if both players run Algorithm 1 with gradient feedback, then there exist finite game-dependent constant $\mathsf{C}_{x,y}$ and integer $t_\star \in \mathbb{N}$ such that, for any $T \ge t_\star$, we have*

$$\mathrm{Gap}\Big(\frac{1}{T}\sum_{t=1}^T x^t, \frac{1}{T}\sum_{t=1}^T y^t\Big) \le \frac{(\mathsf{C}_{x,y})(T+1)^\alpha}{T} = O(T^{-(1-\alpha)}). \tag{6}$$

For sufficiently small $\alpha > 0$ (e.g., $\alpha = 0.01$), the average-iterate convergence rate of Algorithm 1 is $O(T^{-1+\alpha})$, which is close to the optimal rate $O(T^{-1})$. This optimal average-convergence rate has recently been attained by two scale-invariant regret-matching algorithms (Zhang et al., 2026; Meng et al., 2026) for matrix games. In the following sections, we further show that the scale-invariant Algorithm 1 also achieves non-asymptotic last-iterate convergence in matrix games.

## 5 INSTANCE-SPECIFIC LAST-ITERATE CONVERGENCE

We first investigate the matrix games with gradient feedback, and present the instance-specific last-iterate convergence rate for Algorithm 1 as follows. Combining this with Eq. 4 reveals that, when $\alpha \in (0, 1/2)$, the duality gap of last-iterate strategy in Radial IREG-PRM$^+$ vanishes at a sub-exponential rate $O(e^{-t^{1-2\alpha}})$; when $\alpha = 1/2$, the rate becomes sublinear $O(t^{-1})$.

**Theorem 5.1.** *For any two-player zero-sum game, there exist game-dependent constants $C, T_\star, \mu > 0$ such that the current self-play strategy $z^t$ of Algorithm 1 converges to $Z^\star$ for any $\alpha \in (0, \frac{1}{2}]$:*

$$\mathrm{Gap}(z^t) \le C \exp\Big[-\frac{\mu}{2}\mathcal{H}_\alpha(t, T_\star)\Big], \qquad t \ge T_\star. \tag{7}$$

For the reader's convenience, we next present the proof of the convergence rate for the case where both players are activated at some finite time, i.e., $\tau < \infty$. The proofs for the case in which at least one player is never activated follow a similar argument and therefore deferred to Appendix D. Specifically, we first choose any integer $T_0$ satisfying $T_0 \ge \tau + 1$ and $T_0^{2\alpha} \ge \frac{8\sqrt{D}C_D^4 L^2}{\underline{\sigma}^2}$. Since $\alpha > 0$, we have $t^{2\alpha} \to \infty$, such that a finite $T_0$ always exists whenever both players activate in finite time. Then, the most critical step in the proof is to construct the *Lyapunov potential function* as

$$\Phi_t = \Psi_t^* + \frac{1}{2}\mathcal{V}_{t-1}, \qquad t \ge T_0. \tag{8}$$

where $\Psi_t^* = \min_{z^\star \in Z^\star} \Psi_t(z^\star)$, the function $\Psi_t(z) = \frac{1}{g_t}\sum_{i\in\{x,y\}} P(R_i^t, z_i)$ is the weighted version of Lyapunov potential function in Eq. 3, and $\mathcal{V}_t = \sum_{i\in\{x,y\}} \|E_i^t\|_2^2/(g_t S_i^t)$ also corresponds to the weighted version of $\|E_i^t\|_2^2/S_i^t$ in Eq. 3. Lyapunov function always serves as the upper bound on the exploitability gap of strategy profile (Pérolat et al., 2021). Therefore, it remains to analyze the upper bound on each term of Lyapunov function in Eq. 8. The following lemma shows that the minimum $\Psi_t^*$ has the same order of the distance between intermediate strategy profile $\widetilde{z}^t$ and the set $Z^*$ of NEs.

**Lemma 5.2.** *If $\sigma_i^t = S_i^t/g_t \in [\underline{\sigma}, \overline{\sigma}]$, then $c_D\underline{\sigma}\ \mathrm{dist}(\widetilde{z}^t, Z^\star)^2 \le \Psi_t^* \le 2D\overline{\sigma}\ \mathrm{dist}(\widetilde{z}^t, Z^\star)^2$.*

The upper bound on $\Psi_t^*$ in Lemma 5.2 implies that we can pay more attention to bounding the distance $\mathrm{dist}(\widetilde{z}^t, Z^\star)$ between intermediate strategy profile $\widetilde{z}^t$, instead of the real strategy profile $z^t$, to the NE set $Z^*$. Then, using the equivalence between the duality gap and Euclidean distance $\mathrm{Gap}(\widetilde{z}) = O(\mathrm{dist}(\widetilde{z}, Z^*))$ in Lemma C.5, we next pay more attention to the duality gap $\mathrm{Gap}(\widetilde{z}^t)$ and give an upper bound on the duality gap based on the prediction error $a_t = \|z^t - \widetilde{z}^t\|_2$.

**Lemma 5.3.** *Define $K_0 \triangleq 4\sqrt{2D}\overline{\sigma} + 2L$, then for any $t \ge T_0$, $\mathrm{Gap}(\widetilde{z}^t) \le K_0 g_t a_t$.*

Besides the above two lemmas, the following descent inequality of Lyapunov function $\Phi_t$ plays an important role in establishing the Lyapunov contraction inequality and the last-iterate convergence.

**Lemma 5.4** (Descent inequality of Lyapunov function)**.** *Define $\overline{\Delta}_t = \frac{\delta_x^t + \delta_y^t}{g_t}$. Then, for $t \ge T_0$,*

$$\Phi_{t+1} \le \Phi_t - \frac{1}{2}\overline{\Delta}_t - \frac{1}{4}\mathcal{V}_{t-1}, \quad \textit{and} \quad \overline{\Delta}_t \ge c_D\underline{\sigma}\, a_t^2. \tag{9}$$

Starting from the descent inequality of Lyapunov function, and applying the above three useful lemmas, we can achieve the following game-dependent/instance-dependent last-iterate convergence rate for Algorithm 1 in the case where both players are activated in the finite time $\tau < \infty$.

**Theorem 5.5** (Last-iterate convergence in the positive-state case)**.** *Assume that there exists a finite time $T_0$ such that, for every $t \ge T_0$, both players are active, i.e., $S_x^t > 0, S_y^t > 0$. Then the actual iterate sequence of Algorithm 1 satisfies the explicit last-iterate convergence bound*

$$\Phi_t \le \Phi_{T_0} \exp[-\mu_A \mathcal{H}_\alpha(t, T_0)], \qquad \mathrm{Gap}(z^t) \le C_A \exp\Big[-\frac{\mu_A}{2}\mathcal{H}_\alpha(t, T_0)\Big]. \tag{10}$$

*where $\mu_A$ and $C_A$ are instance-dependent and given in Theorem C.6.*

*Proof Sketch.* By Lemmas 5.2 and 5.3, we can bound the Lyapunov function for some $K_1 \ge 1$ as

$$\Phi_t = \Psi_t^* + \frac{1}{2}\mathcal{V}_{t-1} \le 2D\overline{\sigma}\,\mathrm{dist}(\widetilde{z}^t, Z^\star)^2 + \frac{1}{2}\mathcal{V}_{t-1} \le K_1 g_t^2 a_t^2 + \frac{1}{2}\mathcal{V}_{t-1} \le K_1 g_t^2 (a_t^2 + \mathcal{V}_{t-1}).$$

Thus $a_t^2 + \mathcal{V}_{t-1} \geq \frac{\Phi_t}{K_1 g_t^2}$. Then, by the potential-descent relation in Eq. 9, there exists some $\lambda_0 > 0$,

$$\Phi_{t+1} \leq \Phi_t - \lambda_0(a_t^2 + \mathcal{V}_{t-1}) \leq \Phi_t - \frac{\lambda_0}{K_1 g_t^2}\Phi_t = \Big(1 - \frac{\lambda_0}{K_1 g_t^2}\Big)\Phi_t. \tag{11}$$

Recall that $\mu_A = \min\left\{\frac{1}{2}, \frac{\lambda_0}{K_1}\right\}$. Since $g_t^2 = t^{2\alpha}$, by recursion we obtain the following expression

$$\Phi_{t+1} \leq \Big(1 - \frac{\mu_A}{t^{2\alpha}}\Big)\Phi_t \leq \Phi_{T_0} \exp\Big(-\mu_A \sum_{s=T_0}^{t-1} s^{-2\alpha}\Big) \leq \Phi_{T_0} \exp[-\mu_A \mathcal{H}_\alpha(t, T_0)]\,.$$

Next, by the lower bound on Lyapunov function, $\Phi_t \geq \Psi_t^* \geq c_D \underline{\sigma}\, \mathrm{dist}(\widetilde{z}^t, Z^\star)^2$, and Lemma 5.2,

$$\mathrm{dist}(z^t, Z^\star) \leq \|z^t - \widetilde{z}^t\|_2 + \mathrm{dist}(\widetilde{z}^t, Z^\star) = a_t + \mathrm{dist}(\widetilde{z}^t, Z^\star) \leq \Big(\frac{1}{\sqrt{\lambda_0}} + \frac{1}{\sqrt{c_D \underline{\sigma}}}\Big)\sqrt{\Phi_t}.$$

Using again the equivalence between the duality gap and Euclidean distance $\mathrm{Gap}(z) = O(\mathrm{dist}(z, Z^*))$ in Lemma C.5, gives the final upper bound on $\mathrm{Gap}(z^t)$. □

Cai et al. (2025a) establish the first last-iterate convergence result for two non-scale-invariant RM algorithms, ExRM$^+$ and SPRM$^+$, without explicit rates, via the Minty condition on the regret operator. They further show that, when combined with a restarting mechanism, these two algorithms attain a linear last-iterate convergence rate. Another algorithm, SOGRM$^+$ (Meng et al., 2026), also relies on a restarting mechanism to achieve a linear convergence rate. However, the restarting technique requires checking whether the current strategy satisfies the best-iterate convergence condition at each iteration. In contrast, our Algorithm 1 keeps track of the radial regret vector $R_i^t$ to construct an increasing regret term, without normalizing $R_i^t$ to the simplex for subsequent optimization (as in the restarting mechanism). Besides, unlike prior works that rely on best-iterate convergence results, our analysis uses a Lyapunov function argument to directly obtain last-iterate convergence for Algorithm 1.

## 6 Uniform Worst-Case Lower Bound for Last-Iterate Convergence

When $\alpha \in (0, 1/2)$, Theorems 5.1 and 5.5 establish fast instance-specific last-iterate convergence for Algorithm 1. However, the constants and burn-in times in these guarantees may depend on the underlying payoff matrix; in particular, they may deteriorate when instance-specific quantities such as $S_i^\tau > 0$ become small. Thus, these results do not by themselves yield a convergence guarantee that is uniform over the entire class of normalized matrix games. In this section, we therefore study the uniform worst-case last-iterate complexity of Algorithm 1. Specifically, we ask whether the instance-specific guarantees above can be strengthened to a stretched-exponential rate whose constants are independent of the payoff matrix. The following result gives a negative answer.

**Theorem 6.1** (Uniform $\Omega(T^{-(1-\alpha)})$ worst-case lower bound for the last-iterate)**.** *Consider Algorithm 1 with a fixed exponent $\alpha \in (0, 1/2]$, and fix full-support uniform anchors $a_x = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, $a_y = \left(\frac{1}{2}, \frac{1}{2}\right)$. Then, for every integer $T \geq 2$, there exists a normalized zero-sum matrix $A_T \in \left\{A \in \mathbb{R}^{3\times 2} : |A_{jk}| \leq 1\right\}$ such that the current last iterate of Algorithm 1 satisfies*

$$\mathrm{Gap}_{A_T}(z_{A_T}^T) \geq \frac{1-\alpha}{16}\, T^{-(1-\alpha)}. \tag{12}$$

Theorem 6.1 guarantees that $\sup_{\|A\|_\infty \leq 1} \mathrm{Gap}_A(z_A^T) \geq \Omega(T^{-1+\alpha})$, establishing a minimax lower bound on the uniform worst-case last-iterate gap of Algorithm 1. Consequently, for every fixed $\alpha > 0$, the algorithm does not admit a game-independent $O(T^{-1})$ last-iterate guarantee over normalized matrix games. This does not contradict the faster instance-specific convergence established in Section 5, because the hard instance $A_T$ may depend on the horizon $T$. In comparison, $O(T^{-1})$ last-iterate convergence is attainable and matches known first-order oracle lower bounds in broader smooth monotone saddle-point settings (Ouyang & Xu, 2021; Cai & Zheng, 2023). We next establish a matching instance-independent upper bound for the one-zero branch.

**Theorem 6.2** (Uniform $O(T^{-(1-\alpha)})$ upper bound on the last-iterate convergence rate with only one player activated)**.** *Assume the matrix payoff satisfies $|A_{jk}| \leq B$. If exactly one player never activates, then for all $T \geq 2$, the current last iterate of Algorithm 1 satisfies the instance-independent bound*

$$\mathrm{Gap}(z^T) \leq \frac{2B\sqrt{D}\,(T+1)^\alpha}{T-1} = O(T^{-(1-\alpha)}). \tag{13}$$

The hard family constructed in the proof of Theorem 6.1 actually belongs to the one-zero branch: the column player never activates, whereas the row player activates at the first iteration. Therefore, combining Theorems 6.1 and 6.2, we conclude that the $T^{-(1-\alpha)}$ dependence is worst-case tight, up to constant factors, for Algorithm 1 on this branch. In addition, Example E.1 provides a fixed, horizon-independent payoff matrix in which running Algorithm 1 can exactly activate one player. Nevertheless, we should point out that at the current stage we are unable to derive an instance-independent upper bound in the case where both players are activated, i.e. $\tau < \infty$. The reason is that, in this one-zero branch, the current analysis uses Hoffman's error bound (Hoffman, 1952) to translate a Lyapunov decrease into last-iterate convergence. This yields a rate necessarily depending on the game-specific Hoffman constant, which may be prohibitively large.

## 7 Last-Iterate Convergence with Bandit Feedback

In this section, we extend Algorithm 1 to the matrix games with bandit feedback and derive non-asymptotic last-iterate convergence in expectation. We first introduce basic notations.

**Matrix Games with Bandit Feedback.** At each round $t \geq 1$, the players select sampling distributions $p_x^t \in \Delta_m$ and $p_y^t \in \Delta_n$ using only information from previous rounds. They then sample independently indices $I_t \sim p_x^t$ and $J_t \sim p_y^t$ and jointly observe the single scalar payoff $r_t = A_{I_t J_t}$. The single observation does not reveal the full expected utility vectors $Ap_y^t$ and $-A^\top p_x^t$. For $t \geq 1$, let $\mathcal{F}_t = \sigma\big(I_s, J_s, r_s : 1 \leq s \leq t\big)$ denote the $\sigma$-algebra generated by the observation history through round $t$. We write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. To estimate the unavailable utility vectors, we use the implicit-exploration (IX) estimators (Neu, 2015), with a positive offset $\theta_t$ trading a controlled IX bias for lower importance-weighting variance: for $j \in [m]$ and $k \in [n]$,

$$\widehat{u}_{x,j}^t = \frac{r_t \mathbf{1}\{I_t = j\}}{p_{x,j}^t + \theta_t}, \qquad \widehat{u}_{y,k}^t = \frac{-r_t \mathbf{1}\{J_t = k\}}{p_{y,k}^t + \theta_t}. \tag{14}$$

**Regularized Game.** Single-sample estimates remain noisy even as their bias vanishes, so it is hard to derive last-iterate convergence for Algorithm 1 with merely IX estimators. Thus, we further introduce an anchor-centered Tikhonov term (Abe et al., 2024) that provides a restoring force for the last iterate. Such technique is called *regularization* and has been widely applied to demonstrate the last-iterate convergence in games with bandit feedback (Cai et al., 2023; Zhao et al., 2025; Cai et al., 2025b). Let $a = (a_x, a_y)$ as the full-support anchor profile. For any $\tau > 0$, define the regularized utility operator

$$F_\tau(z) = F(z) - \tau(z - a), \tag{15}$$

which is equal to the gradient operator $F_\tau(z) = (\nabla_x \mathcal{L}_\tau(x,y), -\nabla_y \mathcal{L}_\tau(x,y))$ of strongly-convex-strongly-concave saddle function $\mathcal{L}_\tau(x,y) = x^\top A y - \frac{\tau}{2}\|x - a_x\|_2^2 + \frac{\tau}{2}\|y - a_y\|_2^2$. Thus, the regularized game has a unique saddle point $z_\tau^* = (x_\tau^*, y_\tau^*)$, characterized as the unique solution of $\langle F_\tau(z_\tau^*), z - z_\tau^* \rangle \leq 0, \quad \forall z \in Z$. The algorithm uses a time-varying parameter $\tau_t \downarrow 0$: the regularization stabilizes the dynamics while its perturbation of the original game vanishes asymptotically.

**IX–Regularized Radial IREG$^+$ Algorithm.** Algorithm 2 retains the radial prediction–correction structure of Algorithm 1 while adapting it to single-sample feedback. For each player $i \in \{x, y\}$, it starts from the positive radial regret vector $R_i^1 = a_i / \|a_i\|_2$, with $\widetilde{u}_i^0 = 0$. At round $t$, the prediction step produces an internal strategy $z_i^t$ and forms the actual sampling distribution $p_i^t = (1-\epsilon_t) z_i^t + \epsilon_t a_i$, and samples $I_t$ and $J_t$ as described above. After observing $r_t$, it first computes $\widehat{u}_i^t$ according to Eq. 14 and then constructs the regularized utility estimate $\widetilde{u}_i^t = \widehat{u}_i^t - \tau_t(z_i^t - a_i)$, which is used both to correct the current prediction and as the one-step prediction at the next round. Finally, fix a rate-slack parameter $\varepsilon \in (0, 1/10)$ and use the anytime schedules

$$g_t = t^\alpha, \quad \lambda_t = g_{t+1}/g_t, \quad \epsilon_t = t^{-\frac{1}{5}+2\varepsilon}, \quad \theta_t = t^{-\frac{2}{5}+4\varepsilon}, \quad \tau_t = t^{-\frac{1}{10}+\varepsilon}. \tag{16}$$

Here $\lambda_t$ control radial growth, $\epsilon_t$ controls explicit exploration, $\theta_t$ controls implicit exploration, and $\tau_t$ controls regularization. By setting $\alpha = 7/10$, we obtain the following rate, which is the fastest result achievable across different hyperparameter choices $\{g_t, \epsilon_t, \theta_t, \tau_t\}$ under the proof framework.

**Theorem 7.1** (Near-$t^{-1/10}$ expected last-iterate rate under bandit feedback)**.** *Fix any $\varepsilon \in (0, 1/10)$. For any fixed finite two-player zero-sum matrix $A$ and any full-support anchors $a_x, a_y$, Algorithm 2, run with the schedules in Eq. 16. There exist a finite constant $C_{A,\varepsilon} < \infty$ and an integer $T_{A,\varepsilon}$, depending only on $(A, a_x, a_y, m, n, \varepsilon)$, such that for all $t \geq T_{A,\varepsilon}$,*

$$\mathbb{E}[\mathrm{Gap}(z^t)] \leq C_{A,\varepsilon} t^{-1/10+\varepsilon}, \qquad \mathbb{E}[\mathrm{Gap}(p^t)] \leq C_{A,\varepsilon} t^{-1/10+\varepsilon}. \tag{17}$$

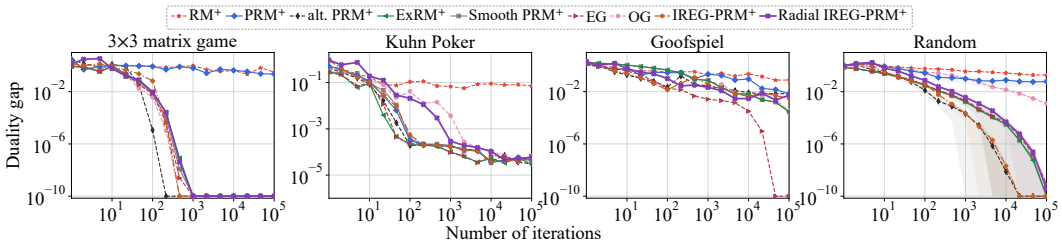


Figure 1: Last-iterate convergence of exploitability gap of our Algorithm 1 and other baselines.

In matrix games with bandit feedback, the first in-expectation last-iterate convergence has been established by Cai et al. (2023) for regularized mirror descent with entropy regularization, with an anytime rate $O(t^{-\frac{1}{6}} \ln t)$. Fiegel et al. (2025) further propose an uncoupled regularized EXP3 algorithm that achieves an in-expectation rate of $O(t^{-\frac{1}{4}} \ln t)$. Our convergence rate $O(t^{-1/10+\epsilon})$ in Theorem 7.1 is slower than previous results, and is still far from the optimal rate $\Omega(t^{-1/4})$ in terms of $L^1$-norm convergence (Fiegel et al., 2025, Theorem 5.1). Nevertheless, Algorithm 2 is the first RM-type algorithms in matrix games with bandit feedback with last-iterate convergence guarantee.

## 8 Experiments

In this section, we conduct experiments to verify the competitiveness of Algorithm 1 against existing methods, in terms of the exploitability/duality gap of last-iterate strategy. We follow the same experimental setting as in Cai et al. (2025a) and use four game instances. Concretely, we use the $3 \times 3$ matrix game instance from Farina et al. (2023), for which RM$^+$ and PRM$^+$ have been theoretically shown to fail to converge in the last iterate, the normal-form representations of Kuhn poker (Kuhn, 1950) and Goofspiel (Ross, 1971). In addition, we evaluate 25 randomly generated matrix games with dimensions $(m, n) = (10, 15)$, and report the average duality gap and the corresponding confidence intervals. For the algorithms that require a stepsize and initialized strategy (ExRM$^+$, Smooth PRM$^+$, EG and OG), we choose the best stepsizs $\eta \in \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ and initialize them at $(1/m\mathbf{1}_m, 1/n\mathbf{1}_n)$. We set the exponent $\alpha = 10^{-3}$ for the proposed Radial IREG-PRM$^+$ in Algorithm 1. According to Fig. 1, we make the following observations. **(1)** Compared with ExRM$^+$, Smooth PRM$^+$, EG, and OG, all of which enjoy asymptotic or finite-time last-iterate convergence guarantees, Radial IREG-PRM$^+$ achieves competitive convergence performance on most game instances. A notable exception is Goofspiel, where EG converges substantially faster than the other methods. **(2)** Compared with IREG-PRM$^+$, which can be viewed as the $\alpha = 0$ counterpart of Radial IREG-PRM$^+$ and enjoys fast average-iterate convergence together with an $O(T^{-1/2})$ best-iterate rate, Radial IREG-PRM$^+$ exhibits comparable performance on the first three game instances, but converges more slowly on the random matrix game. **(3)** Among scale-invariant regret-matching algorithms, Radial IREG-PRM$^+$ outperforms PRM$^+$ on the $3 \times 3$ matrix game, Goofspiel, and the random matrix game, while achieving comparable performance on Kuhn Poker. Interestingly, alternating PRM$^+$, despite lacking both average-iterate and last-iterate convergence guarantees, converges substantially faster than Radial IREG-PRM$^+$ on the $3 \times 3$ matrix game and the random matrix game, while showing similar performance on Kuhn Poker and Goofspiel. Moreover, the sensitivity analysis of Radial IREG-PRM$^+$ w.r.t. different exponents $\alpha$ are deferred to Appendix G.

## 9 Conclusions

In this work, we introduced Radial IREG-PRM$^+$, a scale-invariant regret-matching-type algorithm with non-asymptotic last-iterate convergence guarantees in two-player zero-sum matrix games. The key idea is to apply a deterministic radial expansion to the pseudo-cumulative regret, thereby improving the stability of the learning dynamics without introducing an additional payoff-dependent learning rate or restarting mechanism. We show that this modification retains optimal adversarial regret, yields near-$O(1/T)$ average-iterate convergence in self-play, and leads to fast instance-dependent last-iterate convergence. We further characterize the uniform worst-case behavior through a matching lower-bound analysis on a degenerate branch, and extend the framework to the bandit setting with an in-expectation last-iterate guarantee. More broadly, our results suggest that controlling the *magnitude* of cumulative regret provides a useful mechanism for reconciling scale invariance with last-iterate stability. This perspective may offer a new route for designing online algorithms that preserve practical advantages of regret matching while enjoying last-iterate convergence guarantees.

### AI USE STATEMENT

In this work, we used generative AI tools for [formulate mathematical claims, assist in the writing of proofs, implement methods, assist with translation,interpret results]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks,provide critical ingredients for proving mathematical claims,propose or refine hypotheses], and [generate synthetic data sets,clean and reformat dataset, support qualitative and thematic data analysis] are not applicable to this work. Additionally, we used generative AI tools for [formulate questions for surveys or interviews,create or edit software code,summarize or analyse existing literature]. We have reviewed all AI-assisted work. [We checked LLM-generated research ideas for potential plagiarism through a manual literature survey, LLM-generated code was verified and tested for correctness by two authors]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This paper presents work whose goal is to advance the field of game theory. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

### REPRODUCIBILITY STATEMENT

We build on the code released by Cai et al. (2025a) in `https://proceedings.iclr.cc/paper_files/paper/2025/hash/7c20cba1f6aeeb219f9f9cbdde5c7382-Abstract-Conference.html`. All source code is provided in supplementary material.

## REFERENCES

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. Adaptively perturbed mirror descent for learning in games. In *International Conference on Machine Learning (ICML)*, pp. 31–80, 2024.

James P. Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *ACM Conference on Economics and Computation (EC)*, pp. 321–338, 2018.

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

Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, 365(6456): 885–890, 2019.

Yang Cai and Weiqiang Zheng. Doubly optimal no-regret learning in monotone games. In *International Conference on Machine Learning (ICML)*, pp. 3507–3524, 2023.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

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

Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng. Last-iterate convergence properties of regret-matching algorithms in games. In *International Conference on Learning Representations (ICLR)*, pp. 49634–49661, 2025a.

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

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Darshan Chakrabarti, Julien Grand-Clément, and Christian Kroer. Extensive-form game solving via Blackwell approachability on treeplexes. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. In *Conference on Learning Theory (COLT)*, pp. 807–834, 2019.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Annual Conference on Learning Theory (COLT)*, pp. 6.1–6.20, 2012.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, 2018.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *Innovations in Theoretical Computer Science Conference (ITCS)*, pp. 27:1–27:18, 2019.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations (ICLR)*, 2018.

Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, 2003.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, and Haipeng Luo. Regret matching+: (in)stability and fast convergence in games. In *annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 61546–61572, 2023.

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. In *International Conference on Machine Learning (ICML)*, pp. 17131–17152, 2025.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *Advances in Neural Information Processing Systems*, 2020a.

Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory (COLT)*, pp. 1758–1784, 2020b.

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

Alan J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263–265, 1952.

Harold W. Kuhn. A simplified two-person poker. In *Contributions to the Theory of Games*, volume 1, pp. 97–103. Princeton University Press, 1950.

Chung-Wei Lee, Christian Kroer, and Haipeng Luo. Last-iterate convergence in extensive-form games. In *Annual Conference on Neural Information Processing (NeurIPS)*, pp. 14293–14305, 2021.

Weiming Liu, Huacong Jiang, Bin Li, and Houqiang Li. Equivalence analysis between counterfactual regret minimization and online mirror descent. In *International Conference on Machine Learning (ICML)*, pp. 13717–13745, 2022.

Linjian Meng, Youzhi Zhang, Zhenxing Ge, Tianyu Ding, Shangdong Yang, Zheng Xu, Wenbin Li, and Yang Gao. Last-iterate convergence of smooth regret matching$ˆ+$ variants in learning nash equilibria. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Linjian Meng, Youzhi Zhang, Shangdong Yang, Wenbin Li, Tianyu Ding, and Yang Gao. A faster parameter-free regret matching algorithm. In *International Conference on Learning Representations (ICLR)*, 2026.

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

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 3168–3176, 2015.

John Von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

Francesco Orabona and Dávid Pál. Scale-free online learning. *Theor. Comput. Sci.*, 716:50–69, 2018.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1–2):1–35, 2021.

Julien Pérolat, Rémi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro A. Ortega, Neil Burch, Thomas W. Anthony, David Balduzzi, Bart De Vylder, Georgios Piliouras, Marc Lanctot, and Karl Tuyls. From Poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *International Conference on Machine Learning (ICML)*, pp. 8525–8535, 2021.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory (COLT)*, pp. 993–1019, 2013a.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 3066–3074, 2013b.

Sheldon M. Ross. Goofspiel: The game of pure strategy. *Journal of Applied Probability*, 8(3): 621–625, 1971.

Oskari Tammelin, Neil Burch, Michael Johanson, and Michael Bowling. Solving heads-up limit texas hold'em. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 645–652, 2015.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Brian Hu Zhang, Ioannis Anagnostides, and Tuomas Sandholm. Scale-invariant regret matching and online learning with optimal convergence: Bridging theory and practice in zero-sum games. In *ACM Conference on Economics and Computation (EC)*, 2026.

Canzhe Zhao, Yutian Cheng, Jing Dong, Baoxiang Wang, and Shuai Li. Learning imperfect information extensive-form games with last-iterate convergence under bandit feedback. In *International Conference on Machine Learning (ICML)*, 2025.

# A BASIC LEMMAS FOR RADIAL IREG ALGORITHM

**Lemma A.1** ($\ell_1$-norm of predicted regret vector and prediction movement)**.** *Let $R \in \mathbb{R}^d_+$, $S = \|R\|_2 > 0$, and $m \in \mathbb{R}^d$. Then there exists a unique $\gamma \in \mathbb{R}$ such that $Q = [R + m - \gamma\mathbf{1}]_+$ satisfies $\|Q\|_2 = S$. Let $z = \mathcal{N}(Q)$ and $\delta = \langle m, z\rangle - \gamma$. Then*

$$\boxed{\delta\|Q\|_1 = \frac{1}{2}\|Q - R\|_2^2, \qquad \delta \geq 0.} \tag{18}$$

*Moreover, if $\widetilde{z} = \mathcal{N}(R)$, then for any $p \in \Delta_d$,*

$$\langle m, p - \widetilde{z}\rangle \leq \big(2\sqrt{d}\, S + \|m\|_2\big)\|z - \widetilde{z}\|_2. \tag{19}$$

*Proof.* The function $f(\gamma) = \|[R + m - \gamma\mathbf{1}]_+\|_2$ is continuous and strictly decreasing on the region where it is positive, with $f(\gamma) \to \infty$ as $\gamma \to -\infty$, and $f(\gamma) = 0$ whenever $\gamma \geq \max_j(R_j + m_j)$. Since $S > 0$, there exists a unique root satisfying $f(\gamma) = S$.

On coordinates where $Q_j > 0$, $m_j - \gamma = Q_j - R_j$; on coordinates where $Q_j = 0$, the product of $m_j - \gamma$ and $Q_j$ equals 0. Therefore

$$\begin{aligned}\delta\|Q\|_1 &= \langle m - \gamma\mathbf{1}, Q\rangle = \langle Q - R, Q\rangle \\ &= \frac{1}{2}\big(\|Q - R\|_2^2 + \|Q\|_2^2 - \|R\|_2^2\big) = \frac{1}{2}\|Q - R\|_2^2,\end{aligned}$$

where the last equality holds due to the fact that $\|Q\|_2^2 = \|R\|_2^2 = S$. Moreover, $m - \gamma\mathbf{1} \leq Q - R$ holds coordinatewise, and $\gamma = \gamma\,\langle\mathbf{1}, p\rangle$, so

$$\langle m, p - z\rangle = \langle m, p\rangle - \gamma - \delta \leq \langle Q - R, p\rangle - \delta \leq \|Q - R\|_2, \tag{20}$$

where the last inequality uses $\delta \geq 0$ in Eq. 28 and the Cauchy-Schwartz inequality $\langle Q - R, p\rangle \leq \|Q - R\|_2\|p\|_2$. Let $\phi(q) = q/\|q\|_2$. From $R = S\phi(\widetilde{z})$, $Q = S\phi(z)$, and

$$\|\phi(p) - \phi(q)\|_2 \leq 2\sqrt{d}\,\|p - q\|_2, \qquad p, q \in \Delta_d, \tag{21}$$

we obtain

$$\|Q - R\|_2 \leq 2\sqrt{d}S\|z - \widetilde{z}\|_2 \tag{22}$$

Finally, for the decomposition $\langle m, p - \widetilde{z}\rangle = \langle m, p - z\rangle + \langle m, z - \widetilde{z}\rangle$, applying Eq. 20 and Eq. 22 to bound the first term, and applying Cauchy–Schwartz inequality to bound the second term, derives Eq. 19. □

**Lemma A.2** (Positive correction and normalization bound)**.** *Let $Q \in \mathbb{R}^d_+ \setminus \{0\}$, $S = \|Q\|_2$, and $z = \mathcal{N}(Q)$, and define*

$$E = e - \langle e, z\rangle\mathbf{1}, \qquad H = [Q + E]_+.$$

*Then*

$$\langle Q, E\rangle = 0, \tag{23}$$

$$H - Q \geq E \quad \textit{coordinatewise}, \tag{24}$$

$$H \neq 0, \qquad \|H\|_2 \geq S, \tag{25}$$

$$\|H\|_2^2 \leq S^2 + \|E\|_2^2, \tag{26}$$

$$\|\mathcal{N}(H) - \mathcal{N}(Q)\|_2 \leq (1 + \sqrt{d})\frac{\|E\|_2}{S}. \tag{27}$$

*Proof.* Use the definition $E = e - \langle e, z\rangle\mathbf{1}$ and $Q = \|Q\|_1 z$,we have $\langle Q, E\rangle = \|Q\|_1(\langle e, z\rangle - \langle e, z\rangle) = 0$. For each coordinate, $[Q_j + E_j]_+ - Q_j \geq E_j$, hence $\langle Q, H - Q\rangle \geq \langle Q, E\rangle = 0$. Therefore

$$\|H\|_2^2 = S^2 + 2\langle Q, H - Q\rangle + \|H - Q\|_2^2 \geq S^2 + \|H - Q\|_2^2 > 0, \tag{28}$$

and thus $H \neq 0$. On the other hand, the positive-part projection is Euclidean nonexpansive, so

$$\|H\|_2^2 \leq \|Q + E\|_2^2 = S^2 + \|E\|_2^2, \qquad \|H - Q\|_2 \leq \|E\|_2.$$

Finally, for any nonzero $h, q \in \mathbb{R}^d_+$,

$$\left\| \frac{h}{\|h\|_1} - \frac{q}{\|q\|_1} \right\|_2 \le \frac{\|h-q\|_2}{\|q\|_1} + \frac{\|h\|_2 \, |\|q\|_1 - \|h\|_1|}{\|h\|_1 \|q\|_1} \le \frac{(1+\sqrt{d})\|h-q\|_2}{\|q\|_1},$$

where the last inequality uses the fact that $\|h\|_2 \le \|h\|_1 \le \sqrt{d}\|h\|_2$. Taking $h = H, q = Q$ and then using $\|Q\|_1 \ge \|Q\|_2 = S$ yields Eq. 27. □

For any $w = (w_x, w_y) \in Z$, recall the perspective function

$$\boxed{\Psi_t(w) = \frac{1}{g_t} \sum_{i \in \{x,y\}} P(R_i^t, w_i), \qquad P(r,p) = \|r\|_2 \|p\|_2 - \langle r, p \rangle, \quad r \in \mathbb{R}^d_+, \; p \in \Delta_d.} \tag{29}$$

**Lemma A.3** (potential function descent inequality)**.** *In the positive-branch, for any $p \in \Delta_d$,*

$$P(H,p) \le P(R,p) + \langle u, z - p \rangle - \delta + \frac{\|E\|_2^2}{2S}. \tag{30}$$

*Proof.* Since $\|Q\|_2 = \|R\|_2 = S$ and $Q - R \ge m - \gamma \mathbf{1}$ coordinatewise,

$$P(Q,p) - P(R,p) = -\langle Q - R, p \rangle \le \gamma - \langle m, p \rangle = \langle m, z - p \rangle - \delta.$$

By Lemma A.2, $\|H\|_2 - S \le \|E\|_2^2/(2S)$ and $H - Q \ge E$, so

$$P(H,p) - P(Q,p) \le -\langle E, p \rangle + \frac{\|E\|_2^2}{2S}.$$

Also, $E = e - \langle e, z \rangle \mathbf{1}$ and $e = u - m$, hence $-\langle E, p \rangle = \langle e, z - p \rangle$. Adding the two displays gives Eq. 30. □

**Lemma A.4** (perspective function descent inequality)**.** *For any comparator $w \in Z$,*

$$\Psi_{t+1}(w) \le \Psi_t(w) + \frac{1}{g_t} \sum_i \langle u_i^t, z_i^t - w_i \rangle - \frac{\delta_x^t + \delta_y^t}{g_t} + \frac{1}{2} \mathcal{V}_t. \tag{31}$$

*Proof.* For any fixed $i, t$, applying Eq. 30 in Lemma A.3 yields

$$P(H_i^t, w_i) \le P(R_i^t, w_i) + \langle u_i^t, z_i^t - w_i \rangle - \delta_i^t + \frac{\|E_i^t\|_2^2}{2S_i^t}. \tag{32}$$

Using $R_i^{t+1} = \lambda_t H_i^t$, $\lambda_t = g_{t+1}/g_t$, and $P(cr, w) = cP(r, w)(c \ge 0)$,

$$\frac{P(R_i^{t+1}, w_i)}{g_{t+1}} = \frac{P(H_i^t, w_i)}{g_t}.$$

Then summing Eq. 32 over indices $i \in \{x, y\}$ and dividing both sides with $g_{t+1}$ obtains Eq. 31. □

**Lemma A.5** (Equivalence between perspective and distance to an arbitrary comparator)**.** *Assume that $\underline{\sigma} \le \sigma_i^t = \frac{\|R_i^t\|_2}{g_t} \le \overline{\sigma} < \infty, \quad \forall i \in \{x, y\}, t \ge 1$. Then for any $t \ge 1$ and any $w \in Z$,*

$$\boxed{c_D \underline{\sigma} \|\widetilde{z}^t - w\|_2^2 \le \Psi_t(w) \le 2D\overline{\sigma} \|\widetilde{z}^t - w\|_2^2.} \tag{33}$$

*Moreover, for any $w, w' \in Z$,*

$$|\Psi_t(w) - \Psi_t(w')| \le 2\sqrt{2}\,\overline{\sigma} \|w - w'\|_2. \tag{34}$$

*Proof.* Fix one simplex, denote its dimension by $d \le D$, and write $p = \widetilde{z}_i^t$ and $q = w_i$. Since $p, q \in \Delta_d$, $\phi(p) = \frac{p}{\|p\|_2}$, $\phi(q) = \frac{q}{\|q\|_2}$, and both are Euclidean unit vectors. Hence the exact angular identity holds:

$$\begin{aligned}\|q\|_2 - \langle \phi(p), q\rangle &= \|q\|_2\big(1 - \langle \phi(p), \phi(q)\rangle\big)\\ &= \frac{\|q\|_2}{2}\|\phi(p) - \phi(q)\|_2^2. \end{aligned} \tag{35}$$

We now explicitly relate the unit-direction distance to the Euclidean distance in the simplex. First, $\|p\|_2, \|q\|_2 \in [1/\sqrt{d}, 1]$, and the normalization map satisfies

$$\|\phi(p) - \phi(q)\|_2 \le 2\sqrt{d}\,\|p - q\|_2. \tag{36}$$

Conversely, because $p = \mathcal{N}(\phi(p))$, $q = \mathcal{N}(\phi(q))$, and for nonzero $u, v \in \mathbb{R}_+^d$ we have

$$\|\mathcal{N}(u) - \mathcal{N}(v)\|_2 \le (1 + \sqrt{d})\frac{\|u - v\|_2}{\|v\|_1},$$

Taking $u = \phi(p), v = \phi(q)$ and using $\|\phi(q)\|_1 \ge 1$ gives

$$\|\phi(p) - \phi(q)\|_2 \ge \frac{\|p - q\|_2}{1 + \sqrt{d}}. \tag{37}$$

Substituting Eq. 36–Eq. 37 into Eq. 35 and using $\|q\|_2 \in [1/\sqrt{d}, 1]$ gives

$$\frac{1}{2\sqrt{d}(1 + \sqrt{d})^2}\|p - q\|_2^2 \le \|q\|_2 - \langle \phi(p), q\rangle \le 2d\|p - q\|_2^2. \tag{38}$$

Since $d \le D$, the constant on the left is at least $c_D = [2\sqrt{D}(1 + \sqrt{D})^2]^{-1}$, while the right-hand side is at most $2D$. On the other hand,

$$\frac{R_i^t}{g_t} = \frac{S_i^t}{g_t}\frac{R_i^t}{S_i^t} = \sigma_i^t \phi(\widetilde{z}_i^t),$$

Therefore

$$\frac{1}{g_t}P(R_i^t, w_i) = \sigma_i^t\big(\|w_i\|_2 - \langle \phi(\widetilde{z}_i^t), w_i\rangle\big).$$

Using $\underline{\sigma} \le \sigma_i^t \le \overline{\sigma}$ and Eq. 38, blockwise we obtain

$$c_D\underline{\sigma}\|\widetilde{z}_i^t - w_i\|_2^2 \le \frac{1}{g_t}P(R_i^t, w_i) \le 2D\overline{\sigma}\|\widetilde{z}_i^t - w_i\|_2^2.$$

Summing over $i = x, y$ yields Eq. 33. Finally, for any $w_i, w_i' \in \Delta_{d_i}$, the triangle inequality gives

$$|P(R_i^t, w_i) - P(R_i^t, w_i')| \le S_i^t\big|\|w_i\|_2 - \|w_i'\|_2\big| + S_i^t\|w_i - w_i'\|_2 \le 2S_i^t\|w_i - w_i'\|_2.$$

Dividing by $g_t$, using $\sigma_i^t \le \overline{\sigma}$, and then applying $u + v \le \sqrt{2}(u^2 + v^2)^{1/2}$ to the two blocks yields Eq. 34. □

# B Sublinear Adversarial Regret Bound and Weighted RVU Property

## B.1 Sublinear Adversarial Per-Player Regret Bound

**Theorem B.1** (Coordinatewise cumulative-regret domination)**.** *Algorithm 1 satisfies, for all $T \ge 1$,*

$$R_i^{T+1} \ge \sum_{t=1}^{T} r_i^t \quad \textit{coordinatewise}. \tag{39}$$

*Proof of Theorem B.1.* If $S_i^t > 0$, from $Q_i^t \ge R_i^t + m_i^t - \gamma_i^t\mathbf{1}$ and $E_i^t = u_i^t - m_i^t - \langle u_i^t - m_i^t, z_i^t\rangle\mathbf{1}$,

$$Q_i^t + E_i^t \ge R_i^t + r_i^t + \delta_i^t\mathbf{1} \ge R_i^t + r_i^t.$$

we obtain $H_i^t = [Q_i^t + E_i^t]_+ \ge R_i^t + r_i^t$. If $S_i^t = 0$, then $Q_i^t = 0, m_i^t = 0, E_i^t = r_i^t$, so $H_i^t = [r_i^t]_+ \ge r_i^t = R_i^t + r_i^t$. Finally, $\lambda_t \ge 1$ and $H_i^t \ge 0$, hence $R_i^{t+1} = \lambda_t H_i^t \ge R_i^t + r_i^t$. Induction from $R_i^1 = 0$ proves the claim. □

*Proof for Theorem 4.1.* Let $\sigma_i^t = S_i^t/g_t$. In the positive branch, by Eq. 26; in the zero branch, since $Q = 0$, in both cases we have

$$(\sigma_i^{t+1})^2 = \frac{\|H_i^t\|_2^2}{g_t^2} \le (\sigma_i^t)^2 + \frac{\|E_i^t\|_2^2}{g_t^2}, \tag{40}$$

where the inequality uses the fact that $\|H_i^t\|_2^2 \le \|Q_i^t\|_2^2 + \|E_i^t\|_2^2$ and $\|Q_i^t\|_2^2/g_t^2 = (\sigma_i^t)^2$.

It is not difficult to see that the linear centering map $\mathcal{C}_z(v) = v - \langle v, z\rangle \mathbf{1}$ satisfies $\|\mathcal{C}_z(v)\|_2 \le C_d\|v\|_2$. Taking $u_i^0 = 0$, we have $\|u_i^t - u_i^{t-1}\|_2 \le 2G_i$, and therefore uniformly $\|E_i^t\|_2 = \|\mathcal{C}_{z_i^t}(e_i^t)\|_2 \le 2C_{d_i}G_i$. Because $g_t^2 = t^{2\alpha}$, applying the recursion in Eq. 40 gives

$$(\sigma_i^{T+1})^2 \le 4C_{d_i}^2 G_i^2 \sum_{t=1}^{T} t^{-2\alpha}.$$

If $0 < \alpha < 1/2$, then $\sum_{t=1}^{T} t^{-2\alpha} \le T^{1-2\alpha}/(1-2\alpha)$, which yields Eq. 5 after multiplying by $g_{T+1} = (T+1)^\alpha$. If $\alpha = 1/2$, the harmonic bound $\sum_{t=1}^{T} t^{-1} \le 1 + \log T$ gives Eq. **??**. The first inequality follows from Theorem B.1 and $\max_j R_{i,j}^{T+1} \le \|R_i^{T+1}\|_2$.

Scale invariance follows by induction: if all utilities are multiplied by $c > 0$, then $R, Q, E, H, \gamma$ are all multiplied by $c$, and $S$ is multiplied by $c$, while all normalized strategies remain unchanged; the zero-branch branch is unchanged as well. □

*Proof of Lemma 3.2.* In the zero branch, $H = [u - \langle u, a\rangle \mathbf{1}]_+$. Since $a$ has full support, if $u$ is nonconstant, $\langle u, a\rangle$ is strictly smaller than $\max_j u_j$, so there is at least one positive coordinate; if $u$ is constant, then $H = 0$. The positive-state conclusion follows from $\|H\|_2 \ge S > 0$ in Lemma A.2, and from $S_i^{t+1} = \|R_i^{t+1}\|_2 = \lambda_t \|H_i^t\|_2$ with $\lambda_t = g_{t+1}/g_t$. □

### B.2 Weighted RVU bound on the positive-state prefix

Analogous to previous works (Rakhlin & Sridharan, 2013b;a; Zhang et al., 2026), we also yield a radially weighted predictive RVU bound as follow. For conciseness, we fix one player $i$, suppress the player index when no confusion can arise, and write $d = d_i$.

**Theorem B.2** (Radially weighted predictive RVU bound for Radial IREG$^+$)**.** *Fix integers $T \ge t_0 \ge 1$ such that the player is in the positive-state branch throughout $t = t_0, \dots, T$, i.e., $S^t = \|R^t\|_2 > 0$. Let $\underline{\sigma} := \inf_{t_0 \le t \le T} \frac{S^t}{g_t} > 0$. Then for any comparator $p \in \Delta_d$,*

$$\boxed{\sum_{t=t_0}^{T} \frac{1}{g_t} \left\langle u^t, p - z^t \right\rangle \le \frac{P(R^{t_0}, p)}{g_{t_0}} + \frac{C_d^2}{2\underline{\sigma}} \sum_{t=t_0}^{T} \frac{\|u^t - m^t\|_2^2}{g_t^2} - c_d \underline{\sigma} \sum_{t=t_0}^{T} \|z^t - \widetilde{z}^t\|_2^2.} \tag{41}$$

*For Algorithm 1, once the state has activated, Lemma 3.2 implies that $S^t/g_t$ is nondecreasing. Hence one may take $\underline{\sigma} = S^{t_0}/g_{t_0}$ whenever $t_0$ is at or after activation.*

*Proof.* Recall that $e^t = u^t - m^t$ and $E^t = e^t - \langle e^t, z^t\rangle \mathbf{1}$. By Lemma A.3, for every $t \in \{t_0, \dots, T\}$,

$$\left\langle u^t, p - z^t \right\rangle \le P(R^t, p) - P(H^t, p) - \delta^t + \frac{\|E^t\|_2^2}{2S^t}. \tag{42}$$

Divide by $g_t$ and use $R^{t+1} = \lambda_t H^t$, $\lambda_t = g_{t+1}/g_t$, we have

$$\frac{1}{g_t} \left\langle u^t, p - z^t \right\rangle \le \frac{P(R^t, p)}{g_t} - \frac{P(R^{t+1}, p)}{g_{t+1}} - \frac{\delta^t}{g_t} + \frac{\|E^t\|_2^2}{2g_t S^t}. \tag{43}$$

Summing Eq. 43 from $t = t_0$ to $T$ gives the exact telescoping relation

$$\sum_{t=t_0}^{T} \frac{1}{g_t} \left\langle u^t, p - z^t \right\rangle \le \frac{P(R^{t_0}, p)}{g_{t_0}} - \frac{P(R^{T+1}, p)}{g_{T+1}} + \sum_{t=t_0}^{T} \frac{\|E^t\|_2^2}{2g_t S^t} - \sum_{t=t_0}^{T} \frac{\delta^t}{g_t}. \tag{44}$$

Since $P(r,p) = \|r\|_2\|p\|_2 - \langle r,p\rangle \geq 0$ by Cauchy–Schwarz, the terminal term in Eq. 44 is nonpositive and may be discarded.

We first bound the prediction-error term. The centering map $\mathcal{C}_z(v) = v - \langle v, z\rangle \mathbf{1}$ satisfies $\|\mathcal{C}_z(v)\|_2 \leq (1+\sqrt{d})\|v\|_2 = C_d\|v\|_2, \quad z \in \Delta_d$. Hence $\|E^t\|_2 \leq C_d\|u^t - m^t\|_2$. Since $S^t \geq \underline{\sigma} g_t$,

$$\frac{\|E^t\|_2^2}{2g_t S^t} \leq \frac{C_d^2}{2\underline{\sigma}} \frac{\|u^t - m^t\|_2^2}{g_t^2}. \tag{45}$$

We next lower-bound the stability gain $\delta^t$. Use $Q^t = S^t\phi(z^t), \quad R^t = S^t\phi(\widetilde{z}^t)$ and $\|\phi(p) - \phi(q)\|_2 \geq \frac{1}{1+\sqrt{d}}\|p-q\|_2$ for any $p, q \in \Delta_d$ in Eq. 37. Also, $\|Q^t\|_1 \leq \sqrt{d}\,\|Q^t\|_2 = \sqrt{d}\,S^t$. Substituting these facts into Lemma A.1 yields

$$\delta^t = \frac{\|Q^t - R^t\|_2^2}{2\|Q^t\|_1} = \frac{(S^t)^2}{2\|Q^t\|_1}\|\phi(z^t) - \phi(\widetilde{z}^t)\|_2^2 \geq \frac{S^t}{2\sqrt{d}(1+\sqrt{d})^2}\|z^t - \widetilde{z}^t\|_2^2 = c_d S^t \|z^t - \widetilde{z}^t\|_2^2.$$

Consequently, we have

$$\frac{\delta^t}{g_t} \geq c_d\underline{\sigma}\|z^t - \widetilde{z}^t\|_2^2. \tag{46}$$

Combining Eq. 44, Eq. 45, and Eq. 46 proves Eq. 41. □

**Corollary B.3** (Path-length form)**.** *Under the assumptions of Theorem B.2,*

$$\sum_{t=t_0}^{T} \frac{1}{g_t}\left\langle u^t, p - z^t\right\rangle \leq \frac{P(R^{t_0}, p)}{g_{t_0}} + \frac{1}{\underline{\sigma}}\left(\frac{C_d^2}{2} + c_d C_d^4\right)\sum_{t=t_0}^{T}\frac{\|u^t - m^t\|_2^2}{g_t^2} - \frac{c_d\underline{\sigma}}{2}\sum_{t=t_0+1}^{T}\|z^t - z^{t-1}\|_2^2. \tag{47}$$

*Proof.* For $t \geq t_0 + 1$, decompose $z^t - z^{t-1} = (z^t - \widetilde{z}^t) + (\widetilde{z}^t - z^{t-1})$. Hence

$$\|z^t - \widetilde{z}^t\|_2^2 \geq \frac{1}{2}\|z^t - z^{t-1}\|_2^2 - \|\widetilde{z}^t - z^{t-1}\|_2^2. \tag{48}$$

Because $R^t = \lambda_{t-1}H^{t-1}$ and positive scaling does not change $\ell_1$ normalization, $\widetilde{z}^t = \mathcal{N}(H^{t-1})$, while $z^{t-1} = \mathcal{N}(Q^{t-1})$. Lemma A.2 therefore gives

$$\|\widetilde{z}^t - z^{t-1}\|_2 \leq C_d\frac{\|E^{t-1}\|_2}{S^{t-1}} \leq \frac{C_d^2}{\underline{\sigma}}\frac{\|u^{t-1} - m^{t-1}\|_2}{g_{t-1}}, \tag{49}$$

where the second inequality uses Eq. 45 and $S^{t-1} \geq \underline{\sigma} g_{t-1}$. Consequently,

$$\sum_{t=t_0+1}^{T}\|\widetilde{z}^t - z^{t-1}\|_2^2 \leq \frac{C_d^4}{\underline{\sigma}^2}\sum_{t=t_0}^{T-1}\frac{\|u^t - m^t\|_2^2}{g_t^2}. \tag{50}$$

Starting from Theorem B.2, discard the nonpositive term corresponding to $t = t_0$ in the stability sum and apply Eq. 48–Eq. 50, we obtain Eq. 47. □

### B.3 Weighted RVU bound on the zero-branch

Theorem B.2 and Corollary B.3 are stated on a positive-state interval because their proofs divide by the effective radius $S^t$. We next show that the zero-branch branch does not destroy the RVU structure. Persistent zero-branch rounds incur exactly zero regret; if activation occurs, there is only one exceptional zero-branch round, which contributes a finite boundary cost.

**Theorem B.4** (Global radially weighted RVU bound including the zero-branch)**.** *Fix one player, suppress the player index, and let* $\tau := \inf\{t \geq 1 : S^t > 0\} \in \{2, 3, \ldots\} \cup \{\infty\}$. *Define the activation cost*

$$A_\tau := \frac{1}{g_{\tau-1}}\max_{q\in\Delta_d}\left\langle u^{\tau-1}, q - a\right\rangle. \tag{51}$$

*Then, for every* $T \geq \tau$ *and every comparator* $p \in \Delta_d$,

$$\boxed{\sum_{t=1}^{T}\frac{1}{g_t}\left\langle u^t, p - z^t\right\rangle \leq A_\tau + \frac{P(R^\tau, p)}{g_\tau} + \frac{C_d^2}{2\underline{\sigma}}\sum_{t=\tau}^{T}\frac{\|u^t - m^t\|_2^2}{g_t^2} - c_d\underline{\sigma}\sum_{t=\tau}^{T}\|z^t - \widetilde{z}^t\|_2^2.} \tag{52}$$

*If $\|u^t\|_2 \le G$ for all $t$, then the boundary term is uniformly bounded by*

$$A_\tau + \frac{P(R^\tau, p)}{g_\tau} \le \frac{(\sqrt{2} + C_d)G}{g_{\tau-1}}, \tag{53}$$

*which is independent of the horizon $T$. If $\tau = \infty$, then for every $T \ge 1$ and every $p \in \Delta_d$,*

$$\boxed{\sum_{t=1}^{T} \frac{1}{g_t} \left\langle u^t, p - z^t \right\rangle = 0.} \tag{54}$$

*Proof.* Assume first that $\tau < \infty$. By definition, $S^t = 0$ for $t \le \tau - 1$ and $S^\tau > 0$, that is, for every $t \le \tau - 2$, we have $S^{t+1} = 0$. Since $R^{t+1} = \lambda_t H^t$ with $\lambda_t > 0$, this implies $H^t = 0$. Lemma 3.2 then implies that $u^t$ is a constant vector. Since the zero-branch branch plays $z^t = a \in \Delta_d$, then for every comparator $p \in \Delta_d$,

$$\left\langle u^t, p - z^t \right\rangle = 0, \qquad t = 1, \dots, \tau - 2. \tag{55}$$

The only zero-branch round that may contribute nonzero regret is therefore the activation round $t = \tau - 1$, and by definition of $A_\tau$,

$$\frac{1}{g_{\tau-1}} \left\langle u^{\tau-1}, p - a \right\rangle \le A_\tau. \tag{56}$$

From round $\tau$ onward the state is positive forever by Lemma 3.2. Moreover, $\sigma^t = S^t/g_t$ is nondecreasing, so $S^t \ge \underline{\sigma} g_t$ for all $t \ge \tau$. Applying Theorem B.2 on the interval $[\tau, T]$ with this lower bound and then combining with Eq. 55–Eq. 56 proves Eq. 52.

For the explicit boundary estimate, the Euclidean diameter of the simplex is $\sqrt{2}$, so

$$A_\tau \le \frac{\sqrt{2}\, G}{g_{\tau-1}}. \tag{57}$$

At round $\tau - 1$ the state is zero, hence $Q^{\tau-1} = m^{\tau-1} = 0$ and $H^{\tau-1} = \left[u^{\tau-1} - \left\langle u^{\tau-1}, a \right\rangle \mathbf{1}\right]_+$. Since $R^\tau = \lambda_{\tau-1} H^{\tau-1}$, $g_\tau = \lambda_{\tau-1} g_{\tau-1}$, and $P$ is positively homogeneous in its first argument,

$$\frac{P(R^\tau, p)}{g_\tau} = \frac{P(H^{\tau-1}, p)}{g_{\tau-1}}. \tag{58}$$

Also, $P(h, p) \le \|h\|_2$ for $h \in \mathbb{R}^d_+$ and $p \in \Delta_d$, while the centering bound gives $\|H^{\tau-1}\|_2 \le C_d \|u^{\tau-1}\|_2 \le C_d G$. Combining this with Eq. 57 gives Eq. 53.

Finally, if $\tau = \infty$, then $S^t = S^{t+1} = 0$ for every $t$. The same argument as above gives $H^t = 0$ and hence, by Lemma 3.2, $u^t$ is a constant vector in every round. Since the algorithm always plays $z^t = a$, each instantaneous regret against every comparator is zero, which proves Eq. 54. □

**Lemma B.5** (Weighted RVU regret bound)**.** *Define* $\Pi = \frac{1}{\underline{\sigma}} \left( \frac{C_d^2}{2} + c_d C_d^4 \right)$, $\Gamma = \frac{c_d \underline{\sigma}}{2}$, $\beta_\tau = \frac{(\sqrt{2}+C_d)L}{g_{\tau-1}} + 2\Gamma$, *when $\tau < \infty$. Then for each player $i \in \{x, y\}$, any $T \ge 1$ and $p \in \Delta_d$,*

$$\sum_{t=1}^{T} \frac{1}{g_t} \left\langle u_i^t, p - z_i^t \right\rangle \le \left[ \beta_\tau + \Pi \sum_{t=1}^{T} \frac{\|u_i^t - m_i^t\|_2^2}{g_t^2} - \Gamma \sum_{t=2}^{T} \|z_i^t - z_i^{t-1}\|_2^2 \right] \mathbf{1}(\tau < \infty). \tag{59}$$

*Proof of Lemma B.5.* Apply Corollary B.3 on the positive-state interval $[\tau, T]$, using the lower bound $\sigma^t \ge \underline{\sigma}$, and combine it with the zero-prefix decomposition from the proof of Theorem B.4. Before activation, $z^1 = \cdots = z^{\tau-1} = a$, so all path-length terms with $2 \le t \le \tau - 1$ vanish. The only missing movement when extending the negative sum from $t = \tau + 1$ down to $t = 2$ is $\|z^\tau - z^{\tau-1}\|_2^2 = \|z^\tau - a\|_2^2 \le 2$. Paying $2\Gamma$ in the boundary constant therefore permits the full negative path-length sum. Finally, extending the nonnegative variation sum from $[\tau, T]$ to $[1, T]$ preserves the upper bound. This proves Eq. 59; the stated explicit boundary constant follows from Eq. 53. □

### B.4 SELF-PLAY CONSEQUENCES OF THE GLOBAL PATH-LENGTH RVU BOUND

We now specialize to the fixed two-player zero-sum matrix game, with $u_x^t = Ay^t$, $u_y^t = -A^\top x^t$, $L = \|A\|_{\mathrm{op}}$, and $g_t = t^\alpha$ with $0 < \alpha \le 1/2$. The next result is the RVU cancellation step that is absent from the adversarial proof of Theorem 4.1.

**Proposition B.6** (Bounded second-order path length in the two-active branch)**.** *Suppose $\tau_x, \tau_y < \infty$, and let $\tau = \max\{\tau_x, \tau_y\}$. For $i \in \{x, y\}$, set*

$$\underline{\sigma}_i := \frac{S_i^{\tau_i}}{g_{\tau_i}} > 0, \qquad \Pi_i := \frac{1}{\underline{\sigma}_i}\left(\frac{C_{d_i}^2}{2} + c_{d_i} C_{d_i}^4\right), \qquad \Gamma_i := \frac{c_{d_i}\underline{\sigma}_i}{2}. \tag{60}$$

*Choose an integer $t_\star = \max\left\{\tau, \left\lceil \left(\frac{2\Pi_x L^2}{\Gamma_y}\right)^{1/(2\alpha)} \right\rceil, \left\lceil \left(\frac{2\Pi_y L^2}{\Gamma_x}\right)^{1/(2\alpha)} \right\rceil\right\}$, such that*

$$\frac{\Pi_x L^2}{g_t^2} \le \frac{\Gamma_y}{2}, \qquad \frac{\Pi_y L^2}{g_t^2} \le \frac{\Gamma_x}{2}, \qquad t \ge t_\star + 1. \tag{61}$$

*For each player, write its activation cost as $A_{i,\tau_i} := \frac{1}{g_{\tau_i - 1}} \max_{q_i \in \Delta_{d_i}} \left\langle u_i^{\tau_i - 1}, q_i - a_i \right\rangle$. Define the comparator-uniform boundary constants $\alpha_i := A_{i,\tau_i} + \max_{p_i \in \Delta_{d_i}} \frac{P(R_i^{\tau_i}, p_i)}{g_{\tau_i}} + 2\Gamma_i$, and the finite prefix variations*

$$V_i^\star := \sum_{t=1}^{t_\star} \frac{\|u_i^t - m_i^t\|_2^2}{g_t^2}, \qquad K_\star := \alpha_x + \alpha_y + \Pi_x V_x^\star + \Pi_y V_y^\star. \tag{62}$$

*Then, for every $T \ge t_\star + 1$,*

$$\boxed{\frac{\Gamma_x}{2} \sum_{t=t_\star+1}^{T} \|x^t - x^{t-1}\|_2^2 + \frac{\Gamma_y}{2} \sum_{t=t_\star+1}^{T} \|y^t - y^{t-1}\|_2^2 \le K_\star.} \tag{63}$$

*Proof.* Let $w_t = 1/g_t$. Define the weighted external regrets

$$\mathrm{Reg}_{x,w}(T) := \max_{p \in \Delta_m} \sum_{t=1}^{T} w_t \left\langle Ay^t, p - x^t \right\rangle, \tag{64}$$

$$\mathrm{Reg}_{y,w}(T) := \max_{q \in \Delta_n} \sum_{t=1}^{T} w_t \left\langle -A^\top x^t, q - y^t \right\rangle. \tag{65}$$

Writing $W_T = \sum_{t=1}^T w_t$ and

$$\bar{x}_w^T := \frac{1}{W_T} \sum_{t=1}^{T} w_t x^t, \qquad \bar{y}_w^T := \frac{1}{W_T} \sum_{t=1}^{T} w_t y^t,$$

a direct expansion gives the exact identity

$$\mathrm{Reg}_{x,w}(T) + \mathrm{Reg}_{y,w}(T) = W_T \,\mathrm{Gap}(\bar{x}_w^T, \bar{y}_w^T) \ge 0. \tag{66}$$

Apply Corollary B.5 to each player and then maximize over the corresponding comparator. For $t \ge t_\star + 1 \ge \tau$, both players are in the positive-state branch, so $m_x^t = u_x^{t-1}$ and $m_y^t = u_y^{t-1}$. Hence

$$\|u_x^t - m_x^t\|_2^2 = \|A(y^t - y^{t-1})\|_2^2 \le L^2 \|y^t - y^{t-1}\|_2^2, \tag{67}$$

$$\|u_y^t - m_y^t\|_2^2 = \|A^\top (x^t - x^{t-1})\|_2^2 \le L^2 \|x^t - x^{t-1}\|_2^2. \tag{68}$$

Splitting each positive variation sum at $t_\star$ and using Eq. 67–Eq. 68, we obtain

$$\begin{aligned} \mathrm{Reg}_{x,w}(T) + \mathrm{Reg}_{y,w}(T) \le K_\star &+ \sum_{t=t_\star+1}^{T} \left(\frac{\Pi_y L^2}{g_t^2} - \Gamma_x\right) \|x^t - x^{t-1}\|_2^2 \\ &+ \sum_{t=t_\star+1}^{T} \left(\frac{\Pi_x L^2}{g_t^2} - \Gamma_y\right) \|y^t - y^{t-1}\|_2^2, \end{aligned} \tag{69}$$

where we discarded the nonpositive path-length terms from the finite prefix. By Eq. 61, the right-hand side of Eq. 69 is at most

$$K_\star - \frac{\Gamma_x}{2} \sum_{t=t_\star+1}^{T} \|x^t - x^{t-1}\|_2^2 - \frac{\Gamma_y}{2} \sum_{t=t_\star+1}^{T} \|y^t - y^{t-1}\|_2^2.$$

Combining this with the nonnegativity in Eq. 66 proves Eq. 63. □

**Theorem B.7** (Restatement of Theorem 4.2)**.** *In any two-player zero-sum matrix game, if both players run Algorithm 1 with gradient feedback, then there exist finite game- and trajectory-dependent constants $\mathsf{C}_x, \mathsf{C}_y$ such that, for all sufficiently large $T$,*

$$\boxed{\mathrm{Reg}_x(T) \le \mathsf{C}_x (T+1)^\alpha, \qquad \mathrm{Reg}_y(T) \le \mathsf{C}_y (T+1)^\alpha.} \tag{70}$$

*Consequently, for the uniform averages $\bar{x}^T := \frac{1}{T}\sum_{t=1}^T x^t$, $\bar{y}^T := \frac{1}{T}\sum_{t=1}^T y^t$, we have*

$$\boxed{\mathrm{Gap}(\bar{x}^T, \bar{y}^T) \le \frac{(\mathsf{C}_x + \mathsf{C}_y)(T+1)^\alpha}{T} = O(T^{-(1-\alpha)}).} \tag{71}$$

*Proof.* We first consider the case $\tau_x, \tau_y < \infty$. Let $t_\star$ and $K_\star$ be as in Proposition B.6. The radius recursion Eq. 40, together with $m_x^t = u_x^{t-1}$ and the centering bound, gives for every $T \ge t_\star + 1$,

$$(\sigma_x^{T+1})^2 \le (\sigma_x^{t_\star+1})^2 + C_{d_x}^2 L^2 \sum_{t=t_\star+1}^{T} \frac{\|y^t - y^{t-1}\|_2^2}{g_t^2} \le (\sigma_x^{t_\star+1})^2 + \frac{2C_{d_x}^2 L^2 K_\star}{\Gamma_y} =: \overline{\sigma}_x^2, \tag{72}$$

where the second inequality uses $g_t \ge 1$ and Eq. 63. Symmetrically,

$$(\sigma_y^{T+1})^2 \le (\sigma_y^{t_\star+1})^2 + \frac{2C_{d_y}^2 L^2 K_\star}{\Gamma_x} =: \overline{\sigma}_y^2. \tag{73}$$

Thus the effective radii are uniformly bounded in time after the finite prefix. By Theorem B.1,

$$\mathrm{Reg}_i(T) \le S_i^{T+1} = g_{T+1}\sigma_i^{T+1} \le \overline{\sigma}_i (T+1)^\alpha, \qquad i \in \{x, y\}. \tag{74}$$

Hence Eq. 70 holds in the two-active branch with $\mathsf{C}_i = \overline{\sigma}_i$.

We next treat branches containing a never-active player. Suppose without loss of generality that $\tau_x = \infty$. By Theorem B.4, every $u_x^t$ is a constant vector, so $r_x^t = 0$ and therefore

$$\mathrm{Reg}_x(T) = 0. \tag{75}$$

Moreover, $x^t = a_x$ for every $t$, hence $u_y^t = -A^\top a_x =: v_y$ is time independent. If $v_y$ is a constant vector, then $\tau_y = \infty$ and $\mathrm{Reg}_y(T) = 0$ as well. Otherwise Lemma 3.2 gives $\tau_y = 2$. For every $t \ge 2$ we then have $m_y^t = u_y^{t-1} = u_y^t$, so $E_y^t = 0$. Consequently $\sigma_y^t = \sigma_y^2$ for all $t \ge 2$, and

$$\mathrm{Reg}_y(T) \le S_y^{T+1} = \sigma_y^2 (T+1)^\alpha. \tag{76}$$

The case $\tau_y = \infty$ is symmetric. This proves Eq. 70 in all branches.

Finally, for the uniform averages, the zero-sum regret identity follows directly from the definitions:

$$\mathrm{Reg}_x(T) + \mathrm{Reg}_y(T) = T\left(\max_{p \in \Delta_m} p^\top A \bar{y}^T - \min_{q \in \Delta_n} (\bar{x}^T)^\top A q\right) = T\,\mathrm{Gap}(\bar{x}^T, \bar{y}^T). \tag{77}$$

Combining Eq. 77 with Eq. 70 proves Eq. 71. □

# C Both Players Activate in Finite Time: Two-Time-Scale Lyapunov Analysis

Define the prediction movement and correction-normalization movement

$$a_t^2 = \|x^t - \widetilde{x}^t\|_2^2 + \|y^t - \widetilde{y}^t\|_2^2, \tag{78}$$

$$b_t^2 = \|\widetilde{x}^{t+1} - x^t\|_2^2 + \|\widetilde{y}^{t+1} - y^t\|_2^2, \qquad t \ge \tau, \tag{79}$$

$$\mathcal{V}_t = \sum_{i \in \{x,y\}} \frac{\|E_i^t\|_2^2}{g_t S_i^t}, \qquad t \ge \tau. \tag{80}$$

**Lemma C.1** (Exact index bound for correction normalization). *For $t \geq \tau$,*

$$\boxed{b_t^2 \leq \frac{C_D^2}{\underline{\sigma}} \mathcal{V}_t.} \tag{81}$$

*Proof.* Since $R_i^{t+1} = \lambda_t H_i^t$, positive radial scaling does not change normalization, so $\widetilde{z}_i^{t+1} = \mathcal{N}(H_i^t)$; meanwhile $z_i^t = \mathcal{N}(Q_i^t)$. Lemma A.2 gives

$$\|\widetilde{z}_i^{t+1} - z_i^t\|_2^2 \leq C_D^2 \frac{\|E_i^t\|_2^2}{(S_i^t)^2}.$$

Since $S_i^t = g_t \sigma_i^t$ and $\sigma_i^t \geq \underline{\sigma}$, we have $\frac{\|E_i^t\|_2^2}{(S_i^t)^2} \leq \frac{1}{\underline{\sigma}} \frac{\|E_i^t\|_2^2}{g_t S_i^t}$. Summing over the two players proves the claim. □

### C.1 Normalized Perspective Potential

**Lemma C.2** (One-step inequality for the normalized potential). *Let $\overline{\Delta}_t = \frac{\delta_x^t + \delta_y^t}{g_t}$. Then, for $t \geq \tau$,*

$$\Psi_{t+1}^* \leq \Psi_t^* - \overline{\Delta}_t + \frac{1}{2}\mathcal{V}_t, \tag{82}$$

$$\overline{\Delta}_t \geq c_D \underline{\sigma}\, a_t^2. \tag{83}$$

*Proof.* We first prove Eq. 82. Fix any NE $z^\star = (x^\star, y^\star) \in Z^\star$. Since $t \geq \tau$, both players are in the positive branch, Applying Eq. 31 in Lemma A.4 with $w_i = z_i^\star$ ($i \in \{x, y\}$) yields

$$\Psi_{t+1}(z^\star) \leq \Psi_t(z^\star) + \frac{1}{g_t} \sum_{i \in \{x,y\}} \langle u_i^t, z_i^t - z_i^\star \rangle - \frac{\delta_x^t + \delta_y^t}{g_t} + \frac{1}{2}\mathcal{V}_t. \tag{84}$$

According to the definition $\frac{\delta_x^t + \delta_y^t}{g_t} = \overline{\Delta}_t$, it remains to deal with term $= \frac{1}{g_t} \sum_{i \in \{x,y\}} \langle u_i^t, z_i^t - z_i^\star \rangle$ in Eq. 84. In matrix-game self-play, $u_x^t = Ay^t, \quad u_y^t = -A^\top x^t$, hence

$$\sum_{i \in \{x,y\}} \langle u_i^t, z_i^t - z_i^\star \rangle = \langle Ay^t, x^t - x^\star \rangle + \langle -A^\top x^t, y^t - y^\star \rangle = (x^t)^\top A y^\star - (x^\star)^\top A y^t \leq 0, \tag{85}$$

where the last inequality holds due to the fact that for any profile strategy $(x^t, y^t)$,

$$(x^t)^\top A y^\star \leq (x^*)^\top A y^* \leq (x^\star)^\top A y^t,$$

Substituting all the preceding terms back into Eq. 84, for any fixed $z^\star \in Z^\star$ we obtain

$$\Psi_{t+1}(z^\star) \leq \Psi_t(z^\star) - \overline{\Delta}_t + \frac{1}{2}\mathcal{V}_t. \tag{86}$$

Let $z_t^\star \in \arg\min_{z^\star \in Z^\star} \Psi_t(z^\star)$, Thus $\Psi_t(z_t^\star) = \Psi_t^*$. Substituting $z^\star = z_t^\star$ into Eq. 86,we have

$$\Psi_{t+1}^* = \min_{z^\star \in Z^\star} \Psi_{t+1}(z^\star) \leq \Psi_{t+1}(z_t^\star) \leq \Psi_t(z_t^\star) - \overline{\Delta}_t + \frac{1}{2}\mathcal{V}_t = \Psi_t^* - \overline{\Delta}_t + \frac{1}{2}\mathcal{V}_t, \tag{87}$$

which obtains Eq. 82. By Eq. 18, for a block of dimension $d_i$,

$$\delta_i^t = \frac{(S_i^t)^2}{2\|Q_i^t\|_1} \|\phi(z_i^t) - \phi(\widetilde{z}_i^t)\|_2^2 \geq \frac{S_i^t}{2\sqrt{d_i}(1+\sqrt{d_i})^2} \|z_i^t - \widetilde{z}_i^t\|_2^2. \tag{88}$$

Here we use $\|Q_i^t\|_1 \leq \sqrt{d_i} S_i^t$ and

$$\|\phi(p) - \phi(q)\|_2 \geq \frac{\|p - q\|_2}{1 + \sqrt{d_i}}. \tag{89}$$

Equation Eq. 89 follows from $p = \mathcal{N}(\phi(p))$ and the same normalization Lipschitz estimate used in the proof of Lemma A.2. Dividing both sides of Eq. 88 by $g_t$, and summing over $i \in \{x, y\}$ with $S_i^t / g_t \geq \underline{\sigma}$, yields Eq. 83. □

*Proof of Lemma 5.2.* Fix any Nash equilibrium $z^\star = (x^\star, y^\star) \in Z^\star$. Plugging $w = z^\star$ into Eq. 33 obtains

$$c_D\underline{\sigma}\|\widetilde{z}^t - z^\star\|_2^2 \le \Psi_t(z^\star) = \frac{1}{g_t}\sum_{i\in\{x,y\}} P(R_i^t, z_i^\star) \le 2D\overline{\sigma}\|\widetilde{z}^t - z^\star\|_2^2. \tag{90}$$

For the lower bound in Lemma 5.2 , let

$$z^\star_{\Psi,t} \in \arg\min_{z^\star\in Z^\star} \Psi_t(z^\star).$$

Then $\Psi_t(z^\star_{\Psi,t}) = \Psi_t^*$. Substituting $z^\star = z^\star_{\Psi,t}$ into the left-hand inequality of Eq. 90, we obtain

$$\Psi_t^* = \Psi_t(z^\star_{\Psi,t}) \ge c_D\underline{\sigma}\|\widetilde{z}^t - z^\star_{\Psi,t}\|_2^2 \ge c_D\underline{\sigma}\,\mathrm{dist}(\widetilde{z}^t, Z^\star)^2. \tag{91}$$

where the last inequality uses the fact that $z^\star_{\Psi,t} \in Z^\star$.

For the upper bound in Lemma 5.2, take a nearest equilibrium

$$z^\star_{\mathrm{dist},t} \in \arg\min_{z^\star\in Z^\star} \|\widetilde{z}^t - z^\star\|_2.$$

Then $\|\widetilde{z}^t - z^\star_{\mathrm{dist},t}\|_2 = \mathrm{dist}(\widetilde{z}^t, Z^\star)$. Since $\Psi_t^*$ is the minimum of the potential over $Z^\star$, we have

$$\Psi_t^* \le \Psi_t(z^\star_{\mathrm{dist},t}) \le 2D\overline{\sigma}\|\widetilde{z}^t - z^\star_{\mathrm{dist},t}\|_2^2 = 2D\overline{\sigma}\,\mathrm{dist}(\widetilde{z}^t, Z^\star)^2. \tag{92}$$

Combining Eq. 91 and Eq. 92 gives exactly the final result. □

### C.2 Lagged Recursion for the Past-Loss Correction

**Lemma C.3** (Two-time-scale recursion)**.** *For $t \ge \tau + 1$,*

$$\boxed{\mathcal{V}_t \le \frac{2C_D^2L^2}{t^{2\alpha}\underline{\sigma}}a_t^2 + \frac{2C_D^4L^2}{t^{2\alpha}\underline{\sigma}^2}\mathcal{V}_{t-1}.} \tag{93}$$

*Proof.* By definition, $\mathcal{V}_s$ and $b_s$ are defined only for $s \ge \tau$; and this recursion contains both $\mathcal{V}_{t-1}$ and $b_{t-1}$, so we must require $t - 1 \ge \tau$. At such times, both players are in the positive branch at rounds $t$ and $t - 1$, and the algorithm's past-loss prediction is $m_i^t = u_i^{t-1} = U_i(z^{t-1})$.

In a zero-sum matrix game,

$$U_x(z) = Ay, \qquad U_y(z) = -A^\top x,$$

Hence the prediction error satisfies

$$e_x^t = u_x^t - m_x^t = A(y^t - y^{t-1}), \tag{94}$$

$$e_y^t = u_y^t - m_y^t = -A^\top(x^t - x^{t-1}). \tag{95}$$

Define the centering operator $\mathcal{C}_z(v) = v - \langle v, z\rangle\mathbf{1}$. It is immediate that if $z \in \Delta_d$, then for any $v \in \mathbb{R}^d$,

$$\|\mathcal{C}_z(v)\|_2 \le (1 + \sqrt{d})\|v\|_2 \le C_D\|v\|_2,$$

where $C_D = 1 + \sqrt{D}$. Therefore $\|\mathcal{C}_z\|_{\mathrm{op}} \le C_D$. Now let $L = \|A\|_{\mathrm{op}} = \|A^\top\|_{\mathrm{op}}$. Then

$$\|E_x^t\|_2 = \|\mathcal{C}_z(e_x^t)\|_2 \le C_DL\|y^t - y^{t-1}\|_2, \tag{96}$$

$$\|E_y^t\|_2 = \|\mathcal{C}_z(e_y^t)\|_2 \le C_DL\|x^t - x^{t-1}\|_2. \tag{97}$$

On the other hand, $S_i^t = g_t\sigma_i^t$ and $g_t = t^\alpha$. Thus, when $t \ge \tau$, $\sigma_i^t \ge \underline{\sigma}$ and

$$g_tS_i^t = g_t^2\sigma_i^t \ge t^{2\alpha}\underline{\sigma}. \tag{98}$$

Substituting Eq. 96, Eq. 97 and Eq. 98 into the definition of $\mathcal{V}_t$, we obtain step by step

$$\mathcal{V}_t = \frac{\|E_x^t\|_2^2}{g_tS_x^t} + \frac{\|E_y^t\|_2^2}{g_tS_y^t} \le \frac{C_D^2L^2}{t^{2\alpha}\underline{\sigma}}\left(\|y^t - y^{t-1}\|_2^2 + \|x^t - x^{t-1}\|_2^2\right) = \frac{C_D^2L^2}{t^{2\alpha}\underline{\sigma}}\|z^t - z^{t-1}\|_2^2. \tag{99}$$

Next, using the basic inequality $(a+b)^2 \le 2a^2 + 2b^2$, we have

$$\|z^t - z^{t-1}\|_2^2 \le 2\|z^t - \widetilde{z}^t\|_2^2 + 2\|\widetilde{z}^t - z^{t-1}\|_2^2 = 2a_t^2 + 2b_{t-1}^2. \tag{100}$$

Substituting Eq. 100 into Eq. 99, we get

$$\mathcal{V}_t \le \frac{2C_D^2 L^2}{t^{2\alpha}\underline{\sigma}} a_t^2 + \frac{2C_D^2 L^2}{t^{2\alpha}\underline{\sigma}} b_{t-1}^2. \tag{101}$$

Now use Lemma C.1. For any $t \ge \tau + 1$, we have

$$b_{t-1}^2 \le \frac{C_D^2}{\underline{\sigma}} \mathcal{V}_{t-1}. \tag{102}$$

Substituting Eq. 102 into Eq. 101 gives

$$\begin{aligned}\mathcal{V}_t &\le \frac{2C_D^2 L^2}{t^{2\alpha}\underline{\sigma}} a_t^2 + \frac{2C_D^2 L^2}{t^{2\alpha}\underline{\sigma}} \left(\frac{C_D^2}{\underline{\sigma}} \mathcal{V}_{t-1}\right) \\ &= \frac{2C_D^2 L^2}{t^{2\alpha}\underline{\sigma}} a_t^2 + \frac{2C_D^4 L^2}{t^{2\alpha}\underline{\sigma}^2} \mathcal{V}_{t-1},\end{aligned}$$

This is exactly Eq. 93. □

### C.3 Automatic Stabilization Time and Genuine Round-by-Round Descent

By Lemma C.3 and Eq. 83, for $t \ge T_0$,

$$\mathcal{V}_t \le \frac{1}{2}\overline{\Delta}_t + \frac{1}{4}\mathcal{V}_{t-1}. \tag{103}$$

**Lemma C.4** (Simultaneous boundedness of the Lyapunov function)**.** *For $t \ge T_0$,*

$$\Phi_{t+1} \le \Phi_t - \frac{1}{2}\overline{\Delta}_t - \frac{1}{4}\mathcal{V}_{t-1}, \tag{104}$$

$$\Phi_{t+1} + \Sigma_{t+1} \le \Phi_t + \Sigma_t - \frac{1}{4}\overline{\Delta}_t - \frac{1}{8}\mathcal{V}_{t-1}, \tag{105}$$

*where $\Sigma_t = \sigma_x^t + \sigma_y^t$. Therefore*

$$\sigma_i^t \le \Sigma_t \le \overline{\sigma}, \qquad \overline{\sigma} = \Phi_{T_0} + \Sigma_{T_0}. \tag{106}$$

*Proof.* By Eq. 82 and $\Phi_t = \Psi_t^* + \frac{1}{2}\mathcal{V}_{t-1}$, we have

$$\Phi_{t+1} = \Psi_{t+1}^* + \frac{1}{2}\mathcal{V}_t \le \Psi_t^* - \overline{\Delta}_t + \mathcal{V}_t = \Phi_t - \overline{\Delta}_t + \mathcal{V}_t - \frac{1}{2}\mathcal{V}_{t-1}.$$

By the stable-stage recursion Eq. 103, $\mathcal{V}_t \le \frac{1}{2}\overline{\Delta}_t + \frac{1}{4}\mathcal{V}_{t-1}$, Therefore

$$\begin{aligned}\Phi_{t+1} &\le \Phi_t - \overline{\Delta}_t + \frac{1}{2}\overline{\Delta}_t + \frac{1}{4}\mathcal{V}_{t-1} - \frac{1}{2}\mathcal{V}_{t-1} \\ &= \Phi_t - \frac{1}{2}\overline{\Delta}_t - \frac{1}{4}\mathcal{V}_{t-1},\end{aligned}$$

which gives Eq. 104. We next prove Eq. 105. The key is to control the effective radii $\Sigma_t = \sum_i \sigma_i^t, =$ $\sum_i \frac{S_i^t}{g_t}$. From the radial update $R_i^{t+1} = \frac{g_{t+1}}{g_t} H_i^t$, we obtain $\sigma_i^{t+1} = \frac{\|H_i^t\|_2}{g_t}$. Therefore

$$\sigma_i^{t+1} - \sigma_i^t = \frac{\|H_i^t\|_2 - S_i^t}{g_t} \le \frac{\sqrt{(S_i^t)^2 + \|E_i^t\|_2^2} - S_i^t}{g_t} \le \frac{\|E_i^t\|_2^2}{2g_t S_i^t}. \tag{107}$$

where the first inequality holds due to $\|H_i^t\|_2^2 \le (S_i^t)^2 + \|E_i^t\|_2^2$ in Eq. 26, and the second inequality since when $t \ge T_0$, $S_i^t > 0$ so $\sqrt{S^2 + r} \le S + \frac{r}{2S}$ for any $S > 0, r \ge 0$. Summing over the two players, we obtain

$$\Sigma_{t+1} - \Sigma_t \le \frac{1}{2}\sum_i \frac{\|E_i^t\|_2^2}{g_t S_i^t} = \frac{1}{2}\mathcal{V}_t.$$

Using Eq. 103 again,

$$\Sigma_{t+1} - \Sigma_t \le \frac{1}{4}\overline{\Delta}_t + \frac{1}{8}\mathcal{V}_{t-1}.$$

Adding this to Eq. 104 gives

$$\Phi_{t+1} + \Sigma_{t+1} = \Phi_t + \Sigma_t - \frac{1}{4}\overline{\Delta}_t - \frac{1}{8}\mathcal{V}_{t-1}.$$

Since $\Delta_t \ge 0$ and $\mathcal{V}_{t-1} \ge 0$, the sequence $\{\Phi_t + \Sigma_t\}_{t \ge T_0}$ is nonincreasing, so

$$\Phi_t + \Sigma_t \le \Phi_{T_0} + \Sigma_{T_0} =: \overline{\sigma},$$

hence $\sigma_i^t \le \Sigma_t \le \overline{\sigma}$. □

### C.4 From Dissipation to the Current Last-Iterate Gap

We show the equivalence between duality gap and Euclidean distance. The lower bound $\kappa_A \operatorname{dist}(z, Z^\star) \le \operatorname{Gap}(z) \le \sqrt{2}L \operatorname{dist}(z, Z^\star)$ actually can be dated back to (Hoffman, 1952) and recently has been defined as *metric subregularity* by (Wei et al., 2021) in game theory.

**Lemma C.5** (Equivalence between Gap Duality and Euclidean distance)**.** *There exists a game-dependent constant $\kappa_A > 0$ such that*

$$\kappa_A \operatorname{dist}(z, Z^\star) \le \operatorname{Gap}(z) \le \sqrt{2}L \operatorname{dist}(z, Z^\star) \qquad \forall z \in \Delta_m \times \Delta_n. \tag{108}$$

*Proof.* First note that the matrix-game gap can be written as the maximum of finitely many affine functions:

$$\operatorname{Gap}(x, y) = \max_{j,k}\{(Ay)_j - (x^\top A)_k\}.$$

Because the Nash set is defined by finitely many linear inequalities together with simplex constraints, it is a polyhedral set, and the Hoffman error bound yields a constant $H_A < \infty$ depending on the payoff matrix $A$ such that

$$\operatorname{dist}(z, Z^\star) \le H_A \operatorname{Gap}(z).$$

Let $\kappa_A = H_A^{-1}$, Then we obtain $\operatorname{Gap}(z) \ge \kappa_A \operatorname{dist}(z, Z^\star)$.

We next prove the upper bound. For any fixed $z^\star = (x^\star, y^\star) \in Z^\star$. Employing the definition of NE $(x^\star, y^\star)$ and the Lipschitz property:

$$\max_p p^\top Ay - (x^\star)^\top Ay^\star \le \max_p \left(p^\top Ay - p^\top Ay^\star\right) + \max_p \left(p^\top Ay^\star - (x^\star)^\top Ay^\star\right) \le L\|y - y^\star\|_2,$$

Analogously, we can obtain $(x^\star)^\top Ay^\star - \min_q x^\top Aq \le L\|x - x^\star\|_2$. Therefore

$$\operatorname{Gap}(x, y) \le L\|x - x^\star\|_2 + L\|y - y^\star\|_2 \le \sqrt{2}L\|z - z^\star\|_2, \tag{109}$$

where the second inequality uses $a + b \le \sqrt{2}\sqrt{a^2 + b^2}$. Taking a nearest point in $Z^\star$ proves the claim. □

*Proof of Lemma 5.3.* Let the gap contributions of the two player blocks be

$$\operatorname{Gap}_i(\widetilde{z}^t) = \max_{p_i \in \Delta_{d_i}} \langle m_i^t, \widetilde{z}_i^t - p_i\rangle.$$

Plug the decomposition $\widetilde{z}_i^t - p_i = (z_i^t - p_i) + (\widetilde{z}_i^t - z_i^t)$ Substituting gives

$$\operatorname{Gap}_i(\widetilde{z}^t) \le \max_{p_i} \langle m_i^t, z_i^t - p_i\rangle + \|m_i^t\|_2 \|\widetilde{z}_i^t - z_i^t\|_2.$$

We first control the first term:

$$\max_{p_i} \langle m_i^t, z_i^t - p_i\rangle \le (2\sqrt{d_i}S_i^t + L)\|z_i^t - \widetilde{z}_i^t\|_2 \le 4\sqrt{D}\, g_t\overline{\sigma}\|z_i^t - \widetilde{z}_i^t\|_2. \tag{110}$$

where the first inequality follows from Lemma A.1, and the second follows from Lemma C.4: $S_i^t = g_t\sigma_i^t \le g_t\overline{\sigma}$. Also, by the stabilization-time condition, $S_i^t \ge g_t\underline{\sigma} \ge L$. hence

$$2\sqrt{d_i}S_i^t + L \le 4\sqrt{D}\, g_t\overline{\sigma}.$$

Summing over the two blocks and using Cauchy–Schwartz,

$$\sum_i \|z_i^t - \widetilde{z}_i^t\|_2 \le \sqrt{2} a_t,$$

we obtain

$$\sum_i \max_{p_i} \langle m_i^t, z_i^t - p_i \rangle \le 4\sqrt{2D}\, g_t \overline{\sigma} a_t.$$

We next control the prediction-error term $\sum_i \|m_i^t\|_2 \|\widetilde{z}_i^t - z_i^t\|_2$. Since $\|m_i^t\|_2 = \|u_i^{t-1}\|_2 \le L$, we have

$$\sum_i \|m_i^t\|_2 \|\widetilde{z}_i^t - z_i^t\|_2 \le 2L \|\widetilde{z}^t - z^t\|_2 = 2L a_t. \tag{111}$$

Combining Eq. 110 and Eq. 111 gives: $\mathrm{Gap}(\widetilde{z}^t) \le 4\sqrt{2D}\, g_t \overline{\sigma} a_t + 2L a_t$. □

**Theorem C.6** (Explicit last-iterate rate in the both-active branch)**.** *Define*

$$K_1 = \frac{2D\overline{\sigma} K_0^2}{\kappa_A^2}, \qquad \lambda_0 = \min\left\{\frac{c_D \underline{\sigma}}{2}, \frac{1}{4}\right\}, \qquad \mu_A = \min\left\{\frac{1}{2}, \frac{\lambda_0}{K_1}\right\}. \tag{112}$$

*Then for all $t \ge T_0$,*

$$\Phi_t \le \Phi_{T_0} \exp[-\mu_A \mathcal{H}_\alpha(t, T_0)], \tag{113}$$

$$\mathrm{Gap}(z^t) \le C_A \exp\Big[-\frac{\mu_A}{2} \mathcal{H}_\alpha(t, T_0)\Big], \tag{114}$$

*where*

$$C_A = \sqrt{2} L \left(\frac{1}{\sqrt{c_D \underline{\sigma}}} + \frac{1}{\sqrt{\lambda_0}}\right) \sqrt{\Phi_{T_0}}. \tag{115}$$

*Equivalently, for $0 < \alpha < 1/2$, $\mathrm{Gap}(z^t) = \exp\{-\Theta(t^{1-2\alpha})\}$ up to game-dependent constants; at $\alpha = 1/2$, $\mathrm{Gap}(z^t) \le C_A (T_0/t)^{\mu_A/2}$.*

*Proof.* First, by Lemma 5.2, Eq. 108, and Lemma 5.3, we obtain the following upper bound on the potential:

$$\begin{aligned}
\Phi_t = \Psi_t^* + \frac{1}{2}\mathcal{V}_{t-1} &\le 2D\overline{\sigma}\, \mathrm{dist}(\widetilde{z}^t, Z^\star)^2 + \frac{1}{2}\mathcal{V}_{t-1} \\
&\le K_1 g_t^2 a_t^2 + \frac{1}{2}\mathcal{V}_{t-1} \\
&\le K_1 g_t^2 (a_t^2 + \mathcal{V}_{t-1}).
\end{aligned}$$

Therefore $a_t^2 + \mathcal{V}_{t-1} \ge \frac{\Phi_t}{K_1 g_t^2}$. On the other hand, by the potential-descent relation Eq. 104 and the dissipation lower bound Eq. 83, we have

$$\Phi_{t+1} \le \Phi_t - \lambda_0 (a_t^2 + \mathcal{V}_{t-1}) \le \Phi_t - \frac{\lambda_0}{K_1 g_t^2} \Phi_t = \left(1 - \frac{\lambda_0}{K_1 g_t^2}\right) \Phi_t. \tag{116}$$

Recall that $\mu_A = \min\left\{\frac{1}{2}, \frac{\lambda_0}{K_1}\right\}$. Using $g_t^2 = t^{2\alpha}$, we obtain

$$\Phi_{t+1} \le \left(1 - \frac{\mu_A}{t^{2\alpha}}\right) \Phi_t.$$

Iterating the preceding inequality yields

$$\begin{aligned}
\Phi_t &\le \Phi_{T_0} \prod_{s=T_0}^{t-1} \left(1 - \frac{\mu_A}{s^{2\alpha}}\right) \le \Phi_{T_0} \exp\left(-\mu_A \sum_{s=T_0}^{t-1} s^{-2\alpha}\right) \\
&\le \Phi_{T_0} \exp[-\mu_A \mathcal{H}_\alpha(t, T_0)].
\end{aligned}$$

The last inequality uses $\sum_{s=T_0}^{t-1} s^{-2\alpha} \ge \int_{T_0}^{t} r^{-2\alpha}\, dr = \mathcal{H}_\alpha(t, T_0)$.

Next we control the distance between the actual strategy and the equilibrium set. By the lower bound on the potential, $\Phi_t \geq \Psi_t^* \geq c_{D\underline{\sigma}} \operatorname{dist}(\widetilde{z}^t, Z^\star)^2$, Therefore

$$\operatorname{dist}(\widetilde{z}^t, Z^\star)^2 \leq \frac{\Phi_t}{c_{D\underline{\sigma}}}.$$

Meanwhile, using Eq. 116 and $\Phi_{t+1} \geq 0$, we have

$$\lambda_0 a_t^2 \leq \lambda_0(a_t^2 + \mathcal{V}_{t-1}) \leq \Phi_t - \Phi_{t+1} \leq \Phi_t,$$

and hence $a_t^2 \leq \frac{\Phi_t}{\lambda_0}$. Furthermore, by Lemma 5.2 and the triangle inequality,

$$\begin{aligned}\operatorname{dist}(z^t, Z^\star) &\leq \|z^t - \widetilde{z}^t\|_2 + \operatorname{dist}(\widetilde{z}^t, Z^\star) \\ &= a_t + \operatorname{dist}(\widetilde{z}^t, Z^\star) \\ &\leq \left(\frac{1}{\sqrt{\lambda_0}} + \frac{1}{\sqrt{c_{D\underline{\sigma}}}}\right)\sqrt{\Phi_t}. \end{aligned} \tag{117}$$

Finally, combining Eq. 117, Eq. 108 in Lemma C.5, and Eq. 113 gives Eq. 114. □

*Proof of Theorem 5.5.* This theorem merely reorganizes the result for the both-active branch stated above. By Theorem C.6, when $t \geq T_0$, the Lyapunov potential satisfies

$$\Phi_t \leq \Phi_{T_0} \exp[-\mu_A \mathcal{H}_\alpha(t, T_0)],$$

and the gap of the actual iterate satisfies

$$\operatorname{Gap}(z^t) \leq C_A \exp[-(\mu_A/2)\mathcal{H}_\alpha(t, T_0)].$$ □

## D The Zero-Branch and Convergence of the Full Algorithm

**Theorem D.1** (Neither player ever activates)**.** *If $\tau_x = \tau_y = \infty$, then $(a_x, a_y) \in Z^\star$, and $z^t = (a_x, a_y)$ and $\operatorname{Gap}(z^t) = 0$ for all $t$.*

*Proof.* At round 1, both players are in the zero branch. According to Algorithm 1, $z^1 = (x^1, y^1) = (a_x, a_y)$. the utility vectors received by the two players at round 1 are therefore

$$u_x^1 = Aa_y, \qquad u_y^1 = -A^\top a_x.$$

By the assumption $\tau_x = \tau_y = \infty$, and Lemma 3.2, in the zero branch, a block remains inactive after the current round if and only if the utility vector it receives is constant. Therefore there exist scalars $c_x, c_y \in \mathbb{R}$ such that

$$Aa_y = c_x \mathbf{1}, \qquad -A^\top a_x = c_y \mathbf{1}.$$

First consider the row player. For any $x \in \Delta_{d_x}$,

$$x^\top A a_y = x^\top (c_x \mathbf{1}) = c_x.$$

In particular, $a_x^\top A a_y = c_x$. Thus all pure strategies, and hence all mixed strategies, obtain the same payoff when the opponent's strategy is fixed at $a_y$, so $a_x \in \arg\max_{x \in \Delta_{d_x}} x^\top A a_y$.

Similarly, for any $y \in \Delta_{d_y}$, $a_y \in \arg\min_{y \in \Delta_{d_y}} a_x^\top A y$.

Therefore $a_x$ is a best response to $a_y$, and $a_y$ is a best response to $a_x$. Hence

$$(a_x, a_y) \in Z^\star.$$

We next prove that the entire trajectory remains at this anchor profile forever. Since both utility vectors at round 1 are constant, Lemma 3.2 gives

$$H_x^1 = H_y^1 = 0.$$

Therefore in the next round we still have

$$R_x^2 = R_y^2 = 0,$$

so the algorithm again chooses $z^2 = (a_x, a_y)$. Repeating this argument and applying induction gives

$$R_x^t = R_y^t = 0, \qquad z^t = (a_x, a_y), \qquad \forall t \geq 1.$$

Since $(a_x, a_y) \in Z^\star$, we finally obtain $\operatorname{Gap}(z^t) = 0, \qquad \forall t \geq 1.$ □

**Notation for the fixed-utility branch.** Given a nonconstant vector $v \in \mathbb{R}^d$, define

$$P_v^\star = \arg\max_{p\in\Delta_d} \langle v, p\rangle, \qquad g_v(p) = \max_j v_j - \langle v, p\rangle.$$

Since $P_v^\star$ is the optimal face of a linear program over the simplex, the Hoffman error bound implies that there exists $\kappa_v > 0$ such that

$$g_v(p) \geq \kappa_v \ \mathrm{dist}(p, P_v^\star), \qquad p \in \Delta_d.$$

**Lemma D.2** (Single-block last-iterate convergence in the fixed-utility positive-state branch)**.** *Consider a single block of Algorithm 1. Assume there exists a finite time $\tau$ such that for all $t \geq \tau$, $S^t > 0, u^t = m^t = v$, where $v \in \mathbb{R}^d$ is nonconstant. Then $\sigma^t = S^t/g_t \equiv \sigma > 0$. Moreover, there exist explicit constants $T_v < \infty$, $\mu_v > 0$, and $C_v > 0$ such that for all $t \geq T_v$,*

$$\boxed{g_v(z^t) \leq C_v \exp\Big[-\frac{\mu_v}{2}\mathcal{H}_\alpha(t, T_v)\Big].}$$

*Specifically, one may take*

$$T_v \geq \tau : \ g_{T_v}\sigma \geq \|v\|_2, \qquad \mu_v = \min\left\{\frac{1}{2}, \frac{c_d\kappa_v^2}{18d^2\sigma^2}\right\}, \qquad C_v = \frac{2\|v\|_2}{\sqrt{c_d}\sigma}\sqrt{\psi^\star_{T_v}}.$$

*Proof.* **Step 1: The correction vanishes and the normalized radius remains constant.**

For any $t \geq \tau$, since $u^t = m^t = v$, we have $e^t = u^t - m^t = 0$. Thus the centered correction is $E^t = e^t - \langle e^t, z^t\rangle \mathbf{1} = 0$. Hence

$$H^t = [Q^t + E^t]_+ = Q^t, \qquad S^{t+1} = \|R^{t+1}\|_2 = \frac{g_{t+1}}{g_t}\|Q^t\|_2 = \frac{g_{t+1}}{g_t}S^t.$$

Therefore

$$\sigma^{t+1} = \frac{S^{t+1}}{g_{t+1}} = \frac{S^t}{g_t} = \sigma^t.$$

Consequently, there exists a constant $\sigma > 0$ such that $\sigma^t \equiv \sigma, \qquad S^t = g_t\sigma, \qquad t \geq \tau.$

**Step 2: Strict one-step descent of the perspective potential.**

Fix any

$$p^\star \in P_v^\star = \arg\max_{p\in\Delta_d} \langle v, p\rangle.$$

Since $p^\star$ maximizes the linear objective $\langle v, p\rangle$, we have $\langle v, z^t - p^\star\rangle \leq 0$. Then, using Lemma A.3 and $E^t = 0$,

$$P(H^t, p^\star) \leq P(R^t, p^\star) + \langle v, z^t - p^\star\rangle - \delta_t \leq P(R^t, p^\star) - \delta_t. \tag{118}$$

On the other hand, $P(\cdot, p^\star)$ is positively homogeneous in its first argument, and $R^{t+1} = \frac{g_{t+1}}{g_t}H^t$, hence $\frac{P(R^{t+1},p^\star)}{g_{t+1}} = \frac{P(H^t,p^\star)}{g_t}$. Plugging such result into Eq. 118, we have

$$\frac{P(R^{t+1}, p^\star)}{g_{t+1}} \leq \frac{P(R^t, p^\star)}{g_t} - \frac{\delta_t}{g_t}.$$

Let $p_t^\star \in P_v^\star$ be a minimizer of $\psi_t^*$, i.e., $\psi_t^* = \frac{P(R^t,p_t^\star)}{g_t}$. Then

$$\psi_{t+1}^* = \min_{p^\star\in P_v^\star} \frac{P(R^{t+1}, p^\star)}{g_{t+1}} \leq \frac{P(R^{t+1}, p_t^\star)}{g_{t+1}} \leq \frac{P(R^t, p_t^\star)}{g_t} - \frac{\delta_t}{g_t} = \psi_t^* - \frac{\delta_t}{g_t}. \tag{119}$$

We next relate $\delta_t$ to the prediction movement $a_t = \|z^t - \widetilde{z}^t\|_2$ Using the equal-$\ell_2$-norm identity,

$$\delta_t\|Q^t\|_1 = \frac{1}{2}\|Q^t - R^t\|_2^2.$$

Since $R^t = S^t\phi(\widetilde{z}^t), \quad Q^t = S^t\phi(z^t)$, where $\phi(p) = \frac{p}{\|p\|_2}$, we obtain

$$\begin{aligned}\delta_t &= \frac{(S^t)^2}{2\|Q^t\|_1}\|\phi(z^t) - \phi(\widetilde{z}^t)\|_2^2 \\ &\geq \frac{S^t}{2\sqrt{d}}\|\phi(z^t) - \phi(\widetilde{z}^t)\|_2^2 \\ &\geq \frac{S^t}{2\sqrt{d}(1+\sqrt{d})^2}\|z^t - \widetilde{z}^t\|_2^2 \\ &= c_d S^t a_t^2 = c_d g_t \sigma a_t^2,\end{aligned}$$

where the first and the second inequalities hold since $\|Q^t\|_1 \leq \sqrt{d}\,\|Q^t\|_2 = \sqrt{d}\,S^t$ and $\|\phi(p) - \phi(q)\|_2 \geq \frac{\|p-q\|_2}{1+\sqrt{d}}, \qquad p, q \in \Delta_d$, Then, we have $\frac{\delta_t}{g_t} \geq c_d\sigma a_t^2$. Plug the above result into Eq. 119, we have

$$\boxed{\psi_{t+1}^* \leq \psi_t^* - c_d\sigma a_t^2.} \tag{120}$$

**Step 3: Use the current movement $a_t$ to control $g_v(\widetilde{z}^t)$.**

Because $g_v(p) = \max_j v_j - \langle v, p\rangle$, we have

$$g_v(\widetilde{z}^t) = g_v(z^t) + \langle v, z^t - \widetilde{z}^t\rangle \leq g_v(z^t) + \|v\|_2 a_t. \tag{121}$$

We now control $g_v(z^t)$. From $Q^t = [R^t + v - \gamma_t\mathbf{1}]_+$ we see that for every coordinate $j$, $Q_j^t \geq R_j^t + v_j - \gamma_t$, so

$$v_j - \gamma_t \leq Q_j^t - R_j^t.$$

On the other hand, $\delta_t = \langle v, z^t\rangle - \gamma_t \geq 0$. Therefore

$$v_j - \langle v, z^t\rangle = (v_j - \gamma_t) - \delta_t \leq Q_j^t - R_j^t.$$

Taking the maximum over $j$ gives

$$g_v(z^t) \leq \|Q^t - R^t\|_\infty \leq \|Q^t - R^t\|_2 = S^t\|\phi(z^t) - \phi(\widetilde{z}^t)\|_2 \leq 2\sqrt{d}\,S^t a_t, \tag{122}$$

where the last inequality uses the Lipschitz property of $\phi$. By the choice of $t \geq T_v$,

$$S^t = g_t\sigma \geq g_{T_v}\sigma \geq \|v\|_2,$$

Thus, combining Eq. 121 and Eq. 122 gives

$$\boxed{\begin{aligned}g_v(\widetilde{z}^t) &\leq 2\sqrt{d}\,S^t a_t + \|v\|_2 a_t \\ &\leq 3\sqrt{d}\,S^t a_t = 3\sqrt{d}\,g_t\sigma a_t.\end{aligned}} \tag{123}$$

**Step 4: A polyhedral error bound converts the gap into distance to the optimal face.**

Since $P_v^\star$ is the set of optimal solutions of a linear program over the simplex, it is a polyhedral face. Hence there exists a constant $\kappa_v > 0$ depending only on $v$ such that

$$g_v(p) \geq \kappa_v\,\mathrm{dist}(p, P_v^\star), \qquad \forall p \in \Delta_d.$$

Therefore, setting $p = \widetilde{z}^t$ and applying Eq. 123 gives

$$\mathrm{dist}(\widetilde{z}^t, P_v^\star) \leq \frac{g_v(\widetilde{z}^t)}{\kappa_v} \leq \frac{3\sqrt{d}\,g_t\sigma}{\kappa_v}a_t.$$

Combining the above result with Lemma 5.2 leads to

$$\psi_t^* \leq 2d\sigma\,\mathrm{dist}(\widetilde{z}^t, P_v^\star)^2 \leq 2d\sigma\left(\frac{3\sqrt{d}\,g_t\sigma}{\kappa_v}\right)^2 a_t^2 = \frac{18d^2\sigma^3}{\kappa_v^2}g_t^2 a_t^2 = K_v g_t^2 a_t^2.$$

Therefore

$$\boxed{\psi_t^* \le K_v g_t^2 a_t^2, \quad \textit{and} \quad a_t^2 \ge \frac{\psi_t^*}{K_v g_t^2}.} \tag{124}$$

**Step 5: Obtain a multiplicative contraction.**

Substituting the preceding lower bound into Eq. 120,

$$\psi_{t+1}^* \le \psi_t^* - c_d\sigma a_t^2 \le \psi_t^* - \frac{c_d\sigma}{K_v g_t^2}\psi_t^* = \left(1 - \frac{c_d\sigma}{K_v g_t^2}\right)\psi_t^* = \left(1 - \frac{c_d\sigma/K_v}{t^{2\alpha}}\right)\psi_t^*.$$

Define $\mu_v = \min\left\{\frac{1}{2}, \frac{c_d\sigma}{K_v}\right\}$. Then $0 < \mu_v \le 1/2$ and $\mu_v \le \frac{c_d\sigma}{K_v}$, Thus

$$\psi_{t+1}^* \le \left(1 - \frac{\mu_v}{t^{2\alpha}}\right)\psi_t^*. \tag{125}$$

Iterating this recursion yields

$$\boxed{\psi_t^* \le \psi_{T_v}^* \prod_{s=T_v}^{t-1}\left(1 - \frac{\mu_v}{s^{2\alpha}}\right) \le \psi_{T_v}^* \exp\left(-\mu_v \sum_{s=T_v}^{t-1} s^{-2\alpha}\right) \le \psi_{T_v}^* \exp[-\mu_v \mathcal{H}_\alpha(t, T_v)]\,.} \tag{126}$$

where the last inequality uses $\sum_{s=T_v}^{t-1} s^{-2\alpha} \ge \int_{T_v}^{t} r^{-2\alpha}\, dr = \mathcal{H}_\alpha(t, T_v)$.

**Step 6: Recover the last-iterate gap of the actual strategy from the perspective potential.**

From Eq. 120 and $\psi_{t+1}^* \ge 0$,we have $c_d\sigma a_t^2 \le \psi_t^*$, we have

$$a_t^2 \le \frac{\psi_t^*}{c_d\sigma}. \tag{127}$$

On the other hand, the single-block version of Lemma 5.2 gives

$$\operatorname{dist}(\widetilde{z}^t, P_v^\star) \le \sqrt{\frac{\psi_t^*}{c_d\sigma}}.$$

Employing the above results, triangle inequality, and $a_t = \|z^t - \widetilde{z}^t\|_2$, we have

$$\operatorname{dist}(z^t, P_v^\star) \le \|z^t - \widetilde{z}^t\|_2 + \operatorname{dist}(\widetilde{z}^t, P_v^\star) \le 2\sqrt{\frac{\psi_t^*}{c_d\sigma}}. \tag{128}$$

Finally, for any $p^\star \in P_v^\star$, $\langle v, p^\star\rangle = \max_j v_j$, we have $g_v(z^t) = \langle v, p^\star - z^t\rangle$. Choosing a nearest $p^\star \in P_v^\star$ to $z^t$, then applying the Cauchy–Schwarz inequality, Eq. 128, and Eq. 126 yields

$$\boxed{g_v(z^t) \le \|v\|_2 \operatorname{dist}(z^t, P_v^\star) \le \frac{2\|v\|_2}{\sqrt{c_d\sigma}}\sqrt{\psi_t^*} \le \frac{2\|v\|_2}{\sqrt{c_d\sigma}}\sqrt{\psi_{T_v}^*}\exp\Big[-\frac{\mu_v}{2}\mathcal{H}_\alpha(t, T_v)\Big]\,.} \quad \square$$

**Theorem D.3** (Exactly one player never activates)**.** *If exactly one player has $\tau_i = \infty$, then the profile gap satisfies an explicit game-dependent last-iterate bound of the form*

$$\operatorname{Gap}(z^t) \le \widehat{C}_A \exp\left[-\frac{\widehat{\mu}_A}{2}\mathcal{H}_\alpha(t, \widehat{T}_A)\right] \tag{129}$$

.

*Proof.* Without loss of generality, suppose the row player never activates, i.e.,

$$\tau_x = \infty, \qquad \tau_y < \infty.$$

By the algorithm's zero-branch rule, the row player remains at the anchor forever:

$$S_x^t = 0, \qquad x^t = a_x, \qquad \forall t \ge 1.$$

**Step 1: The row player's current best-response gap is identically zero.**

At round $t$, the row player receives the utility vector

$$u_x^t = Ay^t.$$

Since $\tau_x = \infty$, the row player must remain in the zero branch after every round. By Lemma 3.2, if $Ay^t$ were not a constant vector, then the zero-branch correction would satisfy $H_x^t \neq 0$, which would imply $S_x^{t+1} > 0$, contradicting $\tau_x = \infty$. Therefore, for every $t$, there exists a scalar $c_t$ such that

$$Ay^t = c_t \mathbf{1}.$$

Thus all pure row actions obtain the same payoff $c_t$ against $y^t$. In particular,

$$\max_{x \in \Delta_{d_x}} x^\top Ay^t = c_t = a_x^\top Ay^t.$$

Hence the row player's current best-response gap at the profile $(a_x, y^t)$ is identically zero.

**Step 2: The column player faces a fixed, nonconstant linear utility.**

Since $x^t = a_x$ for all $t$, the column player's maximizing utility vector at every round is

$$u_y^t = -A^\top x^t = -A^\top a_x =: v.$$

Thus $v$ is time-independent.

We claim that $v$ is not a constant vector. Otherwise, in the column player's initial zero branch, Lemma 3.2 would give $H_y^1 = 0$. Since $x^t = a_x$ remains unchanged thereafter, the column player's utility would continue to equal the same constant vector $v$, and induction would yield $S_y^t = 0$ for all $t$, i.e., $\tau_y = \infty$. This contradicts $\tau_y < \infty$ in the present branch. Therefore

$$v = -A^\top a_x$$

must be nonconstant.

In fact, by Lemma 3.2, the column player is activated upon first receiving this nonconstant utility. From the beginning of its positive-state phase, since

$$u_y^t = u_y^{t-1} = v,$$

the past-utility prediction satisfies

$$m_y^t = v, \qquad E_y^t = 0.$$

the column player is exactly in the fixed-utility positive-state single-block setting of Lemma D.2.

**Step 3: The two-player game gap is exactly the linear gap of this single block.**

By the definition of the zero-sum matrix-game gap,

$$\mathrm{Gap}(z^t) = \max_{x \in \Delta_{d_x}} x^\top Ay^t - \min_{y \in \Delta_{d_y}} (x^t)^\top Ay.$$

By Step 1 and $x^t = a_x$,

$$\max_{x \in \Delta_{d_x}} x^\top Ay^t = a_x^\top Ay^t.$$

On the other hand, from $v = -A^\top a_x$, we have $a_x^\top Ay = -\langle v, y \rangle$. Therefore

$$\begin{aligned}
\mathrm{Gap}(z^t) &= a_x^\top Ay^t - \min_{y \in \Delta_{d_y}} a_x^\top Ay \\
&= -\langle v, y^t \rangle + \max_{y \in \Delta_{d_y}} \langle v, y \rangle \\
&= \max_j v_j - \langle v, y^t \rangle = g_v(y^t) \\
&\leq C_v \exp\left[ -\frac{\mu_v}{2} \mathcal{H}_\alpha(t, T_v) \right], \qquad t \geq T_v,
\end{aligned}$$

where the last inequality uses Lemma D.2 with explicit $T_v < \infty$, $\mu_v > 0$, and $C_v > 0$.

Thus the constants in the theorem may be chosen as

$$\widehat{C}_A = C_v, \qquad \widehat{\mu}_A = \mu_v, \qquad \widehat{T}_A = T_v. \qquad \square$$

*Proof of Theorem 5.1.* If both players activate in finite time, apply Theorem C.6; if neither ever activates, apply Theorem D.1; if exactly one player never activates, apply Theorem D.3. Lemma 3.2 shows that these three cases exhaust all trajectories. □

# E Instance-Independent Upper/Lower Bounds for Last-Iterate Convergence Rate of Algorithm 1

First define

$$\mathcal{A}_{m,n}(B) = \left\{A \in \mathbb{R}^{m\times n} : |A_{jk}| \le B\right\}, \qquad \mathfrak{G}_T^{m,n}(B) = \sup_{A\in\mathcal{A}_{m,n}(B)} \mathrm{Gap}_A(z_A^T), \tag{130}$$

where $z_A^T$ denotes the actual strategy output at round $T$ by Algorithm 1 in self-play on matrix $A$.

*Proof of Theorem 6.1.* Given $T \ge 2$, define

$$\varepsilon_T = \frac{1-\alpha}{4}T^{-(1-\alpha)}, \qquad A_T = A_{\varepsilon_T}, \qquad A_\varepsilon = \begin{pmatrix} 1 & 1 \\ 1-\varepsilon & 1-\varepsilon \\ -1 & -1 \end{pmatrix}. \tag{131}$$

Clearly $0 < \varepsilon_T < 1/4$ and $A_T \in \mathcal{A}_{3,2}(1)$. Below fix $\varepsilon = \varepsilon_T$ and omit the subscript $T$.

**Step 1: The column player remains in the zero branch forever.** Let

$$v_\varepsilon = (1, 1-\varepsilon, -1)^\top. \tag{132}$$

Since the two columns of $A_\varepsilon$ are identical, for any $x \in \Delta_3$,

$$-A_\varepsilon^\top x = -\langle v_\varepsilon, x\rangle \mathbf{1}_2$$

is a constant vector. If the column player is in the zero branch at round $t$, then the algorithm defines $m_y^t = Q_y^t = 0$ and $y^t = a_y$, and

$$E_y^t = u_y^t - \langle u_y^t, a_y\rangle \mathbf{1}_2 = 0.$$

hence $H_y^t = 0$ and $R_y^{t+1} = 0$. Induction from $R_y^1 = 0$ gives

$$R_y^t = 0, \qquad y^t = a_y, \qquad \forall t \ge 1. \tag{133}$$

Therefore the row player receives the same fixed utility vector at every round,

$$u_x^t = A_\varepsilon y^t = v_\varepsilon, \qquad \forall t \ge 1. \tag{134}$$

**Step 2: The row player activates in the first round, and its radial scale can be computed exactly.** At the first round, $R_x^1 = 0$, $x^1 = a_x$, and $m_x^1 = Q_x^1 = 0$. From $\langle v_\varepsilon, a_x\rangle = \frac{1-\varepsilon}{3}$ we obtain

$$E_x^1 = v_\varepsilon - \langle v_\varepsilon, a_x\rangle \mathbf{1}_3 = \left(\frac{2+\varepsilon}{3}, \frac{2(1-\varepsilon)}{3}, -\frac{4-\varepsilon}{3}\right)^\top. \tag{135}$$

Therefore

$$H_x^1 = h_\varepsilon = \left(\frac{2+\varepsilon}{3}, \frac{2(1-\varepsilon)}{3}, 0\right)^\top, \qquad R_x^2 = g_2 h_\varepsilon. \tag{136}$$

Define $S_t = \|R_x^t\|_2, \qquad \sigma_t = \frac{S_t}{g_t}$. Since for $t \ge 2$, $m_x^t = u_x^{t-1} = v_\varepsilon = u_x^t$, we have $e_x^t = E_x^t = 0$. Hence

$$H_x^t = Q_x^t, \qquad R_x^{t+1} = \lambda_t Q_x^t.$$

Combining this with the equal-$\ell_2$-norm condition $\|Q_x^t\|_2 = S_t$ gives

$$S_{t+1} = \lambda_t S_t, \qquad \sigma_{t+1} = \sigma_t.$$

Thus for all $t \ge 2$,

$$\sigma_t \equiv \sigma_\varepsilon = \|h_\varepsilon\|_2. \tag{137}$$

Since $0 < \varepsilon \le 1/4$,

$$\sigma_\varepsilon \ge \frac{2(1-\varepsilon)}{3} \ge \frac{1}{2}. \tag{138}$$

**Step 3: The third coordinate is permanently zero from round 2 onward.** We prove

$$R_{x,3}^t = Q_{x,3}^t = x_3^t = 0, \qquad \forall t \geq 2. \tag{139}$$

At $t = 2$, Eq. 136 already gives $R_{x,3}^2 = 0$. Suppose that for some $t \geq 2$, $R_{x,3}^t = 0$. The positive-state prediction is

$$Q_x^t = [R_x^t + v_\varepsilon - \gamma_t \mathbf{1}_3]_+, \qquad \|Q_x^t\|_2 = \|R_x^t\|_2.$$

If $\gamma_t \leq -1$, then the first two coordinates strictly increase:

$$Q_{x,1}^t > R_{x,1}^t, \qquad Q_{x,2}^t > R_{x,2}^t,$$

because $1-\varepsilon-\gamma_t > 0$; hence $\|Q_x^t\|_2 > \|R_x^t\|_2$, contradicting the equal-$\ell_2$-norm condition. Therefore $\gamma_t > -1$, and thus

$$Q_{x,3}^t = [-1 - \gamma_t]_+ = 0.$$

Since $R_x^{t+1} = \lambda_t Q_x^t$, induction proves the claim. Eq. 139.

**Step 4: As long as the second coordinate remains active, obtain an exact one-dimensional slow-variable recursion.** Define

$$D_t = R_{x,1}^t - R_{x,2}^t, \qquad d_t = \frac{D_t}{g_t}. \tag{140}$$

If $Q_{x,2}^t > 0$, then together with Eq. 139, the first two prediction coordinates are not truncated by the positive part, so

$$Q_{x,1}^t = R_{x,1}^t + 1 - \gamma_t,$$
$$Q_{x,2}^t = R_{x,2}^t + 1 - \varepsilon - \gamma_t.$$

Subtracting the two equations gives

$$Q_{x,1}^t - Q_{x,2}^t = D_t + \varepsilon. \tag{141}$$

Then, using $R_x^{t+1} = \lambda_t Q_x^t$ and $\lambda_t = g_{t+1}/g_t$,

$$d_{t+1} = d_t + \frac{\varepsilon}{g_t}. \tag{142}$$

Meanwhile, Eq. 136 gives

$$d_2 = \frac{R_{x,1}^2 - R_{x,2}^2}{g_2} = \frac{2+\varepsilon}{3} - \frac{2(1-\varepsilon)}{3} = \varepsilon. \tag{143}$$

Therefore, as long as the second coordinate remains active in all preceding rounds,

$$d_t = \varepsilon \left(1 + \sum_{s=2}^{t-1} s^{-\alpha}\right), \qquad \frac{Q_{x,1}^t - Q_{x,2}^t}{g_t} = \varepsilon \left(1 + \sum_{s=2}^{t} s^{-\alpha}\right). \tag{144}$$

**Step 5: For $t \leq T$, the second coordinate is in fact never truncated.** For the decreasing function $x^{-\alpha}$ and $0 < \alpha < 1$,

$$1 + \sum_{s=2}^{t} s^{-\alpha} \leq 1 + \int_1^t x^{-\alpha}\, dx = 1 + \frac{t^{1-\alpha} - 1}{1-\alpha} \leq \frac{t^{1-\alpha}}{1-\alpha}. \tag{145}$$

With $\varepsilon = \varepsilon_T = ((1-\alpha)/4)T^{-(1-\alpha)}$, for any $2 \leq t \leq T$,

$$\varepsilon \left(1 + \sum_{s=2}^{t} s^{-\alpha}\right) \leq \frac{1}{4}. \tag{146}$$

We now argue by induction. Suppose the second prediction coordinate has been positive through round $t-1$, so that Eq. 144 is valid for the current $d_t$. Suppose, for contradiction, that $Q_{x,2}^t = 0$. Then by Eq. 139 and $\|Q_x^t\|_2 = S_t$, we must have $Q_{x,1}^t = S_t$. Since the first coordinate is positive,

$$\gamma_t = R_{x,1}^t + 1 - S_t.$$

whereas $Q^t_{x,2} = 0$ requires

$$R^t_{x,2} + 1 - \varepsilon - \gamma_t \le 0,$$

i.e., $D_t + \varepsilon \ge S_t$. Dividing by $g_t$ gives

$$d_t + \frac{\varepsilon}{g_t} \ge \sigma_\varepsilon. \tag{147}$$

But by the induction formula and Eq. 146, the left-hand side is at most $1/4$; by Eq. 138, the right-hand side is at least $1/2$, a contradiction. Therefore

$$Q^t_{x,2} > 0, \qquad 2 \le t \le T. \tag{148}$$

This also proves in reverse that Eq. 144 is valid throughout the entire interval $[2, T]$.

**Step 6: The suboptimal action still has constant probability at round $T$.** By Eq. 141–Eq. 146, for $2 \le t \le T$,

$$Q^t_{x,1} - Q^t_{x,2} \le \frac{1}{4} g_t.$$

On the other hand, Eq. 138 gives $S_t = g_t \sigma_\varepsilon \ge \frac{1}{2} g_t$. Hence

$$Q^t_{x,1} - Q^t_{x,2} \le \frac{1}{2} S_t. \tag{149}$$

Also, since $Q^t_{x,3} = 0$ and $Q^t_x \ge 0$,

$$S_t = \|Q^t_x\|_2 \le Q^t_{x,1} + Q^t_{x,2}.$$

Therefore, Eq. 149 implies

$$Q^t_{x,1} - Q^t_{x,2} \le \frac{1}{2}(Q^t_{x,1} + Q^t_{x,2}),$$

i.e., $Q^t_{x,1} \le 3 Q^t_{x,2}$. Since the actual strategy is $x^t = \mathcal{N}(Q^t_x)$,

$$x^t_2 = \frac{Q^t_{x,2}}{Q^t_{x,1} + Q^t_{x,2}} \ge \frac{1}{4}, \qquad 2 \le t \le T. \tag{150}$$

**Step 7: Compute the true duality gap at round $T$.** The two columns of $A_\varepsilon$ are identical, so for any $x$, both column actions are best responses; while the row player's best pure action is the first row, with payoff 1. Using $x^T_3 = 0$ and $x^T_1 + x^T_2 = 1$,

$$\begin{aligned} \mathrm{Gap}_{A_\varepsilon}(z^T) &= 1 - \langle v_\varepsilon, x^T \rangle \\ &= 1 - \left[ x^T_1 + (1-\varepsilon) x^T_2 \right] \\ &= \varepsilon x^T_2. \end{aligned} \tag{151}$$

By Eq. 150,

$$\mathrm{Gap}_{A_T}(z^T_{A_T}) \ge \frac{\varepsilon_T}{4} = \frac{1-\alpha}{16} T^{-(1-\alpha)}.$$

This proves Eq. 12, and therefore

$$\boxed{\mathfrak{G}^{3,2}_T(1) \ge \frac{1-\alpha}{16} T^{-(1-\alpha)}.} \tag{152}$$

. □

*Proof of Theorem 6.2.* Without loss of generality, suppose the row player never activates. By the branch structure in Theorem D.3, $x^t = a_x$ for all $t$, and the row player's current best-response gap is identically zero. The column player receives the fixed maximizing utility

$$v = -A^\top a_x \in \mathbb{R}^{d_{\mathrm{act}}}.$$

Since $a_x \in \Delta_m$ and $|A_{jk}| \le B$, coordinatewise we have

$$|v_j| \le B. \tag{153}$$

If $v$ were constant, then by Lemma 3.2 the column player would also never activate, contradicting "exactly one player never activates." Hence $v$ is nonconstant. Therefore the first-round zero-branch correction is already nonzero:

$$E^1 = v - \langle v, a_y \rangle \mathbf{1}, \qquad H^1 = [E^1]_+ \neq 0,$$

so the column player enters the positive branch at round 2. Below we omit the player subscript and write $S^t = \|R^t\|_2, \quad z^t = z_y^t$. The total game gap is exactly the optimality gap for the fixed linear utility,

$$\mathrm{Gap}(z^t) = g_v(z^t) = \max_j v_j - \langle v, z^t \rangle. \tag{154}$$

**Step 1: After activation, the normalized radius is constant and admits a uniform upper bound.** For all $t \geq 2$, both the previous-round utility and the current utility equal $v$, so

$$m^t = v, \qquad e^t = u^t - m^t = 0, \qquad E^t = 0.$$

Thus $H^t = Q^t, \quad R^{t+1} = \lambda_t Q^t$. By the equal-$\ell_2$-norm condition $\|Q^t\|_2 = S^t$, we obtain $S^{t+1} = \lambda_t S^t$. Since $g_{t+1} = \lambda_t g_t$,

$$\frac{S^{t+1}}{g_{t+1}} = \frac{\lambda_t S^t}{\lambda_t g_t} \equiv \sigma = \frac{S^2}{g_2}, \qquad t \geq 2. \tag{155}$$

Moreover, because $R^2 = \lambda_1 H^1$, $g_1 = 1$, and $g_2 = \lambda_1$,

$$\sigma = \frac{\|R^2\|_2}{g_2} = \|H^1\|_2.$$

By Eq. 153, $|\langle v, a_y \rangle| \leq B$, hence

$$|E_j^1| = |v_j - \langle v, a_y \rangle| \leq 2B.$$

Therefore

$$\sigma = \|H^1\|_2 \leq \|E^1\|_2 \leq 2B\sqrt{d_{\mathrm{act}}}. \tag{156}$$

In particular,

$$S^{T+1} = g_{T+1}\sigma \leq 2B\sqrt{d_{\mathrm{act}}}\,(T+1)^{\alpha}. \tag{157}$$

**Step 2: In the fixed-utility branch, the actual utility is monotone increasing.** We prove

$$\langle v, z^{t+1} \rangle \geq \langle v, z^t \rangle, \qquad t \geq 2, \tag{158}$$

equivalently, $g_v(z^{t+1}) \leq g_v(z^t)$.

Let $h(p) = \|p\|_2$ for $p \in \Delta_{d_{\mathrm{act}}}$. Since $R^t, Q^t \neq 0$ and $\widetilde{z}^t = \mathcal{N}(R^t)$, $z^t = \mathcal{N}(Q^t)$, we have the identity

$$R^t = S^t \nabla h(\widetilde{z}^t), \qquad Q^t = S^t \nabla h(z^t). \tag{159}$$

Moreover, $E^t = 0$ gives $R^{t+1} = \lambda_t Q^t$, and positive scaling does not change normalization, so

$$\widetilde{z}^{t+1} = z^t. \tag{160}$$

Consider the equal-$\ell_2$-norm prediction at round $t+1$:

$$Q^{t+1} = [R^{t+1} + v - \gamma_{t+1}\mathbf{1}]_+, \qquad \|Q^{t+1}\|_2 = \|R^{t+1}\|_2 = S^{t+1}.$$

Define

$$c^{t+1} = R^{t+1} + v - Q^{t+1}.$$

If $Q_j^{t+1} > 0$, then $c_j^{t+1} = \gamma_{t+1}$; if $Q_j^{t+1} = 0$, then $c_j^{t+1} \leq \gamma_{t+1}$, so for any $p \in \Delta_{d_{\mathrm{act}}}$, $\langle c^{t+1}, p \rangle \leq \gamma_{t+1}$. Since the support of $z^{t+1} = \mathcal{N}(Q^{t+1})$ is exactly the set of positive coordinates of $Q^{t+1}$, we have $\langle c^{t+1}, z^{t+1} \rangle = \gamma_{t+1}$. Combining these facts, for any $p \in \Delta_{d_{\mathrm{act}}}$,

$$\langle c^{t+1}, p - z^{t+1} \rangle \leq 0. \tag{161}$$

On the other hand, using Eq. 159–Eq. 160,

$$c^{t+1} = v - S^{t+1}\big(\nabla h(z^{t+1}) - \nabla h(z^t)\big).$$

taking $p = z^t$ in Eq. 161 gives

$$\langle v, z^{t+1} - z^t \rangle \geq S^{t+1} \left\langle \nabla h(z^{t+1}) - \nabla h(z^t), z^{t+1} - z^t \right\rangle \geq 0,$$

where the final step uses monotonicity of the gradient of the convex function $h$. Hence Eq. 158 holds, and therefore

$$g_v(z^2) \geq g_v(z^3) \geq \cdots \geq g_v(z^T). \tag{162}$$

**Step 3: Cumulative-regret domination directly yields a last-iterate bound.** Take any

$$j^\star \in \arg\max_j v_j.$$

Under fixed utility, the corresponding coordinate of the true instantaneous regret is exactly

$$r_{j^\star}^t = v_{j^\star} - \langle v, z^t \rangle = g_v(z^t) \geq 0. \tag{163}$$

By the coordinatewise cumulative-regret domination in Theorem B.1,

$$R_{j^\star}^{T+1} \geq \sum_{t=1}^{T} r_{j^\star}^t \geq \sum_{t=2}^{T} g_v(z^t).$$

Using Eq. 162,

$$R_{j^\star}^{T+1} \geq (T-1) g_v(z^T). \tag{164}$$

On the other hand, any coordinate of a nonnegative state is at most its Euclidean norm, so

$$R_{j^\star}^{T+1} \leq S^{T+1}.$$

Combining Eq. 157 and Eq. 164, we obtain

$$(T-1) g_v(z^T) \leq 2B\sqrt{d_{\text{act}}}\,(T+1)^\alpha.$$

Using Eq. 154 then gives the claim. Eq. 13. The case in which the column player never activates while the row player does is completely symmetric. □

**Example E.1** (A matrix game with exactly one player never activating)**.** *Consider the zero-sum matrix game* $A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$. *For any* $y \in \Delta_2$, $Ay = y_1 \mathbf{1}$, *so the row player's centered correction is always zero. Hence, starting from* $R_x^1 = 0$, *we have* $R_x^t = 0, \qquad x^t = a_x, \qquad \forall t$, *and therefore* $\tau_x = \infty$.

*On the other hand, for any full-support anchor* $a_x$, $-A^\top a_x = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$, *which is nonconstant. Thus the column player activates (in fact, at the first update), so* $\tau_y < \infty$. *Hence this game realizes the one-zero branch of Theorem 8.3:* $\tau_x = \infty, \qquad \tau_y < \infty$.

*Moreover, its Nash equilibrium set is* $Z^\star = \Delta_2 \times \{e_2\}$.

# F Regularized Radial IREG-PRM$^+$ under Bandit Feedback: Last-Iterate Convergence in Expectation

**Lemma F.1** (Pathwise well-definedness of the algorithm and validity of conditional expectations)**.** *For every sample path and every* $t \geq 1$, *the quantities* $S_i^t, Q_i^t, z_i^t, H_i^t, R_i^{t+1}$ *in Algorithm 2 are all well defined and*

$$S_i^t > 0, \qquad Q_i^t \neq 0, \qquad H_i^t \neq 0.$$

*Moreover,* $\gamma_i^t, Q_i^t, z_i^t, p_i^t$ *are* $\mathcal{F}_{t-1}$*-measurable. Therefore all subsequent* $\mathbb{E}_t[\cdot]$ *conditioned on* $z^t, p^t$ *are well defined.*

*Proof.* The initialization satisfies $S_i^1 = \|a_i/\|a_i\|_2\|_2 = 1 > 0$. Suppose $S_i^t > 0$ at some round. For given $(R_i^t, m_i^t)$, define

$$f_t(\gamma) = \left\| [R_i^t + m_i^t - \gamma \mathbf{1}]_+ \right\|_2.$$

$f_t$ is continuous and strictly decreasing on the region where $f_t > 0$; as $\gamma \to -\infty$, $f_t(\gamma) \to \infty$, while if $\gamma \geq \max_j (R_{i,j}^t + m_{i,j}^t)$, then $f_t(\gamma) = 0$. Hence the equation $f_t(\gamma) = S_i^t > 0$ has a unique solution $\gamma_i^t$. By definition, $\|Q_i^t\|_2 = S_i^t > 0$, so $Q_i^t \neq 0$ and $z_i^t = \mathcal{N}(Q_i^t)$ is well defined.

Next let $E_i^t = e_i^t - \langle e_i^t, z_i^t \rangle \mathbf{1}$. Since $Q_i^t = \|Q_i^t\|_1 z_i^t$,

$$\langle Q_i^t, E_i^t \rangle = 0.$$

coordinatewise, $[Q_i^t + E_i^t]_+ - Q_i^t \geq E_i^t$, hence

$$\langle Q_i^t, H_i^t - Q_i^t \rangle \geq 0$$

and

$$\|H_i^t\|_2^2 = \|Q_i^t\|_2^2 + 2\langle Q_i^t, H_i^t - Q_i^t \rangle + \|H_i^t - Q_i^t\|_2^2 \geq (S_i^t)^2 > 0.$$

Thus $H_i^t \neq 0$, and therefore $R_i^{t+1} = \lambda_t H_i^t \neq 0$. Induction proves that all rounds are well defined.

Finally, $R_i^t, m_i^t$ are finite-dimensional $\mathcal{F}_{t-1}$-measurable random vectors. The unique root can be written as

$$\gamma_i^t = \inf\{q \in \mathbb{Q} : f_t(q) \leq S_i^t\},$$

where the right-hand side is an infimum over countably many $\mathcal{F}_{t-1}$-measurable events. Hence $\gamma_i^t$ is measurable; subsequently, $Q_i^t, z_i^t, p_i^t$ are obtained through continuous/coordinatewise continuous maps and thus are also $\mathcal{F}_{t-1}$-measurable. □

### F.1 Second Moment and Pathwise Bound of the IX Estimator

**Lemma F.2** (Conditional mean, bias, and second moment of the single-sample IX estimator)**.** *For any $t \geq 1$, conditioned on $\mathcal{F}_{t-1}$, we have*

$$\mathbb{E}_t[\widehat{u}_{x,j}^t] = \frac{p_{x,j}^t}{p_{x,j}^t + \theta_t}(Ap_y^t)_j, \qquad \mathbb{E}_t[\widehat{u}_{y,k}^t] = -\frac{p_{y,k}^t}{p_{y,k}^t + \theta_t}(A^\top p_x^t)_k. \tag{165}$$

*Define $b_i^t = \mathbb{E}_t[\widehat{u}_i^t] - U_i(z^t), \qquad b^t = (b_x^t, b_y^t)$. Then*

$$\boxed{\|b^t\|_2 \leq B_t = \sqrt{2}L\left(\sqrt{2}\,\epsilon_t + \frac{\theta_t}{\alpha_a \epsilon_t}\right) \leq C_b t^{-1/5+2\varepsilon},} \tag{166}$$

*where $C_b = \sqrt{2}L(\sqrt{2} + \alpha_a^{-1})$. Moreover, for $i \in \{x, y\}$,*

$$\mathbb{E}_t\|\widehat{u}_i^t\|_2^2 \leq \frac{B^2 d_i}{\alpha_a \epsilon_t}, \qquad \|\widehat{u}_i^t\|_2 \leq \frac{B}{\alpha_a \epsilon_t} \qquad \textit{pathwise}. \tag{167}$$

*Proof.* For the row player, the $j$th coordinate is nonzero only when $I_t = j$. Therefore, using conditional independence,

$$\mathbb{E}_t[\widehat{u}_{x,j}^t] = \sum_{k=1}^n p_{x,j}^t p_{y,k}^t \frac{A_{jk}}{p_{x,j}^t + \theta_t} = \frac{p_{x,j}^t}{p_{x,j}^t + \theta_t}(Ap_y^t)_j.$$

The column-player case gives Eq. 165 in the same way. Let $D_x^t = \mathrm{diag}(p_{x,j}^t/(p_{x,j}^t + \theta_t))$. Then $\mathbb{E}_t \widehat{u}_x^t = D_x^t A p_y^t$, Hence

$$\begin{aligned}
\|b_x^t\|_2 &= \|D_x^t A p_y^t - A z_y^t\|_2 \\
&\leq \|A(p_y^t - z_y^t)\|_2 + \|(I - D_x^t)Ap_y^t\|_2 \\
&\leq L\epsilon_t \|a_y - z_y^t\|_2 + \max_j \frac{\theta_t}{p_{x,j}^t + \theta_t}\,\|Ap_y^t\|_2 \\
&\leq \sqrt{2}L\epsilon_t + L\frac{\theta_t}{\alpha_a \epsilon_t}.
\end{aligned}$$

Here we use $p_{x,j}^t \geq \epsilon_t a_{x,j} \geq \alpha_a \epsilon_t$, $\|p_y^t\|_2 \leq 1$, and the fact that the simplex diameter is at most $\sqrt{2}$. The column-player case is identical. Therefore

$$\|b^t\|_2 \leq \sqrt{\|b_x^t\|_2^2 + \|b_y^t\|_2^2} \leq \sqrt{2}L\left(\sqrt{2}\epsilon_t + \frac{\theta_t}{\alpha_a \epsilon_t}\right).$$

By Eq. 16, $\epsilon_t = t^{-1/5+2\varepsilon}$ and $\theta_t/\epsilon_t = t^{-1/5+2\varepsilon}$, which yields Eq. 166.

For the row player, moreover,

$$\begin{aligned}\mathbb{E}_t\|\widehat{u}_x^t\|_2^2 &= \sum_{j=1}^{m}\sum_{k=1}^{n} p_{x,j}^t p_{y,k}^t \frac{A_{jk}^2}{(p_{x,j}^t+\theta_t)^2} \\ &\le B^2 \sum_{j=1}^{m} \frac{p_{x,j}^t}{(p_{x,j}^t+\theta_t)^2} \\ &\le B^2 \sum_{j=1}^{m} \frac{1}{p_{x,j}^t+\theta_t} \le \frac{B^2 m}{\alpha_a \epsilon_t}.\end{aligned}$$

The column-player case is analogous. Finally, since $\widehat{u}_i^t$ has at most one nonzero coordinate in each round, $|r_t| \le B$, and $p_{i,j}^t + \theta_t \ge p_{i,j}^t \ge \alpha_a \epsilon_t$, we directly obtain $\|\widehat{u}_i^t\|_2 \le B/(\alpha_a \epsilon_t)$. □

**Lemma F.3** (Uniform moment bounds for the regularized estimator and correction)**.** *There exist finite constants $K_u, K_E, K_\infty$ depending only on $(A, D, \alpha_a)$ such that for all $t \ge 1$ and $i \in \{x, y\}$,*

$$\mathbb{E}\|\widetilde{u}_i^t\|_2^2 \le \frac{K_u}{\epsilon_t}, \qquad \mathbb{E}\|E_i^t\|_2^2 \le \frac{K_E}{\epsilon_t}, \qquad \textit{and} \qquad \|E_i^t\|_2 \le \frac{K_\infty}{\epsilon_t} \quad \textit{pathwise}. \tag{168}$$

*For example, one may take*

$$K_u = \frac{2B^2 D}{\alpha_a} + 4, \qquad K_E = 4C_D^2 K_u, \qquad K_\infty = 2C_D\left(\frac{B}{\alpha_a} + \sqrt{2}\right).$$

*Proof.* Using $\|z_i^t - a_i\|_2 \le \sqrt{2}$, $\tau_t \le 1$, and $(a+b)^2 \le 2a^2 + 2b^2$, Lemma F.2 gives

$$\mathbb{E}\|\widetilde{u}_i^t\|_2^2 \le 2\mathbb{E}\|\widehat{u}_i^t\|_2^2 + 4\tau_t^2 \le \frac{2B^2 D/\alpha_a + 4}{\epsilon_t},$$

which is Eq. 168. For $t \ge 2$, $m_i^t = \widetilde{u}_i^{t-1}$, and $\epsilon_t$ is decreasing, so

$$\mathbb{E}\|e_i^t\|_2^2 \le 2\mathbb{E}\|\widetilde{u}_i^t\|_2^2 + 2\mathbb{E}\|\widetilde{u}_i^{t-1}\|_2^2 \le \frac{4K_u}{\epsilon_t}.$$

When $t = 1$, $m_i^1 = 0$, and the same upper bound still holds. The centering map $\mathcal{C}_z(v) = v - \langle v, z\rangle \mathbf{1}$ satisfies $\|\mathcal{C}_z(v)\|_2 \le C_D\|v\|_2$. Hence $\mathbb{E}\|E_i^t\|_2^2 = \mathbb{E}\|\mathcal{C}_{z_i^t}(e_i^t)\|_2^2 \le C_D^2 \mathbb{E}\|e_i^t\|_2^2$, so Eq. 168 follows.

Pathwise, by Eq. 167, $\tau_t \le 1$, and $\epsilon_t \le 1$,

$$\|\widetilde{u}_i^t\|_2 \le \frac{B}{\alpha_a \epsilon_t} + \sqrt{2} \le \frac{B/\alpha_a + \sqrt{2}}{\epsilon_t}.$$

Since $\epsilon_{t-1}^{-1} \le \epsilon_t^{-1}$, $\|e_i^t\|_2 \le 2(B/\alpha_a + \sqrt{2})/\epsilon_t$; applying the centering bound again yields Eq. 168. □

### F.2 Deterministic Upper and Lower Bounds on the Normalized Radius and Prediction Movement

**Lemma F.4** (Pathwise boundedness of the normalized radius)**.** *For every sample path and every $t \ge 1$, Algorithm 2 satisfies*

$$\boxed{1 \le \sigma_i^t \le \overline{\sigma} < \infty, \qquad i \in \{x, y\},} \tag{169}$$

*where one may take*

$$\overline{\sigma} = \left(1 + K_\infty^2 \sum_{s=1}^{\infty} s^{-1-4\varepsilon}\right)^{1/2}. \tag{170}$$

*Therefore $S_i^t \ge g_t$, and the algorithm remains in the positive branch forever, requiring no restart or activation branch.*

*Proof.* The initialization gives $\sigma_i^1 = \|R_i^1\|_2/g_1 = 1$. Whenever $S_i^t > 0$, Lemma A.2 applies verbatim to the present $(Q_i^t, E_i^t, H_i^t)$, so

$$\|H_i^t\|_2 \geq S_i^t, \qquad \|H_i^t\|_2^2 \leq (S_i^t)^2 + \|E_i^t\|_2^2.$$

Then, using $R_i^{t+1} = \lambda_t H_i^t$ and $\lambda_t = g_{t+1}/g_t$,we have

$$(\sigma_i^{t+1})^2 = \frac{\|H_i^t\|_2^2}{g_t^2} \leq (\sigma_i^t)^2 + \frac{\|E_i^t\|_2^2}{g_t^2}, \tag{171}$$

$$\sigma_i^{t+1} = \frac{\|H_i^t\|_2}{g_t} \geq \frac{S_i^t}{g_t} = \sigma_i^t. \tag{172}$$

we see that $S_i^t$ is always positive and $\sigma_i^t \geq 1$. Using the pathwise bound in Lemma F.3 and Eq. 16,

$$\frac{\|E_i^t\|_2^2}{g_t^2} \leq \frac{K_\infty^2}{\epsilon_t^2 g_t^2} = K_\infty^2 t^{-1-4\varepsilon}.$$

Summing Eq. 171 from 1 to $t-1$, and using $\sum_{s\geq 1} s^{-1-4\varepsilon} < \infty$ for every fixed $\varepsilon > 0$, gives Eq. 169–Eq. 170. □

**Lemma F.5** (Second moment of the prediction movement)**.** *Define*

$$a_t^2 = \|z_x^t - \widetilde{z}_x^t\|_2^2 + \|z_y^t - \widetilde{z}_y^t\|_2^2. \tag{173}$$

*Then there exists a finite constant $K_a$ such that*

$$\boxed{\mathbb{E}[a_t^2] \leq \frac{K_a}{\epsilon_t g_t^2} = K_a t^{-6/5-2\varepsilon}.} \tag{174}$$

*Proof.* Fix a block and temporarily write $R, m, Q, S, \gamma, z, \widetilde{z}$. Let $f(\gamma) = \|[R + m - \gamma\mathbf{1}]_+\|_2$. At $\gamma = \|m\|_\infty$, $[R + m - \gamma\mathbf{1}]_+ \leq R$ coordinatewise, so $f(\gamma) \leq S$; at $\gamma = -\|m\|_\infty$, $R + m - \gamma\mathbf{1} \geq R \geq 0$, so $f(\gamma) \geq S$. By uniqueness of the equal-$\ell_2$-norm root,

$$|\gamma| \leq \|m\|_\infty \leq \|m\|_2. \tag{175}$$

The positive-part projection is Euclidean nonexpansive, hence

$$\|Q - R\|_2 \leq \|m - \gamma\mathbf{1}\|_2 \leq (1 + \sqrt{d_i})\|m\|_2 \leq C_D\|m\|_2. \tag{176}$$

At the same time, applying the bound in Lemma A.2, $\|\mathcal{N}(Q) - \mathcal{N}(R)\|_2 \leq C_D \frac{\|Q-R\|_2}{\|R\|_2}$. and combining Eq. 176 with $S_i^t \geq g_t$, we obtain

$$\|z_i^t - \widetilde{z}_i^t\|_2^2 = \|\mathcal{N}(Q_i^t) - \mathcal{N}(R_i^t)\|_2^2 \leq \frac{C_D^4\|m_i^t\|_2^2}{g_t^2}. \tag{177}$$

For the upper bound of $\|m_i^t\|_2^2$, notice that when $t = 1$, $m_i^1 = 0$; when $t \geq 2$, $m_i^t = \widetilde{u}_i^{t-1}$. then using Lemma F.3,

$$\mathbb{E}\|m_i^t\|_2^2 = \mathbb{E}\|\widetilde{u}_i^{t-1}\|_2^2 \leq \frac{K_u}{\epsilon_{t-1}} \leq \frac{K_u}{\epsilon_t}. \tag{178}$$

Combining Eq. 177-Eq. 178, and summing over player indices $i \in \{x, y\}$ obtains Eq. 174. □

### F.3 Strong Restoring Force and Duality Gap of the Regularized Equilibrium

$$\langle F_\tau(z_\tau^*), z - z_\tau^* \rangle \leq 0, \qquad \forall z \in Z. \tag{179}$$

**Lemma F.6** (Regularized equilibrium: existence, uniqueness, and strong restoring inequality)**.** *For every $\tau > 0$, the VI Eq. 179 has a unique solution $z_\tau^*$. Moreover, for all $z \in Z$,*

$$\boxed{\langle F_\tau(z), z - z_\tau^* \rangle \leq -\tau\|z - z_\tau^*\|_2^2.} \tag{180}$$

*Proof.* We first give an existence proof that does not rely on an external VI theorem. Fix any $\eta > 0$ and define the continuous map

$$T_\eta(z) = \Pi_Z\big(z + \eta F_\tau(z)\big),$$

where $\Pi_Z$ denotes Euclidean projection. Since $Z$ is a nonempty compact convex set and $T_\eta : Z \to Z$ is continuous, Brouwer's fixed-point theorem gives some $z^* \in Z$ such that $z^* = T_\eta(z^*)$.Using the first-order optimality condition of Euclidean projection, we have

$$\langle z^* - (z^* + \eta F_\tau(z^*)), z - z^* \rangle \geq 0, \quad \iff \quad \langle F_\tau(z^*), z - z^* \rangle \leq 0, \qquad \forall z \in Z.$$

Thus the VI has at least one solution. For any $z = (x, y), z' = (x', y') \in Z$, the bilinear part satisfies the exact skew identity

$$\langle F(z) - F(z'), z - z' \rangle = \langle A(y - y'), x - x' \rangle + \langle -A^\top (x - x'), y - y' \rangle = 0.$$

Therefore

$$\langle F_\tau(z) - F_\tau(z'), z - z' \rangle = -\tau \|z - z'\|_2^2. \tag{181}$$

If both $z, z'$ are VI solutions, then

$$\langle F_\tau(z), z' - z \rangle \leq 0, \qquad \langle F_\tau(z'), z - z' \rangle \leq 0.$$

Adding the two inequalities gives

$$\langle F_\tau(z) - F_\tau(z'), z - z' \rangle \geq 0.$$

Combining this with Eq. 181 yields $-\tau \|z - z'\|_2^2 \geq 0$, so $z = z'$, proving uniqueness.

Finally, fix any $z \in Z$. In Eq. 181, set $z' = z_\tau^*$ to obtain

$$\langle F_\tau(z), z - z_\tau^* \rangle = \langle F_\tau(z_\tau^*), z - z_\tau^* \rangle - \tau \|z - z_\tau^*\|_2^2.$$

VI optimality makes the first term at most 0, hence $\langle F_\tau(z), z - z_\tau^* \rangle \leq -\tau \|z - z_\tau^*\|_2^2$, which completes the proof of Eq. 180. □

**Lemma F.7** (Sensitivity of the regularized equilibrium with respect to $\tau$)**.** *Let $D_a = \sup_{z \in Z} \|z - a\|_2 \leq 2$. Then for any $\tau, \tau' > 0$,*

$$\boxed{\|z_\tau^* - z_{\tau'}^*\|_2 \leq D_a \frac{|\tau - \tau'|}{\min\{\tau, \tau'\}}.} \tag{182}$$

*In particular, for $\tau_t = t^{-1/10+\varepsilon}$,*

$$\|z_{\tau_{t+1}}^* - z_{\tau_t}^*\|_2 \leq \frac{(1/10 - \varepsilon) D_a}{t}. \tag{183}$$

*Proof.* Let $z = z_\tau^*$, $z' = z_{\tau'}^*$, and $d = z - z'$. The two VI optimality conditions respectively give

$$\langle F(z) - \tau(z - a), z' - z \rangle \leq 0, \qquad \langle F(z') - \tau'(z' - a), z - z' \rangle \leq 0.$$

equivalently,

$$\langle F(z) - \tau(z - a) - F(z') + \tau'(z' - a), d \rangle \geq 0.$$

The bilinear skew term has zero inner product with $d$, hence

$$-\tau \|d\|_2^2 + (\tau' - \tau) \langle z' - a, d \rangle \geq 0.$$

and therefore

$$\tau \|d\|_2^2 \leq |\tau' - \tau| D_a \|d\|_2.$$

If $d \neq 0$, canceling $\|d\|_2$ gives $\|d\|_2 \leq D_a |\tau' - \tau| / \tau$; swapping $\tau, \tau'$ and taking the smaller denominator gives the claim. Eq. 182.

For the schedule used in this section, $\tau_t = t^{-1/10+\varepsilon}$ is decreasing, so

$$\frac{\tau_t - \tau_{t+1}}{\tau_{t+1}} = \left(1 + \frac{1}{t}\right)^{1/10-\varepsilon} - 1 \leq \frac{1/10 - \varepsilon}{t},$$

where the inequality uses Bernoulli's inequality $(1 + x)^a \leq 1 + ax$ for $a \in [0, 1]$. Substituting into Eq. 182 yields Eq. 183. □

**Lemma F.8** (Gap bias of the regularized equilibrium relative to the original zero-sum game)**.** *For any $\tau > 0$,*

$$\boxed{\mathrm{Gap}(z_\tau^*) \le 4\tau.} \tag{184}$$

*Moreover, for any $z, z' \in Z$,*

$$|\,\mathrm{Gap}(z) - \mathrm{Gap}(z')| \le \sqrt{2}L\|z - z'\|_2. \tag{185}$$

*Proof.* Write $z_\tau^* = (x_\tau^*, y_\tau^*)$. In the VI Eq. 179, first take the test point $z = (x, y_\tau^*)$. Then, for all $x \in \Delta_m$,

$$\langle Ay_\tau^* - \tau(x_\tau^* - a_x), x - x_\tau^*\rangle \le 0.$$

we obtain

$$\max_{x\in\Delta_m} \langle Ay_\tau^*, x - x_\tau^*\rangle \le \tau\|x_\tau^* - a_x\|_2 \max_{x\in\Delta_m} \|x - x_\tau^*\|_2 \le 2\tau,$$

because the Euclidean diameter of the simplex is at most $\sqrt{2}$. Next take the test point $z = (x_\tau^*, y)$. We obtain

$$\langle -A^\top x_\tau^* - \tau(y_\tau^* - a_y), y - y_\tau^*\rangle \le 0,$$

Thus

$$\max_{y\in\Delta_n} \langle -A^\top x_\tau^*, y - y_\tau^*\rangle \le 2\tau.$$

while

$$\max_x \langle Ay_\tau^*, x - x_\tau^*\rangle + \max_y \langle -A^\top x_\tau^*, y - y_\tau^*\rangle = \max_x x^\top Ay_\tau^* - \min_y (x_\tau^*)^\top Ay = \mathrm{Gap}(z_\tau^*).$$

Hence Eq. 184 follows. Now let $f(y) = \max_{p\in\Delta_m} p^\top Ay$, $h(x) = \min_{q\in\Delta_n} x^\top Aq$. The Euclidean norms of all $p, q$ are at most 1, so

$$|f(y) - f(y')| \le L\|y - y'\|_2, \qquad |h(x) - h(x')| \le L\|x - x'\|_2.$$

Therefore

$$|\,\mathrm{Gap}(z) - \mathrm{Gap}(z')| \le L(\|x - x'\|_2 + \|y - y'\|_2) \le \sqrt{2}L\|z - z'\|_2. \qquad \square$$

### F.4 Moving-Comparator Perspective Potential

Define Lyapunov potential function as

$$\boxed{\Phi_t = \Psi_t(z_{\tau_t}^*).} \tag{186}$$

**Lemma F.9** (Regularized single-sample perspective one-step inequality)**.** *For any $\mathcal{F}_{t-1}$-measurable comparator $w \in Z$, pathwise we have*

$$\Psi_{t+1}(w) \le \Psi_t(w) + \frac{1}{g_t}\sum_i \langle \widetilde{u}_i^t, z_i^t - w_i\rangle - \frac{\delta_x^t + \delta_y^t}{g_t} + \frac{1}{2}\mathcal{V}_t. \tag{187}$$

*and there exists a constant $K_V$ such that*

$$\boxed{\mathbb{E}[\mathcal{V}_t] \le \frac{K_V}{\epsilon_t g_t^2} = K_V t^{-6/5-2\varepsilon}.} \tag{188}$$

*Proof.* Applying Lemma A.4 to the bandit feedback setting obtains Eq. 187. Combining $\mathbb{E}\|E_i^t\|_2^2 \le \frac{K_E}{\epsilon_t}$ in Lemma F.3 and $S_i^t \ge g_t$ in Lemma F.4 obtains

$$\mathbb{E}\mathcal{V}_t = \sum_i \mathbb{E}\left[\frac{\|E_i^t\|_2^2}{g_t S_i^t}\right] \le \sum_i \frac{\mathbb{E}\|E_i^t\|_2^2}{g_t^2} \le \frac{2K_E}{\epsilon_t g_t^2}.$$

Taking $K_V = 2K_E$, we obtain Eq. 188. $\square$

### F.5 LYAPUNOV RECURSION

**Lemma F.10** (Expected Lyapunov Function Descent Inequality)**.** *There exist constants $c_{A,\varepsilon} > 0$ (one may take $c_{A,\varepsilon} = \frac{1}{8D\overline{\sigma}}$) and $C_{A,\varepsilon} < \infty$ such that for all $t \geq 1$,*

$$\boxed{\mathbb{E}[\Phi_{t+1}] \leq \left(1 - c_{A,\varepsilon} t^{-4/5+\varepsilon}\right) \mathbb{E}[\Phi_t] + C_{A,\varepsilon} t^{-1+3\varepsilon}.} \tag{189}$$

*Proof.* Fix $t$. Then $\Phi_t, z^t, p^t, \tau_t$ are all $\mathcal{F}_{t-1}$-measurable (see Lemma F.1). In Lemma F.9, take $w = z^*_{\tau_t}$, and using $\delta_i^t \geq 0$,

$$\begin{aligned}\Psi_{t+1}(z^*_{\tau_t}) \leq& \Psi_t(z^*_{\tau_t}) + \frac{1}{g_t}\sum_i \langle \widetilde{u}_i^t, z_i^t - z^*_{\tau_t,i}\rangle - \frac{\delta_x^t + \delta_y^t}{g_t} + \frac{1}{2}\mathcal{V}_t \\ \leq& \Psi_t(z^*_{\tau_t}) + \frac{1}{g_t}\sum_i \langle \widetilde{u}_i^t, z_i^t - z^*_{\tau_t,i}\rangle + \frac{1}{2}\mathcal{V}_t. \end{aligned} \tag{190}$$

Applying Lemma F.2 $\mathbb{E}_t[\widetilde{u}^t] = F_{\tau_t}(z^t) + b^t, \quad \|b^t\|_2 \leq B_t$. and Lemma F.6 with $d_t = \|z^t - z^*_{\tau_t}\|_2$,

$$\begin{aligned}\mathbb{E}_t \sum_i \langle \widetilde{u}_i^t, z_i^t - z^*_{\tau_t,i}\rangle &= \langle F_{\tau_t}(z^t), z^t - z^*_{\tau_t}\rangle + \langle b^t, z^t - z^*_{\tau_t}\rangle \\ &\leq -\tau_t d_t^2 + B_t d_t \leq -\frac{\tau_t}{2} d_t^2 + \frac{B_t^2}{2\tau_t},\end{aligned}$$

where the final step uses Young's inequality $B_t d_t \leq \frac{\tau_t}{2} d_t^2 + \frac{B_t^2}{2\tau_t}$. Thus, taking expectation over $\mathcal{F}_{t-1}$ for Eq. 190,

$$\mathbb{E}_t \Psi_{t+1}(z^*_{\tau_t}) \leq \Phi_t - \frac{\tau_t}{2g_t} d_t^2 + \frac{B_t^2}{2\tau_t g_t} + \frac{1}{2}\mathbb{E}_t \mathcal{V}_t. \tag{191}$$

We next convert $d_t^2$ into the current perspective potential. By $\widetilde{z}^t - z^*_{\tau_t} = (\widetilde{z}^t - z^t) + (z^t - z^*_{\tau_t})$, $\|\widetilde{z}^t - z^*_{\tau_t}\|_2^2 \leq 2a_t^2 + 2d_t^2$, we have

$$d_t^2 \geq \frac{1}{2}\|\widetilde{z}^t - z^*_{\tau_t}\|_2^2 - a_t^2. \tag{192}$$

On the other hand, by the upper bound in Lemma A.5, $\Phi_t \leq 2D\overline{\sigma}\|\widetilde{z}^t - z^*_{\tau_t}\|_2^2$. substituting into Eq. 192 gives

$$d_t^2 \geq \frac{\Phi_t}{4D\overline{\sigma}} - a_t^2. \tag{193}$$

Substituting Eq. 193 into Eq. 191 yields

$$\mathbb{E}_t \Psi_{t+1}(z^*_{\tau_t}) \leq \left(1 - \frac{\tau_t}{8D\overline{\sigma} g_t}\right)\Phi_t + \frac{\tau_t}{2g_t} a_t^2 + \frac{B_t^2}{2\tau_t g_t} + \frac{1}{2}\mathbb{E}_t \mathcal{V}_t. \tag{194}$$

It remains to handle the movement of the comparator from $z^*_{\tau_t}$ to $z^*_{\tau_{t+1}}$. By the Lipschitz bound in Lemma A.5 and Lemma F.7, pathwise we have

$$\Phi_{t+1} = \Psi_{t+1}(z^*_{\tau_{t+1}}) \leq \Psi_{t+1}(z^*_{\tau_t}) + 2\sqrt{2}\overline{\sigma}\|z^*_{\tau_{t+1}} - z^*_{\tau_t}\|_2. \tag{195}$$

Taking expectation over both sides of Eq. 195, and then applying Eq. 194 gives

$$\mathbb{E}\Phi_{t+1} \leq \left(1 - \frac{\tau_t}{8D\overline{\sigma} g_t}\right)\mathbb{E}\Phi_t + \frac{\tau_t}{2g_t}\mathbb{E}a_t^2 + \frac{B_t^2}{2\tau_t g_t} + \frac{1}{2}\mathbb{E}\mathcal{V}_t + 2\sqrt{2}\overline{\sigma}\|z^*_{\tau_{t+1}} - z^*_{\tau_t}\|_2. \tag{196}$$

Using Eq. 166, Eq. 174, Eq. 188, the order of each term in Eq. 196 is

$$\begin{aligned}\frac{\tau_t}{g_t} &= t^{-1/10+\varepsilon-7/10} = t^{-4/5+\varepsilon}, \\ \frac{\tau_t}{g_t}\mathbb{E}[a_t^2] &\leq K_a t^{-4/5+\varepsilon} t^{-6/5-2\varepsilon} = K_a t^{-2-\varepsilon}, \\ \frac{B_t^2}{\tau_t g_t} &\leq C_b^2 t^{-2/5+4\varepsilon}\, t^{1/10-\varepsilon}\, t^{-7/10} = C_b^2 t^{-1+3\varepsilon}, \\ \mathbb{E}\mathcal{V}_t &\leq K_V t^{-6/5-2\varepsilon}, \\ \|z^*_{\tau_{t+1}} - z^*_{\tau_t}\|_2 &= O(t^{-1}).\end{aligned}$$

Since $0 < \varepsilon < 1/10$, $\frac{6}{5} + 2\varepsilon > 1 - 3\varepsilon$, $\quad 2 + \varepsilon > 1 - 3\varepsilon$, $\quad 1 > 1 - 3\varepsilon$, all faster-decaying terms can be absorbed into a single $C_{A,\varepsilon} t^{-1+3\varepsilon}$ term. The constants may depend on $\varepsilon$ through $\overline{\sigma}$ in Eq. 170. Taking $c_{A,\varepsilon} = (8D\overline{\sigma})^{-1}$ gives Eq. 189. □

**Lemma F.11.** *Suppose a nonnegative sequence $x_t$ satisfies*

$$x_{t+1} \le (1 - ct^{-\alpha})x_t + Ct^{-\beta}, \qquad t \ge 1, \tag{197}$$

*where $c, C > 0$ and $0 < \alpha < \beta < 1$. Let $q = \beta - \alpha > 0$. Then there exist a finite constant $K$ and an integer $T_0$ such that*

$$x_t \le Kt^{-q}, \qquad t \ge T_0. \tag{198}$$

*Proof.* Since $\alpha > 0$ and $\alpha - 1 < 0$, we can take $T_0$ sufficiently large so that

$$ct^{-\alpha} \le \frac{1}{2} \quad \text{and} \quad qt^{\alpha-1} \le \frac{c}{2}, \qquad \forall t \ge T_0. \tag{199}$$

Then choose

$$K \ge \max\left\{2C/c,\ x_{T_0} T_0^q\right\}.$$

We next prove $x_t \le Kt^{-q}, \forall t \ge T_0$ by induction.The claim holds evidently in the basic case $t = T_0$ since $K \ge x_{T_0} T_0^q$. If $x_t \le Kt^{-q}$, consider the inductive case $t + 1$. The definition of $K$ gives $Kc/2 > C$, then

$$\begin{aligned} x_{t+1} &\le (1 - ct^{-\alpha})Kt^{-q} + Ct^{-\beta} \\ &= Kt^{-q} - (cK - C)t^{-\beta} \\ &\le Kt^{-q} - \frac{cK}{2} t^{-\beta}. \end{aligned}$$

On the other hand, the mean value theorem for $s \mapsto s^{-q}$ gives

$$t^{-q} - (t+1)^{-q} \le qt^{-q-1} = qt^{\alpha-1}t^{-\alpha-q} \le \frac{c}{2} t^{-\beta}.$$

where the last inequality uses Eq. 199 for $t \ge T_0$, Therefore

$$x_{t+1} \le Kt^{-q} - K\big(t^{-q} - (t+1)^{-q}\big) = K(t+1)^{-q}. \qquad \square$$

*Proof for Theorem 7.1.* Fix $\varepsilon \in (0, 1/10)$. In Lemma F.10, set

$$x_t = \mathbb{E}\Phi_t, \qquad \alpha = \frac{4}{5} - \varepsilon, \qquad \beta = 1 - 3\varepsilon.$$

Then $0 < \alpha < \beta < 1$ and

$$q = \beta - \alpha = \frac{1}{5} - 2\varepsilon > 0.$$

Applying Lemma F.11, there exist finite $C_{A,\varepsilon}, T_{A,\varepsilon}$ such that

$$\mathbb{E}\Phi_T \le C_{A,\varepsilon} T^{-1/5+2\varepsilon},$$

By the lower bound in Lemma A.5, pathwise we have $c_D \|\widetilde{z}^T - z^*_{\tau_T}\|_2^2 \le \Phi_T$. Hence Jensen/Cauchy–Schwarz gives

$$\mathbb{E}\|\widetilde{z}^T - z^*_{\tau_T}\|_2 \le \sqrt{\mathbb{E}\|\widetilde{z}^T - z^*_{\tau_T}\|_2^2} \le c_D^{-1/2} \sqrt{\mathbb{E}\Phi_T} \le C_{A,\varepsilon} T^{-1/10+\varepsilon}.$$

On the other hand, Lemma F.5 gives $\mathbb{E}a_T \le \sqrt{\mathbb{E}a_T^2} \le \sqrt{K_a}\, T^{-3/5-\varepsilon}$. Thus

$$\mathbb{E}\|z^T - z^*_{\tau_T}\|_2 \le \mathbb{E}a_T + \mathbb{E}\|\widetilde{z}^T - z^*_{\tau_T}\|_2 \le C_{A,\varepsilon} T^{-1/10+\varepsilon},$$

because $3/5 + \varepsilon > 1/10 - \varepsilon$. This proves $\mathbb{E}\|z^T - z^*_{\tau_T}\|_2 \le C_{A,\varepsilon} T^{-1/10+\varepsilon}$.

Finally, by Lemma F.8,

$$\mathrm{Gap}(z^T) \le \mathrm{Gap}(z^*_{\tau_T}) + \sqrt{2}L\|z^T - z^*_{\tau_T}\|_2 \le 4\tau_T + \sqrt{2}L\|z^T - z^*_{\tau_T}\|_2.$$

Taking expectation and using $\tau_T = T^{-1/10+\varepsilon}$ gives Eq. 17. For the actual sampling strategy, from $p_i^T - z_i^T = \epsilon_T(a_i - z_i^T)$,

$$\|p^T - z^T\|_2 \le 2\epsilon_T = 2T^{-1/5+2\varepsilon}.$$

Using the Lipschitz property of the gap again,

$$\mathrm{Gap}(p^T) \le \mathrm{Gap}(z^T) + 2\sqrt{2}L\, T^{-1/5+2\varepsilon}.$$

Since $1/5 - 2\varepsilon > 1/10 - \varepsilon$, taking expectation finishes the whole proof. □

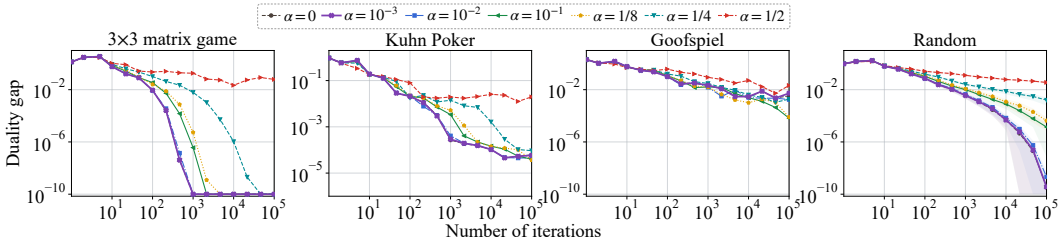


Figure 2: Last-iterate convergence of exploitability gap of Algorithm 1 with different $\alpha$.

## G Additional Experiments

Figure 2 further investigates the effect of the radial exponent $\alpha$ on the last-iterate convergence of Radial IREG-PRM$^+$. Overall, the algorithm exhibits stable convergence behavior across a broad range of $\alpha$ values on all four game instances, although the convergence speed is clearly affected by the choice of $\alpha$. On the $3 \times 3$ matrix game, smaller positive values of $\alpha$, such as $\alpha = 10^{-3}$ and $\alpha = 10^{-2}$, lead to substantially faster convergence, whereas larger values result in noticeably slower decay of the duality gap. A similar trend can be observed on the random matrix game, where small $\alpha$ achieves the best asymptotic performance, while large $\alpha$ values, especially $\alpha = 1/2$, converge much more slowly. On Kuhn Poker and Goofspiel, the dependence on $\alpha$ is less pronounced: most choices of $\alpha$ attain comparable final exploitability, although smaller or moderate values generally yield faster convergence during the transient phase. In contrast, the case $\alpha = 0$, corresponding to the original IREG-PRM$^+$ update without radial expansion, may fail to exhibit last-iterate convergence and can stagnate at a nonzero duality gap. These observations are qualitatively consistent with our theoretical analysis, which predicts that positive radial expansion promotes last-iterate stability and that smaller $\alpha$ leads to a faster stretched-exponential decay.