HIGH-PROBABILITY LAST-ITERATE CONVERGENCE FOR ADAPTIVELY REGULARIZED MIRROR DESCENT IN GAMES WITH BANDIT FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Regularization is a powerful technique for establishing last-iterate convergence, typically by augmenting the payoff functions with auxiliary terms that induce strong monotonicity of the game operator to stabilize learning dynamic. Existing approaches commonly rely on a gradually vanishing regularization coefficient so that the resulting last-iterate strategies can approach an NE of the original unregularized game. In this work, we study two-player zero-sum games under bandit feedback, where each player observes only the payoff associated with its sampled action, and introduce a fixed-coefficient regularization framework that perturbs the payoff based on the distance to an anchoring strategy. Specifically, we propose an adaptively regularized mirror descent algorithm equipped with a log-barrier regularizer, which periodically updates the anchoring strategy while keeping the regularization coefficient fixed. Through a dual-norm analysis, we prove that the resulting uncoupled algorithm achieves an anytime high-probability last-iterate convergence rate of $\tilde{\mathcal{O}}(t^{-1/4})$ in both matrix games and extensive-form games (EFGs). We further establish a constant-probability lower bound of $\Omega(T^{-1/2})$ over a fixed horizon under the same strictly uncoupled bandit information model, revealing the remaining gap between our upper bound and the best possible convergence rate suggested by the lower bound. Experiments on four EFG benchmarks demonstrate the competitive performance of our method against existing baselines.

1 INTRODUCTION

A two-player zero-sum game models a competitive interaction in which one player's gain is exactly the other player's loss. A central objective in such games is to compute a minimax strategy, or equivalently, a Nash equilibrium (NE) (Neumann, 1928), at which neither player can improve their payoff by unilaterally deviating from the current strategy. A classical approach to equilibrium computation is regret minimization (Cesa-Bianchi & Lugosi, 2006), where each player independently employs a no-regret online learning algorithm over their strategy space. The time-averaged strategies of the two players then converge to an approximate NE (Hart & Mas-Colell, 2001). This paradigm has played a central role in the development of superhuman AI agents for competitive games, such as Go (Silver et al., 2017), Poker (Brown & Sandholm, 2018), and Stratego (Perolat et al., 2022).

Despite their success in guaranteeing convergence of time-averaged strategies, many classical no-regret algorithms fail to converge in their last iterates. In particular, Mertikopoulos et al. (2018); Bailey & Piliouras (2018); Cai et al. (2025a) rigorously demonstrate the divergence for widely used algorithms, including online mirror descent and regret matching. Last-iterate convergence is nevertheless particularly desirable in practice, as it avoids the need to compute averaged policies. Motivated by this observation, there has been growing interest in developing learning dynamics with last-iterate convergence guarantees, particularly for *uncoupled* online algorithms that are implemented through self-play without communication among players (Daskalakis & Panageas, 2019; Wei et al., 2021; Lee et al., 2021; Cai et al., 2024). Among the existing approaches, regularization-based algorithms have emerged as a particularly effective paradigm, offering both last-iterate convergence guarantees and strong empirical performance (Perolat et al., 2021; Abe et al., 2023).

Regularization perturbs the original payoffs with auxiliary terms that induce strong monotonicity of the underlying game operator (Facchinei & Pang, 2003), thereby ensuring a unique NE of the

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.

regularized game. To recover an NE of the original game, existing methods under gradient feedback can be broadly divided into two groups. **(1)** The first gradually decreases the regularization coefficient over time so that the last-iterate strategy of the regularized games approach a NE of the original game, as studied in matrix games (Cen et al., 2021) and extensive-form games (EFGs) (Liu et al., 2023; Meng et al., 2025). **(2)** The second keeps regularization coefficient fixed while periodically updating an anchoring strategy, thereby adapting the effective regularization through the distance between the anchor and current strategy. This approach has been studied in continuous-time imperfect-information games (Perolat et al., 2021) and discrete-time monotone games (Abe et al., 2024; 2025).

Compared with the gradient-feedback setting, last-iterate convergence under the more challenging bandit feedback model (Auer et al., 2002; Zimmert & Seldin, 2021), where each player observes only the payoff associated with their sampled action, remains much less understood. Using entropy regularization, Cai et al. (2023) first establish a high-probability last-iterate convergence rate of $\mathcal{O}(t^{-1/8})$ for matrix games, which is subsequently extended to EFGs using a weighted negentropy regularizer (Zhao et al., 2025). More recently, Fiegel et al. (2026) employ log-barrier regularization and obtain a high-probability convergence rate of $\tilde{\mathcal{O}}(t^{-1/4})$ for both matrix games and EFGs. However, all these regularization-based approaches rely on carefully designed schedules of vanishing regularization coefficients. This naturally raises the following question: *Can an uncoupled bandit algorithm achieve last-iterate convergence without decreasing the regularization coefficient, as is possible under gradient feedback?* We answer this question in the affirmative. Specifically, our main contributions for two-player zero-sum games under bandit feedback are summarized as follows:

- We introduce a fixed-coefficient regularization framework in which the regularization effect is controlled through an adaptively updated anchoring strategy rather than a vanishing coefficient. Specifically, we propose an adaptively regularized mirror descent algorithm equipped with a log-barrier regularizer. The resulting uncoupled algorithm periodically updates the anchor while keeping the regularization coefficient fixed, and achieves an anytime high-probability last-iterate convergence rate of $\tilde{\mathcal{O}}(t^{-1/4})$, matching the best known rate (Fiegel et al., 2026).
- We establish a constant-probability lower bound of $\Omega(T^{-1/2})$ for last-iterate convergence over a fixed horizon for any uncoupled algorithm under the considered bandit information model. Consequently, no algorithm in this class can guarantee an anytime last-iterate convergence rate of $o(t^{-1/2})$ with sufficiently high probability. This result characterizes the remaining gap between the lower bound and our current upper bound of $\tilde{\mathcal{O}}(t^{-1/4})$ on the last-iterate convergence rate.
- We extend both the proposed algorithm and its anytime high-probability convergence guarantee of $\tilde{\mathcal{O}}(t^{-1/4})$ to perfect-recall EFGs by employing a dilated log-barrier regularizer.
- Compared with the related work of Fiegel et al. (2026), which also relies on a dual-space analysis, we develop a refined matrix-analytic argument for controlling the dual norm of the projected residual in Section 3.1. This analysis yields a smaller Lipschitz constant for the game operator with respect to the dual norm in both matrix games and EFGs (see Lemmas 3.3 and 4.1). Beyond its role in our convergence analysis, this result provides further insight into the advantages of dual-norm geometry for the analysis of learning dynamics in games.

2 Preliminary and Algorithms

For any integer $d \in \mathbb{N}_+$, let $\Delta_d = \{x \in \mathbb{R}^d_+ : \sum_{i=1}^d x_i = 1\}$ denote the $(d-1)$-dimensional simplex, $\mathbf{1}_n \in \mathbb{R}^n$ denote the all-ones vector. This work mainly focuses on two-player zero-sum games.

Matrix Games. Let the payoff matrix be denoted by $A \in [0,1]^{m\times n}$, where m is the number of actions available to the minimizing player (i.e. the *min-player*) and n is the number of actions available to the maximizing player (i.e. the *max-player*). The min-player chooses a mixed strategy $\mu \in \Delta_m$, and the max-player chooses a mixed strategy $\nu \in \Delta_n$. The expected loss under the mixed strategy profile (μ, ν) is given by the bilinear form

$$\ell(\mu, \nu) = \mathbb{E}_{a\sim\mu, b\sim\nu}[\ell(a, b)] = \mu^\top A\nu,$$

where $\ell(a, b) \in [0, 1]$ denotes the stochastic loss realized when the min-player plays action a and the max-player plays action b. Let $\mathcal{W} = \Delta_m \times \Delta_n$ denote the set of all mixed strategy profiles. A profile $w = (\mu, \nu) \in \mathcal{W}$ is called a *Nash equilibrium* (equivalently, a minimax profile) if it satisfies

$$\mu \in \arg\min_{\mu'\in\Delta_m} \ell(\mu', \nu), \qquad \nu \in \arg\max_{\nu'\in\Delta_n} \ell(\mu, \nu').$$

Algorithm 1: Adaptively Regularized OMD with Log-Barrier

Input: Dimensions m, n; fixed coefficient $\lambda > 0$; update times $\{T_k\}$; parameters $\eta, a, T_{\text{off}}, \delta$.

Initialize $w_0 = (\mu_0, \nu_0)$ to satisfy $\|\Pi_0 (F(w_0) + \alpha_0 \nabla\Phi(w_0))\|_{*,w_0} \le \alpha_0/8$. Set $k \leftarrow 0$;

for $t = 0, 1, 2, \ldots$ **do**

  if $t = T_k$ **then**

    Set epoch-start anchor $\sigma^k \leftarrow w_t$, decreasing factor $\rho_k \leftarrow \alpha_{T_k}/\lambda$.

    For each simplex block, find the unique β_k satisfying $\sum_i \frac{1}{(1-\rho_k)/\sigma_i^k + \beta_k} = 1$.

    Set $\widehat{\sigma}_i^k \leftarrow \frac{1}{(1-\rho_k)/\sigma_i^k + \beta_k}$; // $\widehat{\sigma}^k = \arg\min_{w \in \mathrm{ri}(\mathcal{W})} \Phi(w) - (1-\rho_k)\langle \nabla\Phi(\sigma^k), w\rangle$

  Set $\eta_t \leftarrow \eta(t + T_{\text{off}})^{-3/4}$, $\alpha_t \leftarrow a \log((t + T_{\text{off}})/\delta)(t + T_{\text{off}})^{-1/4}$.

  Sample $a_t \sim \mu_t$ and $b_t \sim \nu_t$; observe the single loss sample $\ell_t \in [0, 1]$.

  Construct loss estimators $\widehat{\ell}^t_{\min}, \widehat{\ell}^t_{\max}$ with equation 2 and set $\widehat{F}_t = (\widehat{\ell}^t_{\min}, \widehat{\ell}^t_{\max})$.

  Update $w_{t+1} \leftarrow \arg\min_{w \in \mathcal{W}} \left\{ D_\Phi(w, w_t) + \eta_t \left\langle \widehat{F}_t + \lambda(\nabla\Phi(w_t) - \nabla\Phi(\widehat{\sigma}^k)), w \right\rangle \right\}$.

  if $t + 1 = T_{k+1}$ **then** $k \leftarrow k + 1$;

To measure the proximity of a profile w to the set of minimax profiles, we use the *exploitability gap*:

$$\mathrm{EG}(\mu, \nu) = \max_{\nu' \in \Delta_n} \ell(\mu, \nu') - \min_{\mu' \in \Delta_m} \ell(\mu', \nu).$$

This quantity is zero if and only if (μ, ν) is a Nash equilibrium, in which neither player can improve their utilities by unilaterally deviating from the current profile.

Variational Inequality Formulation. Following the standard approach in the literature Wei et al. (2021), we reformulate the equilibrium problem as a monotone variational inequality. Let $K = m+n$. Define the *pseudo-gradient* operator $F : \mathcal{W} \to \mathbb{R}^K$ in nonnegative-coordinate form as

$$F(w) = \begin{pmatrix} \nabla_\mu \ell(\mu, \nu) \\ \mathbf{1}_n - \nabla_\nu \ell(\mu, \nu) \end{pmatrix} = Jw + c, \qquad J = \begin{pmatrix} 0 & A \\ -A^\top & 0 \end{pmatrix}, \qquad c = \begin{pmatrix} 0_m \\ \mathbf{1}_n \end{pmatrix}. \tag{1}$$

Observe that $J^\top = -J$, i.e., J is skew-symmetric, which implies that F is a monotone operator. A point $w^\star \in \mathcal{W}$ is said to solve the *variational inequality* problem associated with F if

$$\langle F(w^\star), w^\star - u \rangle \le 0 \qquad \forall u \in \mathcal{W}.$$

For the bilinear zero-sum game under consideration, the set of solutions to this variational inequality coincides exactly with the set of Nash equilibria. Moreover, the exploitability gap admits the following equivalent variational characterization: $\mathrm{EG}(w) = \max_{u \in \mathcal{W}} \langle F(w), w - u \rangle$.

Bandit Online Learning. We now turn to the online learning setting. At each round $t \ge 1$, the two players independently sample actions (a_t, b_t) according to their current strategies $(\mu_t, \nu_t) \in \mathcal{W}$. They then jointly observe a random loss $\ell_t \in [0, 1]$ drawn from the distribution associated with the action pair (a_t, b_t). Critically, the players do not observe the losses that would have resulted from other action pairs—this is the hallmark of *bandit feedback* (Lattimore & Szepesvári, 2020).

Let $\mathcal{F}_t = \sigma(\omega, a_1, b_1, \ell_1, \ldots, a_{t-1}, b_{t-1}, \ell_{t-1})$ denote the natural filtration (i.e. σ-algebra) generated by the observations up to time t. The sequence of profiles (μ_t, ν_t) is required to be predictable with respect to (w.r.t.) this filtration, meaning that each profile is chosen based only on past observations. Since the pseudo-gradient $F(w_t)$ is not directly observable under bandit feedback, we rely on the importance-sampling estimators for min-player/max-player players, which are defined as

$$\widehat{\ell}^{\min}_t(a) = \frac{\ell_t}{\mu_t(a)} \mathbf{1}\{a_t = a\}, \qquad \widehat{\ell}^{\max}_t(b) = \frac{1 - \ell_t}{\nu_t(b)} \mathbf{1}\{b_t = b\}, \tag{2}$$

and let $\widehat{F}_t = (\widehat{\ell}^{\min}_t, \widehat{\ell}^{\max}_t) \in \mathbb{R}^K$. These estimators are unbiased in the sense that $\mathbb{E}[\widehat{F}_t \mid \mathcal{F}_t] = F(w_t)$, where the expectation is taken w.r.t. the randomization in actions a_t, b_t and the loss realization ℓ_t.

Log-Barrier Regularization and Primal/Dual Norms. To solve the variational inequality in this bandit-feedback setting, we employ an online mirror descent algorithm. Let $\Phi : \mathrm{int}(\mathcal{W}) \to \mathbb{R}$ be the *log-barrier* regularizer defined by $\Phi(w) = -\sum_{i=1}^K \log w_i, \qquad w = (\mu_1, \ldots, \mu_m, \nu_1, \ldots, \nu_n)$.

Its gradient and Hessian are given by $\nabla\Phi(w)_i = -{w_i}^{-1}, H(w) = \nabla^2\Phi(w) = \mathrm{diag}(w_i^{-2})$. The associated Bregman divergence is $D_\Phi(x, y) = \Phi(x) - \Phi(y) - \langle\nabla\Phi(y), x - y\rangle$. We also define the local primal and dual norms induced by the Hessian at $w \in \mathbb{R}^K_{>0}$:

$$\|x\|_w^2 = \langle x, H(w)x\rangle = \sum_{i=1}^K x_i^2/w_i^2, \qquad \|g\|_{*,w}^2 = \left\langle g, H(w)^{-1}g\right\rangle = \sum_{i=1}^K w_i^2 g_i^2.$$

Accordingly, the induced inner-products are $\langle x, g\rangle_w = \langle x, H(w)g\rangle$ and $\langle x, g\rangle_{*,w} = \left\langle x, H(w)^{-1}g\right\rangle$ respectively. A crucial structural property of the product-of-simplices domain $\mathcal{W} = \Delta_m \times \Delta_n$ is that its normal space at any relative interior point $w \in \mathrm{ri}(\mathcal{W})$ is given by the blockwise-constant subspace

$$\mathcal{G} = \left\{g \in \mathbb{R}^K : g_i = \gamma_1 \text{ for } i \le m,\ g_{m+j} = \gamma_2 \text{ for } j \le n\right\}.$$

For any $g \in \mathcal{G}$, any strategy profile $w = (\mu, \nu) \in \mathcal{W}$, the inner product $\langle g, w\rangle$ is constant, since $\langle \mathbf{1}_m, \mu\rangle = \langle \mathbf{1}_n, \nu\rangle = 1$. Hence, two update directions that differ by an element of $\mathcal{G}$ yield the same mirror-descent update when projected onto the simplex.

Adaptively Regularized Mirror Descent with Log-Barrier. At the t-th iteration, the regularized gradient of mirror descent update is $\xi_t = \widehat{F}_t + \lambda(\nabla\Phi(w_t) - \nabla\Phi(\widehat{\sigma}^k))$, where $k = k(t)$ represents the k-th epoch the step t locates in, λ is the fixed regularization coefficient, and $\widehat{\sigma}^k$ is the boosted anchoring strategy that is repeatedly at a predefined interval $t = T_k$, by solving a constrained problem $\widehat{\sigma}^k = \arg\min_{w\in\mathrm{ri}(\mathcal{W})} \Phi(w) - (1-\rho_k)\langle\nabla\Phi(\sigma^k), w\rangle$ with $\rho_k = \alpha_{T_k}/\lambda$. The construction of anchoring strategy $\widehat{\sigma}^k$ is equivalent to $\nabla\Phi(\widehat{\sigma}^k) = (1-\rho_k)\nabla\Phi(\sigma^k) + \beta_k\mathbf{1}$, which is inspired from the recent anchor-strategy-boosting technique (Abe et al., 2025; Cai & Zheng, 2023) to accelerate convergence under the fixed-coefficient regularized games. The pseudo code is shown in Algorithm 1.

3 Last-Iterate Convergence in Matrix Games

In this section, we give the demonstration framework for the high-probability last-iterate convergence of Algorithm 1 in matrix games. Let $\Pi_t : \mathbb{R}^K \longrightarrow \mathcal{G}^{\perp,*,w_t}$ be the projection operator that maps K-dimensional space $\mathbb{R}^K$ into the space $\mathcal{G}^{\perp,*,w_t}$, which is orthogonal to the normal space $\mathcal{G}$ under the induced inner-product $\langle\cdot,\cdot\rangle_{*,w_t}$. Write $H_t = \nabla^2\Phi(w_t)$ for shortness when the context is clear.

3.1 Bounding Exploitability Gap by Minimizing Dual Residual

Define projected residual $p_t = \Pi_t\big(F(w_t) + \alpha_t\nabla\Phi(w_t)\big)$, then the upper bound on the dual norm $\|p_t\|_{*,w_t}$ of projected residual can translate into an upper bound on the exploitability gap of strategy w_t, as below, which is a direct application of (Fiegel et al., 2026, Lemma 6.4).

Lemma 3.1 (Dual Residual of Log-Barrier Controls Exploitability)**.** *Define* $p = \Pi_w\big(F(w) + \alpha\nabla\Phi(w)\big), d = \|p\|_{*,w}^2$. *If* $d \le \alpha^2$, *then* $\mathrm{EG}(w) \le K(\sqrt{d} + \alpha) \le 2K\alpha$.

The above result holds in both gradient and bandit feedback setting. In the bandit feedback, let $s_t = s_0 + t, l_t = \log(s_t/\delta)$, we define the adapted process $\{U_t = \frac{d_t}{a^2 l_t}\}_{t\ge0}$ and leverage stopping time theory (Chung, 1974) to obtain the following upper bound on residual's dual norm $d_t = \|p_t\|_{*,w_t}^2$.

Proposition 3.2 (High-probability Bound for Dual Residual)**.** *Assume* $s_0 \ge 4l_0$. *Define the stopping time* $\mathcal{T}_U := \inf\left\{t \ge 0 : U_t > \vartheta^2\frac{l_t}{\sqrt{s_t}}\right\}$ *with* $\vartheta \in (0, 1/2)$. *Then, for any step* $t < \mathcal{T}_U$,

$$U_{t+1} \le \left(1 - \frac{1}{s_t}\right)U_t + Bs_{t+1}^{-3/2} + W_{t+1}, \quad \textit{with} \quad \mathbb{E}_t[W_{t+1}] = 0, \quad |W_{t+1}|^2 \le Vs_{t+1}^{-3/2}U_t. \tag{3}$$

If the parameters further satisfy $U_0 \le \frac{\vartheta^2}{4}\frac{l_0}{\sqrt{s_0}}, \quad B \le \frac{\vartheta^2 l_0}{16}, \quad V \le \frac{\vartheta^2}{128}$, *then* $\mathbb{P}(\mathcal{T}_U < \infty) \le \delta$.

equation 3 in Proposition 3.2 tells us to establish the recursive expression of $d_{t+1} - d_t$. To this end, our work employs the *Taylor expansion analysis* to derive elegant results. Denote $\mathcal{A}_\alpha(w) := F(w) + \alpha\nabla\Phi(w)$. Since Π_t is the $\langle\cdot,\cdot\rangle_{*,w_t}$-orthogonal projection, there exists a unique $g_t \in \mathcal{G}$ such that $p_t = \mathcal{A}_{\alpha_t}(w_t) - g_t$, and $d_t = \|p_t\|_{*,w_t}^2$. In this section, g_t is regarded as a fixed vector. Let $q_t(w, \alpha) := \mathcal{A}_\alpha(w) - g_t$, then define the (frozen) residual potential function

$$\Psi_t(w, \alpha) := \|q_t(w, \alpha)\|_{*,w}^2. \tag{4}$$

Then for any $t \geq 1$, we have $\Psi_t(w_t, \alpha_t) = d_t$, and

$$d_{t+1} = \min_{g \in \mathcal{G}} \left\| \mathcal{A}_{\alpha_{t+1}}(w_{t+1}) - g \right\|_{*, w_{t+1}}^2 \leq \left\| \mathcal{A}_{\alpha_{t+1}}(w_{t+1}) - g_t \right\|_{*, w_{t+1}}^2 = \Psi_t(w_{t+1}, \alpha_{t+1}). \quad (5)$$

Therefore, $d_{t+1} - d_t \leq \Psi_t(w_{t+1}, \alpha_{t+1}) - \Psi_t(w_t, \alpha_t)$. It suffices to study the variation of the same scalar function Ψ_t between (w_t, α_t) and (w_{t+1}, α_{t+1}). Denote the update vector $\Delta_t := w_{t+1} - w_t$, and the scalar variation $\Delta_{\alpha,t} = \alpha_{t+1} - \alpha_t$. Let $\delta_t = \begin{pmatrix} \Delta_t \\ \Delta_{\alpha,t} \end{pmatrix}$, and $w_{t,s} = w_t + s\Delta_t$, $\alpha_{t,s} = \alpha_t + s\Delta_{\alpha,t}$ for any $s \in [0, 1]$. From equation 5 and the integral-form Taylor formula,

$$d_{t+1} - d_t \leq \Psi_t(w_{t+1}, \alpha_{t+1}) - \Psi_t(w_t, \alpha_t) = \langle \nabla \Psi_t(w_t, \alpha_t), \delta_t \rangle + \mathcal{R}_t^{\Psi}, \quad (6)$$

where $\mathcal{R}_t^{\Psi} = \int_0^1 (1-s) \delta_t^\top \nabla^2 \Psi_t(w_{t,s}, \alpha_{t,s}) \delta_t \, \mathrm{d}s$ is the second-order remainder term of the Taylor expansion. Then, we introduce a useful lemma that is used to bound each term in the Taylor formula of equation 6. Such a result may be of independent interest in the study of game theory.

Lemma 3.3 (Uniform Upper Bound for the Whitened Game Operator)**.** *Denote game operator* $J = \begin{pmatrix} 0_{m \times m} & A \\ -A^\top & 0_{n \times n} \end{pmatrix}$, $W(w) = \mathrm{Diag}(w)$. *Define* $\|A\|_{\max} := \max_{1 \leq i \leq m,\, 1 \leq j \leq n} |A_{ij}|$, *Then, for any* $w \in \mathrm{ri}(\mathcal{W})$, *the whiten game operator satisfies* $\left\| W(w) J W(w) \right\|_{\mathrm{op}} \leq \|A\|_{\max} \leq 1$.

Remark 3.4 (Two improvements of Lemma 3.3)**. (1)** When using the common ℓ_2-norm, $\|F(w) - F(w')\|_2 \leq \|A\|_{\mathrm{op}} \|w - w'\|_2$, so F is $\sqrt{nm}$-Lipschitz continuous w.r.t. the ℓ_2 norm. In contrast, using Lemma 3.3. $\|F(w) - F(w')\|_{*,w} = \left\| W(w) J W(w) W(w)^{-1} (w - w') \right\|_2 \leq \|w - w'\|_w$, so F is 1-Lipschitz continuous w.r.t. the primal norm $\|\cdot\|_w$, leading to a much smaller Lipschitz factor. **(2)** Previous work (Fiegel et al., 2026, Lemma 6.2) also utilizes the dual norm, but yields a Lipschitz factor of $\sqrt{n+m}$. In contrast, our Lemma 3.3 employs a more technical matrix analysis to derive a much smaller Lipschitz factor 1 (see proof in Sect D.1).

Let $M := 2 + 4a\sqrt{K} + \frac{\lambda\sqrt{K}}{2}$ and $\mathcal{B}_t := 2\sqrt{d_t} + (2 + \alpha_t) \|\Delta_t\|_{w_t} + \sqrt{K} |\Delta_{\alpha,t}|$. Define the update residual $r_t^{\mathrm{md}} := -H_t^{-1} \Pi_t(\nabla \Phi(w_{t+1}) - \nabla \Phi(w_t) - H_t \Delta_t)$. Then we can obtain two following bounds on the first-order and second-order terms of Taylor expansion in equation 6.

Proposition 3.5 (Upper Bound for the First-Order Taylor Approximation)**.** *Assume that* $\eta_{T_k} \leq \frac{1}{6M(H_k+1)}$ *for any epoch index k. Then almost surely,*

$$\begin{aligned} \langle \nabla \Psi_t(w_t, \alpha_t), \delta_t \rangle \leq & -2\eta_t \alpha_t d_t + 2\eta_t d_t^{3/2} + \eta_t \alpha_t d_t + \eta_t \alpha_t (1 + \alpha_t) \sqrt{d_t} + 2\sqrt{K} |\Delta_{\alpha,t}| \sqrt{d_t} \\ & + 2(1 + \alpha_t) \left\| r_t^{\mathrm{md}} \right\|_{w_t} \sqrt{d_t} + 2 \left\| r_t^{\mathrm{md}} \right\|_{w_t} d_t + \mathcal{M}_{t+1}^{\Psi}, \end{aligned} \quad (7)$$

where martingale difference $\mathcal{M}_{t+1}^{\Psi}$ *satisfies* $\mathbb{E}_t[\mathcal{M}_{t+1}^{\Psi}] = 0$, *and* $|\mathcal{M}_{t+1}^{\Psi}| \leq 6\eta_t \big[(1 + \alpha_t) \sqrt{d_t} + d_t \big]$.

Proposition 3.6 (Upper Bound for the Second-Order Hessian Remainder)**.** *If* $\eta_{T_k} \leq \frac{1}{6M(H_k+1)}$, *then*

$$\mathcal{R}_t^{\Psi} \leq 4 \big[(1 + \alpha_t^2) + 4\mathcal{B}_t + \mathcal{B}_t^2 + 2\alpha_t \mathcal{B}_t \big] \|\Delta_t\|_{w_t}^2 + K(\Delta_{\alpha,t})^2. \quad (8)$$

Therefore, it remains to bound the primal norm of update vector $\|w_{t+1} - w_t\|_{w_t}$ in equation 8 and the primal norm of update residual $\left\| r_t^{\mathrm{md}} \right\|_{w_t}$ in equation 7.

3.2 Bounding Primal Norms of Update Vector and Update Residual

For one-step mirror descent with log-barrier regularization, we first give the following lemma.

Lemma 3.7. *Let* $w^+ = \arg\min_{u \in \mathcal{W}} \{ D_\Phi(u, w) + \eta \langle \xi, u \rangle \}$. *If* $\eta \|\xi\|_{*,w} \leq 1/4$, *then we have* $\|w^+ - w\|_w \leq 2\eta \|\xi\|_{*,w}$ *and coordinatewise* $\frac{1}{2} w_i \leq w_i^+ \leq 2w_i$.

Then it is used to prove the boundedness of the following intra-epoch relative-coordinate variation.

Proposition 3.8. *If the step size satisfies* $\eta_{T_k} \leq \frac{1}{6M(H_k+1)}$ *for any k, then* $\forall t \in [T_k, T_k + H_k]$,

$$R_t := \max_{i \in [K]} \left| \frac{w_{t,i}}{w_{T_k,i}} - 1 \right| \leq \frac{1}{2}. \quad (9)$$

Furthermore, we have $\eta_t \|\xi_t\|_{*,w_t} \leq 1/4$ *and* $\|w_{t+1} - w_t\|_{w_t} \leq 2\eta_t \|\xi_t\|_{*,w_t}$.

The proof for Proposition 3.8 is the most technical part in this work. Its proof sketch is as follow. We prove $R_t \le 1/2$ by induction. The base case $R_{T_k} = 0 \le 1/2$ is obvious. Then in the inductive step $t+1$, we use the assumption $R_t \le 1/2$ to demonstrate the boundedness of dual norm of the gradient $\|\xi_t\|_{*,w_t} \le M$. Therefore, let the step size η_t be small enough to satisfy the condition of Lemma 3.7, we have $\|w_{t+1} - w_t\|_{*,w_t} \le 2\eta_t M$. Expanding the left-hand side of this inequality coordinate-wise yields $R_{t+1} \le R_t + 3\eta_s M \le 3\eta_s M(H_k+1) \le 1/2$, completing the whole induction proof. Then, we use the result $R_t \le 1/2$ to prove the boundedness of gradient's dual norm $\|\xi_t\|_{*,w_t} \le M$ at every step t, finally leading to the bound on primal norm of update vector $\|w_{t+1} - w_t\|_{w_t} \le 2\eta_t \|\xi_t\|_{*,w_t}$ based on Lemma 3.7 (see more details in the proof for Proposition D.5-D.6).

On the other hand, since $\left\|r_t^{\mathrm{md}}\right\|_{w_t} \le \|\nabla\Phi(w_{t+1}) - \nabla\Phi(w_t) - H_t\Delta_t\|_{*,w_t}$, expanding the right-hand side of this inequality coordinate-wise and using the relation $\frac{1}{2}w_{t,i} \le w_{t+1,i} \le 2w_{t,i}$ gives the following upper bound for the primal norm of the update residual $\left\|r_t^{\mathrm{md}}\right\|_{w_t}$.

Proposition 3.9 (Projected Path Expansion)**.** *Recall in Algorithm 1 at step t, the gradient is $\xi_t = \lambda(\nabla\Phi(w_t) - \nabla\Phi(\bar{\sigma}^k))$. Then, we have the update vector decomposition $\Delta_t = -\eta_t H_t^{-1}\Pi_t(\xi_t) + r_t^{\mathrm{md}}$. Furthermore, if $\eta_{T_k} \le \frac{1}{6M(H_k+1)}$ for any epoch index k, then*

$$\left\|r_t^{\mathrm{md}}\right\|_{w_t} \le 2\|w_{t+1} - w_t\|_{w_t}^2 \le 8\eta_t^2 \|\xi_t\|_{*,w_t}^2 . \tag{10}$$

The update vector decomposition is derived without using the exact expression of regularization function $\Phi(w)$, revealing that the residual r_t^{md} measures the discrepancy between the actual update Δ_t and the primal-space update obtained by mapping the projected dual gradient through H_t^{-1}.

3.3 High-Probability Upper Bound on the Exploitability Gap

Define the stopping time $\mathcal{T} = \inf\left\{t \ge 0 : d_t > \vartheta^2\alpha_t^2\right\}$ with $\vartheta \in (0, 1/2)$. Define $\chi = 1 - \vartheta - c_e$ with $c_e \le 1/2$. Then plugging the upper bound on $\|w_{t+1} - w_t\|_{w_t}$ from Proposition 3.8 into equation 8, plugging the bound on $\left\|r_t^{\mathrm{md}}\right\|_{w_t}$ from Proposition 3.9 into equation 7, and then combining the Taylor expansion in equation 6, we can derive a descent inequality for the distance d_{t+1}. Concretely, there exist constants $C_0, C_1, C_2 > 0$, such that when T_{off} is large enough (see Proposition F.1):

$$d_{t+1} \le \big(1 - \chi\eta_t\alpha_t\big)d_t + C_0\eta_t^2 + C_1\frac{\alpha_t}{\eta_t s_t^2} + M_{t+1}, \tag{11}$$

where M_{t+1} is the martingale difference satisfying $|M_{t+1}|^2 \le C_2\eta_t^2 d_t$. Then, by normalizing $U_t = \frac{d_t}{a^2 l_t}$, we derive the following normalized residual recursive expression.

Proposition 3.10. *Suppose the conditions of equation 11 hold. Define $W_{t+1} = M_{t+1}/(a^2 l_{t+1})$. Then there exist constants $B_* = \frac{C_0\eta^2}{a^2} + \frac{2C_1}{a\eta}, V_\star = \frac{2C_2\eta^2}{a^2 l_0}$, such that for all $t < \mathcal{T}$,*

$$U_{t+1} \le \left(1 - \frac{1}{s_t}\right)U_t + B_* s_{t+1}^{-3/2} + W_{t+1}, \qquad |W_{t+1}|^2 \le V_\star s_{t+1}^{-3/2} U_t \tag{12}$$

Then, letting $l_0 = \log\frac{s_0}{\delta}$ be large enough to ensure $B_\star \le \frac{\vartheta^2 l_0}{16}, V_\star \le \frac{\vartheta^2}{128}$, and letting the initialization satisfy $U_0 = \frac{d_0}{a^2\ell_0} \le \frac{\vartheta^2}{4}\frac{l_0}{\sqrt{s_0}}$, applying Proposition 3.2 yields the following high-probability bound $\tilde{\mathcal{O}}(t^{-1/4})$ on the exploitability gap of last-iterate strategy.

Theorem 3.11 (High-Probability Anytime Last-Iterate convergence Rate)**.** *Consider Algorithm 1 with fixed epoch length $H_k \equiv H$, step size $\eta_t = \eta(t + T_{\mathrm{off}})^{-3/4}$, $\alpha_t = a\log\left(\frac{t+T_{\mathrm{off}}}{\delta}\right)(t + T_{\mathrm{off}})^{-1/4}$. Assume the parameters satisfy for any $k \in \mathbb{N}_+$:*

$$(\mathbf{1}) \quad \eta_{T_k} = \eta\,(T_k + T_{\mathrm{off}})^{-3/4} \le \frac{1}{6M(H_k+1)}, \quad T_{\mathrm{off}} \ge \max\left\{32HM\sqrt{K}, \frac{\lambda^2\eta^2}{a^2}\,32^2H^2M^2K\right\}$$

$$(\mathbf{2}) \quad T_{\mathrm{off}} \ge \max\left\{\mathcal{O}\left(\exp\left\{K(a+\lambda)^2\left(\frac{\sigma^2\eta^2}{a^2} + K^2H^2a\eta\right)\right\}\right), \mathcal{O}\left(\exp\left\{\frac{1}{\eta a}\right\}\right)\right\}.$$

Then, with probability at least $1-\delta$, we have for every $t \ge 0$:

$$\mathrm{EG}(w_t) \le 2Ka\log\left(\frac{t+T_{\mathrm{off}}}{\delta}\right)(t + T_{\mathrm{off}})^{-1/4}. \tag{13}$$

The main differences between ours and the most related work Fiegel et al. (2026) with best-known rate $\tilde{\mathcal{O}}(t^{-1/4})$ last-iterate convergence rate are summarized in two aspects:

- We use a different demonstration framework to minimize $d_t = \|p_t\|_{*,w_t}$. equation 6 uses Taylor expansion to bound $d_{t+1} - d_t$, facilitating more elegant matrix analysis (like smaller Lipschitz constant of game operator in Lemma 3.3). Besides, the core breakthrough point is to prove the intra-epoch relative-coordinate variation $R_t \leq 1/2$ in Proposition 3.8. In contrast, (Fiegel et al., 2026, Sect D) chooses to decompose $d_{t+1} - d_t$ into several terms, and the core point is to show $\left\|\Delta_t + \eta_t H_t^{-1} p_t\right\|_{w_t} \leq O(\eta_t^2)$ in their equation (2).
- We design a different uncoupled algorithm with fixed regularization strength and periodically updated reference strategy. In contrast, (Fiegel et al., 2026, Algorithm 1) decreases the regularization strength in every step without reference strategy. Due to the usage of reference strategy, our proof need to tackle an additional term e_t (see Lemma D.3) in the regularized payoff, and our result degenerates to the convergence rate in Fiegel et al. (2026) when epoch length $H_k = 1$ (Remark D.4).

3.4 Constant-Probability Lower Bound on the Exploitability Gap

In this section, we further provide lower bound analysis for the last-iterate convergence rate of uncoupled algorithms in games with bandit feedback. We first introduce necessary notations. The filtration $\mathcal{F}_t$ records the full analysis history before round t. For an *uncoupled* algorithm $\mathcal{A}$, each player is restricted to its own local history. Let $\omega_{\min}$ and $\omega_{\max}$ denote the players' private algorithmic seeds; their joint law is independent of the unknown game. Define the local filtrations $\mathcal{F}_t^{\min} = \sigma(\omega_{\min}, a_1, \ell_1, \ldots, a_{t-1}, \ell_{t-1})$, $\mathcal{F}_t^{\max} = \sigma(\omega_{\max}, b_1, \ell_1, \ldots, b_{t-1}, \ell_{t-1})$. Obviously, $\mathcal{F}_t^{\min} \subseteq \mathcal{F}_t$ and $\mathcal{F}_t^{\max} \subseteq \mathcal{F}_t$. The uncoupled measurability requirement is μ_t is $\mathcal{F}_t^{\min}$-measurable, ν_t is $\mathcal{F}_t^{\max}$-measurable. Given μ_t and ν_t, the current actions are generated from the current distributions, i.e., $a_t \sim \mu_t, b_t \sim \nu_t$. The two action draws are conditionally independent given the history preceding the actions. Formally, this means that

$$\mathbb{P}(a_t = a \,|\, \mathcal{F}_t^{\min}) = \mu_t(a), \quad \mathbb{P}(b_t = b \,|\, \mathcal{F}_t^{\max}) = \nu_t(b), \quad \mathbb{P}(a_t = a, b_t = b \,|\, \mathcal{F}_t) = \mu_t(a)\nu_t(b). \tag{14}$$

We then use the hard Bernoulli family from Fiegel et al. (2025) for in-probability lower bound analysis. For $\varepsilon \in [-1/12, 1/12]$, define a stationary stochastic game G^ε by the cell-wise Bernoulli observations and its corresponding mean matrix as

$$\mathsf{G}^\varepsilon = \begin{pmatrix} \mathrm{Ber}(2/3 - \varepsilon) & \mathrm{Ber}(1/3 + \varepsilon) \\ \mathrm{Ber}(1/3) & \mathrm{Ber}(2/3) \end{pmatrix}, \qquad A^\varepsilon = \begin{pmatrix} 2/3 - \varepsilon & 1/3 + \varepsilon \\ 1/3 & 2/3 \end{pmatrix}. \tag{15}$$

Conditionally on $\mathcal{F}_t$ and (a_t, b_t), the observation ℓ_t is an independent Bernoulli random variable with the parameter in the corresponding cell of equation 15. For each ε, let $\mathbb{P}_\varepsilon$ denote the law induced jointly by the fixed learning algorithm and the game G^ε. Then, we give the following fixed-horizon lower-bound for any uncoupled bandit learning algorithm.

Theorem 3.12 (Fixed-horizon constant-probability lower bound)**.** *Let $T \geq 4$. For every uncoupled bandit learning algorithm $\mathcal{A}$ satisfying equation 14, there exists $\varepsilon_T \in \{-\sqrt{3/(128T)}, +\sqrt{3/(128T)}\}$ such that $\mathbb{P}_{\varepsilon_T}\left(\mathrm{EG}_{A^{\varepsilon_T}}(\mu_T, \nu_T) \geq T^{-1/2}/8\right) \geq 1/4$.*

Such constant-probability lower bound $\Omega(T^{-1/2})$ guarantees that there is no uncoupled algorithm that can achieve *anytime* high-probability upper bound $o(t^{-1/2})$ on the last-iterate convergence rate. To the best of our knowledge, this is the first *in-probability* lower bound for the last-iterate convergence of *strictly uncoupled algorithms* in games with bandit feedback. Previous lower bound for this class of algorithms either holds in L_p-convergence (Fiegel et al., 2025, Thm 5.1) or presupposes the last-iterate duality gap can be written in a certain form $\Theta(t^{-\alpha})$ (Zhao et al., 2025, Thm 5.6). Furthermore, we show in the next proposition that, our fixed-horizon lower bound $\Omega(T^{-1/2})$ cannot be strengthened over the same algorithm class, to a larger order such as $\Omega(T^{-1/4})$.

Theorem 3.13 (Matching fixed-horizon upper bound)**.** *Fix $T \geq 8$ and $\delta \in (0, 1)$. There exists a horizon-aware pair of uncoupled algorithms satisfying equation 14 such that, for every stationary stochastic 2×2 zero-sum game whose cell observations are independent Bernoulli random variables with mean matrix $A \in [0, 1]^{2\times 2}$, $\mathbb{P}_A\left(\mathrm{EG}_A(\mu_T, \nu_T) \leq 4\sqrt{\log(8/\delta)/T}\right) \geq 1 - \delta$.*

Therefore, the fixed-horizon minimax rate is $\Theta(T^{-1/2})$ up to confidence-dependent constants, partially solving the open problem of the lower bound tightness for uncoupled convergence with high

probability Fiegel et al. (2025). Besides, the tightness of horizon-fixed lower bound $\Theta(T^{-1/2})$ also implies that our anytime last-iterate convergence $\tilde{O}(t^{-1/4})$ may not be optimal.

4 Last-Iterate Convergence in Extensive-Form Games

In this section, we further extend the algorithm and analysis to the EFGs with bandit feedback.

Extensive-Form Game (EFG) and Sequence-Form Strategy. Let $D_{\mathrm{g}} \in \mathbb{N}$ denote the depth of the extensive-form game. Consider a two-player zero-sum extensive-form game with perfect recall (Von Stengel, 1996). Let $\mathcal{X}$ and $\mathcal{Y}$ be the information sets of the *min-* and *max-players*. Let $\Sigma_{\min} = \{(x,a)|x \in \mathcal{X}, a \in \mathcal{A}(x)\}$ be set of all inforset-action pairs for *min-player*. $\Sigma_{\max}$ is defined analogously. For $(x,a) \in \Sigma_{\min}$, let $r(x,a)$ denote the minimizing player's realization weight of the sequence ending with (x,a). Similarly, let $s(y,b)$ denote the maximizing player's realization weight. The two sequence-form polytopes can be written as

$$\mathcal{Q}_{\min} = \{r \in \mathbb{R}_+^{K_{\min}} : E_{\min} r = e_{\min}\}, \quad \mathcal{Q}_{\max} = \{s \in \mathbb{R}_+^{K_{\max}} : E_{\max} s = e_{\max}\}, \tag{16}$$

where the matrix $E_{\min}$ encode the usual realization-flow equalities $\sum_{a \in \mathcal{A}(x)} r(x,a) = r(p(x))$, with $p(x)$ as the parent infoset-action pair before reaching x. The same holds for $E_{\max}$. Define

$$\mathcal{W}_{\mathrm{E}} = \mathcal{Q}_{\min} \times \mathcal{Q}_{\max}, \qquad w = (r,s), \qquad K_{\mathrm{E}} = K_{\min} + K_{\max}.$$

Equivalently, $\mathcal{W}_{\mathrm{E}} = \{w \in \mathbb{R}_+^{K_{\mathrm{E}}} : Ew = e\}$ for a block-diagonal flow matrix E. Its tangent space is $\mathcal{W}_{\mathrm{E},0} = \ker E$, and relative-interior normal space are $\mathcal{G}_{\mathrm{E}} = \mathrm{range}(E^\top)$.

For every terminal history $z \in \mathcal{Z}$, let $\sigma_{\min}(z) \in \Sigma_{\min}$ and $\sigma_{\max}(z) \in \Sigma_{\max}$ denote the two players' terminal inforset-action pair along z, and let $\pi_c(z) \geq 0$ denote the product of all chance and environment-transition probabilities along z after the two players' own realization factors have been removed. Let $L(z)$ be the total loss accumulated along z. We write $U_{\mathrm{E}} := \max_{z \in \mathcal{Z}} |L(z)|$. Each stage loss lies in $[0,1]$ and there are at most D_{g} stages, hence $U_{\mathrm{E}} \leq D_{\mathrm{g}}$.

For $i \in \Sigma_{\min}$, $j \in \Sigma_{\max}$, define $(A_{\mathrm{E}})_{ij} = \sum_{z \in \mathcal{Z}: \sigma_{\min}(z)=i,\, \sigma_{\max}(z)=j} \pi_c(z) L(z)$. Perfect recall implies that the reaching probability of a terminal history is $\mathbb{P}_{r,s}(z) = r_{\sigma_{\min}(z)} s_{\sigma_{\max}(z)} \pi_c(z)$. So,

$$\ell(r,s) = \sum_{z \in \mathcal{Z}} \mathbb{P}_{r,s}(z) L(z) = \sum_{i \in \Sigma_{\min}} \sum_{j \in \Sigma_{\max}} r_i (A_{\mathrm{E}})_{ij} s_j = r^\top A_{\mathrm{E}} s. \tag{17}$$

We use the linear pseudo-gradient $F_{\mathrm{E}}(w)$, which is monotone due to the skew-symmetry $J_{\mathrm{E}}^\top = -J_{\mathrm{E}}$,

$$F_{\mathrm{E}}(w) = \begin{pmatrix} A_{\mathrm{E}} s \\ -A_{\mathrm{E}}^\top r \end{pmatrix} = J_{\mathrm{E}} w, \qquad \text{with} \quad J_{\mathrm{E}} = \begin{pmatrix} 0 & A_{\mathrm{E}} \\ -A_{\mathrm{E}}^\top & 0 \end{pmatrix}. \tag{18}$$

Bandit Sequence-Form Estimator. Every information set belongs to a unique depth. Write $\mathcal{X} = \bigsqcup_{h=0}^{D_{\mathrm{g}}-1} \mathcal{X}_h$, $\mathcal{Y} = \bigsqcup_{h=0}^{D_{\mathrm{g}}-1} \mathcal{Y}_h$. Equivalently, the sequence coordinates split as $\Sigma_{\min} = \bigsqcup_{h=0}^{D_{\mathrm{g}}-1} \Sigma_{\min,h}$, $\Sigma_{\max} = \bigsqcup_{h=0}^{D_{\mathrm{g}}-1} \Sigma_{\max,h}$, where $(x_h,a) \in \Sigma_{\min,h}$ means that $x_h \in \mathcal{X}_h$ and $a \in \mathcal{A}(x_h)$, and analogously $(y_h,b) \in \Sigma_{\max,h}$ for the maximizing player.

At episode t, A trajectory of depth at most D_{g} is sampled. The current realization plans r_t, s_t induce behavioral policies by $\pi_t(a \mid x) = \frac{r_t(x,a)}{r_t(p(x))}$, $\varrho_t(b \mid y) = \frac{s_t(y,b)}{s_t(p(y))}$, with the root parent realization equal to one. At depth h, Let $x_h^t \in \mathcal{X}_h$, $a_h^t \in \mathcal{A}(x_h^t)$, $y_h^t \in \mathcal{Y}_h$, $b_h^t \in \mathcal{B}(y_h^t)$ denote the infosets visited and the actions selected, and let $\ell_{t,h} \in [0,1]$ be the observed stage loss.

Define sequence-form importance-sampling estimator as $\widehat{F}_t = (\widehat{F}_t^{\min}, \widehat{F}_t^{\max})$, whose coordinatewise components, for every h and every $(x_h,a) \in \Sigma_{\min,h}$ and $(y_h,b) \in \Sigma_{\max,h}$, are as follows:

$$\widehat{F}_t^{\min}(x_h,a) = \frac{\ell_{t,h}}{r_t(x_h,a)} \mathbf{1}\{(x_h^t,a_h^t) = (x_h,a)\}, \quad \widehat{F}_t^{\max}(y_h,b) = -\frac{\ell_{t,h}}{s_t(y_h,b)} \mathbf{1}\{(y_h^t,b_h^t) = (y_h,b)\}. \tag{19}$$

The whole pesudo code for the adaptively regularized mirror descent with log-barrier regularization in solving two-player zero-sum EFG is shown in Algorithm 2. The constructed bandit estimator in equation 19 is unbiased $\mathbb{E}_t[\widehat{F}_t] = \widetilde{F}_{\mathrm{E}}(w_t)$ modulo the normal space, i.e. $\widetilde{F}_{\mathrm{E}}(w_t) - F_{\mathrm{E}}(w_t) \in \mathcal{G}_{\mathrm{E}}$. Therefore, let Π_w^{E} denote the local-dual orthogonal projection onto the orthogonal complement of $\mathcal{G}_{\mathrm{E}}$ at a point w, then $\Pi_{w_t}^{\mathrm{E}} \widetilde{F}_{\mathrm{E}}(w_t) = \Pi_{w_t}^{\mathrm{E}} F_{\mathrm{E}}(w_t)$. Thus, we can adapt the convergence analysis from matrix game in Section 3 to give a high-probability bound on $d_t^{\mathrm{E}} = \left\| \Pi_t^E \left(F_{\mathrm{E}}(w_t) + \alpha_t \nabla \Phi_{\mathrm{E}}(w_t) \right) \right\|_{*,w_t}$. The only difference lies in the following upper bound on the norm of whiten game operator.

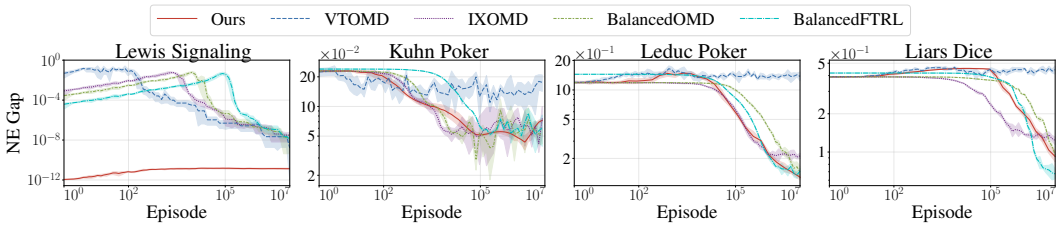


Figure 1: Last-iterate convergence of exploitability gap of our Algorithm 2 and other four baselines.

Lemma 4.1. *Let A_{E} be the canonical sequence-form payoff matrix in equation 17, and assume $|L(z)| \leq U_{\mathrm{E}}$ for every terminal history. Then, for every feasible sequence-form strategy profile $w = (r, s)$, let $D_r = \mathrm{Diag}(r)$ and $D_s = \mathrm{Diag}(s)$, we have $\|W(w)J_{\mathrm{E}}W(w)\|_{\mathrm{op}} \leq U_{\mathrm{E}} \leq D_{\mathrm{g}}$.*

As shown in Remark 3.4, Lemma 4.1 indicates that the game operator J_{E} is D_{g}-Lipschitz w.r.t. the dual norm $\|\cdot\|_{*,w}$. Such constant is smaller than the Lipschitz factor $D_{\mathrm{g}}\sqrt{K_{\mathrm{E}}}$ w.r.t. $\|\cdot\|_{*,w}$ in (Fiegel et al., 2026, Lemma 7.3). Then, we can extend the last-iterate convergence rate to the EFG setting.

Theorem 4.2 (high-probability last-iterate convergence in EFGs)**.** *Consider Algorithm 2 in a two-player zero-sum perfect-recall extensive-form game of depth D_{g}. Then, under almost the same conditions of T_{off} as that in Theorem 3.11, with probability at least $1-\delta$, we have for every $t \geq 0$,*

$$\mathrm{EG}(w_t) \leq 2K_{\mathrm{E}}a \log\left(\frac{t+T_{\mathrm{off}}}{\delta}\right)(t+T_{\mathrm{off}})^{-1/4}. \tag{20}$$

The first bandit algorithm to solve EFG is VTOMD (Zhao et al., 2025) that enjoys $\tilde{O}(t^{-1/8})$ high-probability last-iterate convergence, but needs to compute a virtual transition probability across the whole game tree before running OMD updates. In contrast, with log-barrier regularization, Algorithm 2 can directly perform OMD updates on the game tree. Moreover, the associated dual-norm analysis enables a unified framework for last-iterate convergence in matrix games and EFGs.

5 Experiments

This section reports the empirical performance of our proposed Algorithm 2 in EFGs. We compare the last-iterate convergence of our algorithm with that of four previous approaches, namely IXOMD (Kozuno et al., 2021), BalancedOMD (Bai et al., 2022), BalancedFTRL (Fiegel et al., 2023), and virtual transition-based OMD (VTOMD) (Zhao et al., 2025).

Experimental Settings and Results. We use four standard imperfect information EFG benchmarks: Lewis Signaling (Lewis, 2008), Kuhn Poker (Kuhn, 1950), Leduc Poker (Southey et al., 2005), and Liars Dice (Lisý et al., 2015). All game implementations are built on the OpenSpiel library (Lanctot et al., 2019). For other baselines, we follow the same hyperparameter setting as in (Zhao et al., 2025) for fair comparison. For our Algorithm 2, we set $T_{\mathrm{off}} = 100$, the learning rate $\eta = 10^{-2}$, epoch length $H = 10$, decreasing factor $a = 10^{-3}$, regularization coefficient $\lambda = 0.01$ in most experiments. According to Fig. 1, we can observe that: **(1)** Our method provides consistently competitive last-iterate convergence across all four games. Its advantage is especially striking in Lewis Signaling, while in the more challenging poker and Liars Dice environments it remains comparable to the strongest existing baselines. **(2)** The relative convergence speed can be game dependent. For example, BalancedFTRL, with optimal average-iterate complexity, exhibits the strongest last-iterate convergence in both Leduc Poker and Liars Dice, but converges much slower in the small games like Lewis Signaling and Kuhn Poker, whereas our method achieves strong final convergence across different game instances

6 Conclusions

This work proposes an adaptively regularized online mirror descent algorithm with log-barrier regularization to solve two-player zero-sum games under bandit feedback setting. This fully uncoupled algorithm periodically updates the boosted anchoring strategy to manipulate the regularization strength, finally ensuring an anytime last-iterate convergence rate $\tilde{\mathcal{O}}(t^{-\frac{1}{4}})$ in both matrix games and extensive-form games. The refined proofs result in a smaller Lipschitz constant of game operator with respect to the dual norm, shedding more light on the advantages of log-barrier in the last-iterate convergence of game theory. We also characterize the tight horizon-fixed lower bound $\Theta(T^{-1/2})$ that holds in probability for any uncoupled algorithm. Experiments on four EFG instances empirically show that the proposed algorithm achieve competitive performance against existing baselines.

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 at least one author.]. 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 reproduce the results of other baselines by directly running the code released at `https://github.com/ColoeredGalaxy/Last_ite_Convergence_in_EFGs/tree/main`. All source code for reproducing the experimental results in this paper are deferred to supplementary material.

REFERENCES

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, Kentaro Toyoshima, and Atsushi Iwasaki. Last-iterate convergence with full and noisy feedback in two-player zero-sum games. In *International Conference on Artificial Intelligence and Statistics (AISTAT)*, pp. 7999–8028, 2023.

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.

Kenshi Abe, Mitsuki Sakamoto, Kaito Ariu, and Atsushi Iwasaki. Boosting perturbed gradient ascent for last-iterate convergence in games. In *International Conference on Learning Representations (ICLR)*, 2025.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Yu Bai, Chi Jin, Song Mei, and Tiancheng Yu. Near-optimal learning of extensive-form games with imperfect information. In *International Conference on Machine Learning (ICML)*, pp. 1337–1382, 2022.

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.

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

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, 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. Fast last-iterate convergence of learning in games requires forgetful algorithms. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

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 Representation (ICLR)*, 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.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. In *Annual Conference on Neural Information Processing (NeurIPS)*, pp. 27952–27964, 2021.

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

Kai Lai Chung. Elsevier, 1974.

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.

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

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Stochastic regret minimization in extensive-form games. In *International Conference on Machine Learning (ICML)*, pp. 3018–3028, 2020.

Côme Fiegel, Pierre Ménard, Tadashi Kozuno, Rémi Munos, Vianney Perchet, and Michal Valko. Adapting to game trees in zero-sum imperfect information games. In *International Conference on Machine Learning (ICML)*, pp. 10093–10135, 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.

Côme Fiegel, Pierre Ménard, Tadashi Kozuno, Michal Valko, and Vianney Perchet. Optimal last-iterate convergence in matrix games with bandit feedback using the log-barrier. *CoRR*, 2026. doi: 10.48550/ARXIV.2604.15242.

Sergiu Hart and Andreu Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98(1):26–54, 2001.

Tadashi Kozuno, Pierre Ménard, Remi Munos, and Michal Valko. Learning in two-player zero-sum partially observable markov games with perfect recall. In *Advances in Neural Information Processing Systems*, pp. 11987–11998, 2021.

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

Harry Waldo Kuhn. *Extensive Games and the Problem of Information*, pp. 193–216. Princeton University Press, 1953.

Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. Monte carlo sampling for regret minimization in extensive games. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1078–1086, 2009.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *CoRR*, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

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.

David Lewis. *Convention: A Philosophical Study*. Wiley-Blackwell, Cambridge, MA, USA, 2008.

Viliam Lisý, Marc Lanctot, and Michael H. Bowling. Online monte carlo counterfactual regret minimization for search in imperfect information games. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 27–36, 2015.

Mingyang Liu, Asuman E. Ozdaglar, Tiancheng Yu, and Kaiqing Zhang. The power of regularization in solving extensive-form games. In *International Conference on Learning Representations (ICLR)*, 2023.

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

Linjian Meng, Tianpei Yang, Youzhi Zhang, Zhenxing Ge, Shangdong Yang, Tianyu Ding, Wenbin Li, Bo An, and Yang Gao. Efficient last-iterate convergence in solving extensive-form games. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 2703–271, 2018.

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.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 3003–3011, 2013.

Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas 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*, pp. 8525–8535, 2021.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Martin Schmid, Neil Burch, Marc Lanctot, Matej Moravcik, Rudolf Kadlec, and Michael Bowling. Variance reduction in monte carlo counterfactual regret minimization (VR-MCCFR) for extensive form games using baselines. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 2157–2164, 2019.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: Opponent modelling in poker. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 550–558, 2005.

Bernhard Von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 1996.

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 and Tuomas Sandholm. Finding and certifying (near-)optimal strategies in black-box extensive-form games. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 5779–5788, 2021.

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.

Yichi Zhou, Jialian Li, and Jun Zhu. Posterior sampling for multi-agent reinforcement learning: solving extensive games with imperfect information. In *International Conference on Learning Representations (ICLR)*, 2020.

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

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 905–912, 2007.

A Related Works

High-Probability Last-Iterate Convergence with Bandit Feedback. In two-player zero-sum games with bandit feedback, the first high-probability last-iterate convergence rate $\tilde{\mathcal{O}}(t^{-\frac{1}{8}})$ is proposed by Cai et al. (2023) for matrix games by adding entropy regularization into payoff function and implementing mirror descent over the perturbed simplex. Such algorithm is further extended to irreducible Markov games with a last-iterate convergence $\tilde{\mathcal{O}}(t^{-\frac{1}{9+\epsilon}})$ for any $\epsilon > 0$ Cai et al. (2023), and extended to EFGs with a rate $\tilde{\mathcal{O}}(t^{-\frac{1}{8}})$ by introducing a virtual transition probability (Zhao et al., 2025). On the other hand, by designing an average-to-last-iterate reduction algorithm, Cai et al. (2025b) improve the high-probability rate to $\tilde{\mathcal{O}}(t^{-\frac{1}{5}})$ in matrix games without using regularization technique. But such reduction still requires computing the running average of iterates, and it is remain unclear how to generalize this technique to the EFG setting. Using average-to-last reduction, Maiti et al. (2026) derive a last-iterate convergence $\tilde{\mathcal{O}}(t^{-\frac{1}{4}})$ for bilinear saddle-point problems over convex sets with optimistic follow-the-regularization-leader (OFTRL) algorithm. The recent work (Fiegel et al., 2025) derives a lower bound $\Omega(t^{-\frac{1}{4}})$ on exploitability in terms of L_p-convergence rate. The state-of-the-art last-iterate convergence rate for matrix games and EFGs is $\tilde{\mathcal{O}}(t^{-\frac{1}{4}})$ (Fiegel et al., 2026), which uses log-barrier regularization with dual analysis. Different from existing regularization-based works that rely on carefully shrinking the regularization strengths, our work propose a coefficient-fixed regularization framework that perturbs the payoff based on the distance to the anchoring strategy, and the adaptively regularized mirror descent algorithm with log-barrier regularization achieves the same $\tilde{\mathcal{O}}(t^{-\frac{1}{4}})$ convergence rate as the latest work (Fiegel et al., 2026).

Average-Iterate Convergence with Bandit Feedback. In two-player zero-sum games from bandit feedback, regret-minimization-based works mainly investigate the average-iterate convergence in the setting of imperfect-information extensive-form games (Kuhn, 1953). The most celebrated approach for solving this problem is the model-free sampling-based counterfactual regret minimization (CFR) (Zinkevich et al., 2007), which randomly samples one path from the root to a terminal node in the game tree, and updates the behavioral strategies along the path with regret matching (Lanctot et al., 2009; Schmid et al., 2019) or mirror descent algorithms Farina et al. (2020). The above methods typically build an unbiased importance-sampling estimate of the counterfactual regret their averaged policies achieve a high-probability convergence rate of $\tilde{\mathcal{O}}((X\sqrt{A} + Y\sqrt{B})/\sqrt{t})$ to NE, where X, Y are the number of information sets and A, B are the number of actions for the two players. Kozuno et al. (2021) propose to use biased importance-sampling estimator with implicit exploration (Neu, 2015) and attains the same high-probability convergence rate. Bai et al. (2022) further improve this rate to the near-optimal one $\tilde{\mathcal{O}}((\sqrt{XA} + \sqrt{YB})/\sqrt{t})$ by integrating novel balanced exploration policies in both the sampling and the local regret minimization steps. On the other hand, there also exist model-based exploration methods, e.g. Zhou et al. (2020) combines posterior sampling for reinforcement learning (Osband et al., 2013) with CFR, and Zhang & Sandholm (2021) constructs a confidence bound that accounts for uncertainty in the unknown chance-transition distributions, both yielding a rate no faster than $\tilde{\mathcal{O}}(AB\sqrt{S/t})$, where S is the number of states. In contrast, this work focuses on deriving high-probability anytime last-iterate convergence for matrix games and EFGs.

B Proof for Section 3.1: (I) Exploitability-Residual Conversion and High-probability Residual Bound

Proof of Lemma 3.1. By the definition of orthogonal projection, there exists $g \in \mathcal{G}$ such that

$$F(w) + \alpha\nabla\Phi(w) = p + g.$$

Then $\forall i \in [K]$,

$$F_i(w) = p_i + g_i + \frac{\alpha}{w_i}.$$

For any $u \in W$, since $\langle g, w - u\rangle = 0$,

$$\langle F(w), w - u\rangle = \sum_i p_i w_i - \sum_i p_i u_i + \alpha K - \alpha \sum_i \frac{u_i}{w_i}. \tag{21}$$

From

$$|w_i p_i| \leq \left(\sum_j w_j^2 p_j^2 \right)^{1/2} = \sqrt{d} \leq \alpha,$$

we have for any $i \in [K]$,

$$p_i \geq -\frac{\alpha}{w_i}.$$

therefore

$$-p_i u_i - \alpha \frac{u_i}{w_i} \leq 0. \tag{22}$$

At the same time, using Cauchy-Schwartz inequality,

$$\sum_i p_i w_i = \langle p, w \rangle \leq \|p\|_{*,w} \|w\|_w = \sqrt{Kd}. \tag{23}$$

Substituting equation 22 and equation 23 back into equation 21,

$$\langle F(w), w - u \rangle \leq \sqrt{Kd} + K\alpha \leq 2K\alpha.$$

. □

Lemma B.1 (Freedman Inequality)**.** *Let $\{X_j\}_{j \geq 1}$ be a martingale difference sequence with respect to the filtration, and suppose $|X_j| \leq R$ almost surely. Let*

$$S_n = \sum_{j=1}^{n} X_j, \qquad V_n = \sum_{j=1}^{n} \mathbb{E}_{j-1}[X_j^2].$$

Then for any $x, \sigma^2 > 0$,

$$\mathbb{P}\left(S_n \geq x,\ V_n \leq \sigma^2\right) \leq \exp\left(-\frac{x^2}{2(\sigma^2 + Rx/3)} \right). \tag{24}$$

Proof. For $0 \leq \lambda < 3/R$ and $|y| \leq R$, Taylor expansion and the geometric series give

$$e^{\lambda y} \leq 1 + \lambda y + \frac{\lambda^2 y^2}{2(1 - \lambda R/3)}.$$

Taking conditional expectation and using $\mathbb{E}_{j-1}[X_j] = 0$,

$$\mathbb{E}_{j-1}[e^{\lambda X_j}] \leq \exp\left(\frac{\lambda^2 \mathbb{E}_{j-1}[X_j^2]}{2(1 - \lambda R/3)} \right).$$

Therefore

$$Z_n(\lambda) := \exp\left(\lambda S_n - \frac{\lambda^2 V_n}{2(1 - \lambda R/3)} \right)$$

is a nonnegative supermartingale, and $\mathbb{E}[Z_n(\lambda)] \leq 1$. Markov's inequality gives

$$\mathbb{P}(S_n \geq x, V_n \leq \sigma^2) \leq \exp\left(-\lambda x + \frac{\lambda^2 \sigma^2}{2(1 - \lambda R/3)} \right).$$

Taking $\lambda = \frac{x}{\sigma^2 + Rx/3}$, yields equation 24. □

Proof of Proposition 3.2. Define the stopped martingale differences

$$X_{j+1} := s_j W_{j+1} \mathbf{1}\{j < \mathcal{T}_U\}, \qquad S_t := \sum_{j=0}^{t-1} X_{j+1}.$$

According to the recursive expression $U_{t+1} \leq \left(1 - \frac{1}{s_t}\right) U_t + B s_{t+1}^{-3/2} + W_{t+1}$, for any $t \geq 1$,

$$U_{t\wedge\mathcal{T}_U} \leq \underbrace{\frac{s_0 - 1}{s_{t\wedge\mathcal{T}_U - 1}} U_0}_{\text{(I)}} + \underbrace{\frac{1}{s_{t\wedge\mathcal{T}_U - 1}} \sum_{j=0}^{t\wedge\mathcal{T}_U - 1} s_j B s_{j+1}^{-3/2}}_{\text{(II)}} + \underbrace{\frac{S_{t\wedge\mathcal{T}_U}}{s_{t\wedge\mathcal{T}_U - 1}}}_{\text{(III)}}, \tag{25}$$

where we use the fact that $\prod_{r=j+1}^{t-1} \left(1 - \frac{1}{s_r}\right) = \frac{s_j}{s_{t-1}}$.

According to the condition of U_0 in Proposition 3.2, the first term (I) in equation 25 can be bounded as

$$\boxed{\text{(I)} = \frac{s_0 - 1}{s_{t-1}} U_0 \leq \frac{s_0 - 1}{s_{t-1}} \frac{\vartheta^2}{4} \frac{l_0}{\sqrt{s_0}} \leq \frac{\vartheta^2}{4} \frac{l_t}{\sqrt{s_t}},} \tag{26}$$

where the last inequality holds since function $(\sqrt{x} - \frac{1}{\sqrt{x}}) \ln \frac{x}{\delta}$ is increasing when $x \geq 1$ and

$$\frac{s_0 - 1}{s_t - 1} \frac{l_0}{\sqrt{s_0}} \leq \frac{l_t}{\sqrt{s_t}} \quad \Longleftrightarrow \quad \frac{s_0 - 1}{\sqrt{s_0}} l_0 \leq \frac{s_t - 1}{\sqrt{s_t}} l_t.$$

For the second term (II) in equation 25,

$$\sum_{j=0}^{t-1} s_j s_{j+1}^{-3/2} \leq \sum_{j=0}^{t-1} s_j^{-1/2} \leq 2\sqrt{s_t}.$$

Combining $s_t / s_{t-1} \leq 2$ and the condition of B in Proposition 3.2, we have

$$\boxed{\text{(II)} \leq \frac{2\sqrt{s_t}}{s_t - 1} \frac{\vartheta^2 l_0}{16} \leq \frac{\vartheta^2 l_0}{4\sqrt{s_t}} \leq \frac{\vartheta^2}{4} \frac{l_t}{\sqrt{s_t}}.} \tag{27}$$

For the third term (III) in 25, when $j < \mathcal{T}_U$,

$$|X_{j+1}|^2 \leq s_j^2 V s_{j+1}^{-3/2} U_j \leq V \vartheta^2 l_j,$$

Thus, for fixed t,

$$|X_{j+1}| \leq R_t := \vartheta \sqrt{V l_t}, \qquad j < t.$$

the quadratic variation satisfies

$$\sum_{j=0}^{t-1} \mathbb{E}_j[X_{j+1}^2] \leq V \vartheta^2 t l_t \leq V \vartheta^2 s_t l_t =: \sigma_t^2.$$

Define $x_t := \frac{\vartheta^2}{4} \sqrt{s_t}\, l_t$. Since $V \leq \vartheta^2 / 128$ and $s_t \geq 4$, direct computation leads to

$$\frac{x_t^2}{2(\sigma_t^2 + R_t x_t / 3)} \geq 2 l_t.$$

Applying Freedman inequality in Lemma B.1 gives

$$\mathbb{P}(S_t \geq x_t) \leq e^{-2 l_t} = \left(\frac{\delta}{s_t}\right)^2.$$

Using union bound technique, we have

$$\mathbb{P}\left(\exists t \geq 1 :\ S_t \geq x_t\right) \leq \delta^2 \sum_{n=s_0}^{\infty} n^{-2} \leq \delta.$$

Therefore, with probablity at least $1 - \delta$, for any $t \geq 1$,

$$\boxed{\text{(III)} = \frac{S_t}{s_{t-1}} \leq \frac{x_t}{s_{t-1}} \leq \frac{\vartheta^2}{2} \frac{l_t}{\sqrt{s_t}},} \tag{28}$$

where we use $s_t / s_{t-1} \leq 2$ again. Plugging the upper bounds in equation 26-equation 28 into equation 25 , we have

$$U_{t\wedge\mathcal{T}_U} \leq \left(\frac{1}{4} + \frac{1}{4} + \frac{1}{2}\right) \vartheta^2 \frac{l_t}{\sqrt{s_t}} = \vartheta^2 \frac{l_t}{\sqrt{s_t}}.$$

Thus, if $\mathcal{T}_U < \infty$, let $t = \mathcal{T}_U$, then the above result $U_{t\wedge\mathcal{T}_U} \leq \vartheta^2 \frac{l_t}{\sqrt{s_t}}$ contradicts the definition of $\mathcal{T}_U$. Hence the event $\mathcal{T}_U = \infty$ holds with high probability at least $1 - \delta$. □

C Proof for Section 3.1: (II) Residual Potential Function and its Second-order Taylor Expansion

Recall that

$$W := W(w), \quad \nabla\Phi(w) = -W^{-1}\mathbf{1}, \quad H(w) := \nabla^2\Phi(w) = W^{-2}, \quad Q_t(w,\alpha) := \mathrm{Diag}(q_t(w,\alpha)).$$

Finally define

$$B(w,\alpha) := \nabla_w q_t(w,\alpha) = J + \alpha H(w). \tag{29}$$

Proposition C.1 (First-Order Gradient of the Residual Potential)**.** *For any $w \in \mathrm{ri}(W)$ and $\alpha > 0$, we have*

$$\nabla_w \Psi_t(w,\alpha) = 2B(w,\alpha)^\top W^2 q_t(w,\alpha) + 2W\big(q_t(w,\alpha) \odot q_t(w,\alpha)\big), \tag{30}$$

$$\partial_\alpha \Psi_t(w,\alpha) = 2\nabla\Phi(w)^\top W^2 q_t(w,\alpha) = -2w^\top q_t(w,\alpha). \tag{31}$$

Therefore, for any vector (Δ, Δ_α),

$$\langle \nabla\Psi_t(w,\alpha), \begin{pmatrix}\Delta\\ \Delta_\alpha\end{pmatrix}\rangle = 2\,\langle B(w,\alpha)\Delta + \Delta_\alpha \nabla\Phi(w), q_t(w,\alpha)\rangle_{*,w} + 2\sum_{i=1}^K \frac{\Delta_i}{w_i}\big(w_i q_{t,i}(w,\alpha)\big)^2. \tag{32}$$

Proof. By equation 4, for each j,

$$\frac{\partial \Psi_t}{\partial w_j} = 2\sum_{i=1}^K w_i^2 q_{t,i}\frac{\partial q_{t,i}}{\partial w_j} + 2w_j q_{t,j}^2.$$

Using $\nabla_w q_t = B(w,\alpha)$ gives equation 30. Since $\partial_\alpha q_t = \nabla\Phi(w)$,

$$\partial_\alpha \Psi_t = 2\nabla\Phi(w)^\top W^2 q_t.$$

From $W^2\nabla\Phi(w) = -w$ we obtain equation 31. Finally, take the inner product of the joint gradient with the direction (Δ, Δ_α) to obtain equation 32. □

Proposition C.2 (Hessian of the Residual Potential)**.** *Let $q = q_t(w,\alpha)$, $Q = Q_t(w,\alpha)$, and $B = B(w,\alpha)$. Then*

$$\nabla^2_{(w,\alpha)}\Psi_t(w,\alpha) = \begin{pmatrix}\mathcal{H}_{ww}(w,\alpha) & -2(B^\top w + q)\\ -2(w^\top B + q^\top) & 2K\end{pmatrix}, \tag{33}$$

where

$$\mathcal{H}_{ww}(w,\alpha) = 2B^\top W^2 B + 4B^\top WQ + 4WQB + 2Q^2 - 4\alpha W^{-1}Q. \tag{34}$$

Moreover,

$$\nabla^2_{w\alpha}\Psi_t(w,\alpha) = 2(g_t - c) \in \mathcal{G}. \tag{35}$$

Proof. Starting from equation 30, differentiate with respect to w.

$$d(W^2) = 2W\,\mathrm{Diag}(dw), \qquad dq = B\,dw.$$

Thus

$$d\big(2B^\top W^2 q\big) = 2(dB)^\top W^2 q + 4B^\top WQ\,dw + 2B^\top W^2 B\,dw,$$
$$d\big(2W(q\odot q)\big) = 2Q^2\,dw + 4WQB\,dw.$$

Because F is affine, the second derivative of q_i with respect to w comes only from $\alpha(\nabla\Phi(w))_i = -\alpha/w_i$, namely

$$\nabla^2_w q_i(w,\alpha) = -\frac{2\alpha}{w_i^3}e_i e_i^\top.$$

Therefore

$$2\sum_{i=1}^K (W^2 q)_i \nabla^2_w q_i = -4\alpha W^{-1}Q,$$

Combining the above results yields equation 34.

Starting from equation 30, differentiate with respect to α and use $\partial_\alpha B = H = W^{-2}$, $W\nabla\Phi(w) = -\mathbf{1}$, we obtain

$$\begin{aligned}\nabla^2_{w\alpha}\Psi_t &= 2HW^2q + 2B^\top W^2\nabla\Phi(w) + 4WQ\nabla\Phi(w)\\ &= 2q - 2B^\top w - 4q = -2(B^\top w + q).\end{aligned}$$

Also, from Eq.(31), we obtain $\partial^2_{\alpha\alpha}\Psi_t = -2w^\top\partial_\alpha q_t(w,\alpha) = -2w^\top\nabla\Psi(w) = 2K$.

Finally, from $q = Jw + c + \alpha\nabla\Phi(w) - g_t$, $B = J + \alpha H(w)$, $J^\top = -J$ and $H(w)w = -\nabla\Phi(w)$, we have

$$\begin{aligned}B^\top w + q &= (J^\top + \alpha H)w + Jw + c + \alpha\nabla\Phi(w) - g_t\\ &= -Jw - \alpha\nabla\Phi(w) + Jw + c + \alpha\nabla\Phi(w) - g_t\\ &= c - g_t.\end{aligned}$$

This proves equation 35. □

Remark C.3 (Exact Cancellation of the Mixed Second-Order Term along Feasible Directions)**.** Because $\Delta_t = w_{t+1} - w_t \in \mathcal{W}_0$. and $g_t - c \in \mathcal{G}$ is independent of any $w \in \mathcal{W}_0$, we have $\Delta_t^\top(g_t - c) = 0$. Therefore, for any point on the path $(w_{t,s}, \alpha_{t,s})$,

$$2\Delta_{\alpha,t}\Delta_t^\top\nabla^2_{w\alpha}\Psi_t(w_{t,s},\alpha_{t,s}) = 0. \tag{36}$$

D Proof for Section 3.2: Key Properties of Adaptively Regularized Mirror Descent with Log-Barrier Regularization

Given $x \in \mathrm{ri}(\Delta_d)$ and $\rho \in (0,1]$, define the reference $\mathsf{R}_\rho(x) \in \mathrm{ri}(\Delta_d)$: find $\beta \in [0,d]$ such that

$$\sum_{i=1}^d \frac{1}{(1-\rho)/x_i + \beta} = 1,$$

and let

$$\mathsf{R}_\rho(x)_i = \frac{1}{(1-\rho)/x_i + \beta}.$$

Apply this construction separately to the simplex blocks of the two players to obtain the fixed reference of the k-th epoch

$$\widehat{\sigma}^k = \mathsf{R}_{\rho_k}(\sigma^k).$$

Lemma D.1 (Existence and Uniqueness of the Reference Strategy)**.** *For any $x \in \mathrm{ri}(\Delta_d)$ and $\rho \in (0,1]$, the above β exists uniquely and satisfies $0 \le \beta \le d$. Therefore, $\mathsf{R}_\rho(x)$ is a valid strictly interior mixed strategy.*

Proof. Let $c = 1 - \rho \in [0,1)$, and define

$$h(\beta) = \sum_{i=1}^d \frac{1}{c/x_i + \beta}.$$

If $c > 0$, then $h(0) = \frac{1}{c} > 1$, and $h(d) = \sum_i \frac{x_i}{c + dx_i} < \sum_i \frac{1}{d} = 1$. If $c = 0$, then $h(d) = 1$. Moreover,

$$h'(\beta) = -\sum_i (c/x_i + \beta)^{-2} < 0.$$

Thus, there exists a unique $\beta \in (0,d]$ such that $h(\beta) = 1$. By definition, all coordinates of the reference are strictly positive and sum to 1.

The upper bound can also be obtained directly. From

$$\frac{1}{\widehat{\sigma}_i} = \frac{c}{x_i} + \beta$$

Multiplying both sides by $\widehat{\sigma}_i$ and summing gives:

$$d = c \sum_i \frac{\widehat{\sigma}_i}{x_i} + \beta \sum_i \widehat{\sigma}_i,$$

hence $\beta = d - c\sum_i \frac{\widehat{\sigma}_i}{x_i} \le d$. □

Lemma D.2 (Equivalent Transformation of the Regularized Payoff at the Epoch Start)**.** *At the start of the k-th epoch, $w_{T_k} = \sigma^k$, there exists $q_k \in \mathcal{G}$ such that*

$$\lambda \left(\nabla\Phi(\sigma^k) - \nabla\Phi(\widehat{\sigma}^k)\right) = \alpha_{T_k} \nabla\Phi(\sigma^k) + q_k.$$

Proof. On a d-dimensional block,

$$\frac{1}{\widehat{\sigma}_i^k} = \left(1 - \frac{\alpha_{T_k}}{\lambda}\right) \frac{1}{\sigma_i^k} + \beta_k.$$

Since $\nabla\Phi(x)_i = -1/x_i$,

$$\nabla\Phi(\widehat{\sigma}^k) = \left(1 - \frac{\alpha_{T_k}}{\lambda}\right) \nabla\Phi(\sigma^k) - \beta_k \mathbf{1}.$$

therefore

$$\lambda \left(\nabla\Phi(\sigma^k) - \nabla\Phi(\widehat{\sigma}^k)\right) = \alpha_{T_k} \nabla\Phi(\sigma^k) + \lambda\beta_k \mathbf{1}.$$

the blockwise-constant terms form $q_k \in \mathcal{G}$. □

Lemma D.3 (Equivalent Transformation of the Regularized Payoff at Any Intra-Epoch Step)**.** *For $t \in [T_k, T_{k+1})$, there exists $q_k \in \mathcal{G}$ such that*

$$\lambda(\nabla\Phi(w_t) - \nabla\Phi(\widehat{\sigma}^k)) = \alpha_t \nabla\Phi(w_t) + e_t + q_k,$$

where

$$\boxed{e_t = (\lambda - \alpha_t)\big(\nabla\Phi(w_t) - \nabla\Phi(w_{T_k})\big) + (\alpha_{T_k} - \alpha_t)\nabla\Phi(w_{T_k}).}$$

In particular $e_{T_k} = 0$.

Proof. By the blockwise formula in Lemma D.2,

$$\nabla\Phi(\widehat{\sigma}^k) = \left(1 - \frac{\alpha_{T_k}}{\lambda}\right) \nabla\Phi(w_{T_k}) - q_k/\lambda.$$

Substituting into the left-hand side gives:

$$\begin{aligned}
&\lambda\nabla\Phi(w_t) - \lambda\nabla\Phi(\widehat{\sigma}^k) \\
&= \lambda\nabla\Phi(w_t) - (\lambda - \alpha_{T_k})\nabla\Phi(w_{T_k}) + q_k \\
&= \alpha_t\nabla\Phi(w_t) + (\lambda - \alpha_t)\big(\nabla\Phi(w_t) - \nabla\Phi(w_{T_k})\big) \\
&\quad + (\alpha_{T_k} - \alpha_t)\nabla\Phi(w_{T_k}) + q_k.
\end{aligned}$$

□

Remark D.4 (The Exact Update Direction of Algorithm)**.** Thus, modulo the normal space $\mathcal{G}$, the update direction of mirror descent step in Algorithm 1 equals

$$\widehat{F}_t + \alpha_t \nabla\Phi(w_t) + e_t,$$

where e_t is an $\mathcal{F}_t$-measurable predictable bias for any $t \in [T_k, T_k + H_k]$, with $e_{T_k} = 0$. In particular, when $H_k = 1$, $e_{T_k} = e_{T_k+1} = 0$, then our algorithm's update direction degrades exactly to the direction $\widehat{F}_t + \alpha_t\nabla\Phi(w_t)$ as that of (Fiegel et al., 2026, Algorithm 1), and our proof results include the last-iterate convergence of (Fiegel et al., 2026, Algorithm 1) as a special case.

D.1 Primal Norm Bounds for the Primal Update Vector $\|w_{t+1} - w_t\|_{w_t}$

Proof of Lemma 3.7. Let $\widetilde{w}$ be the unconstrained dual step:

$$\widetilde{w} = \arg\min_{u \in \mathbb{R}^K} \{D_\Phi(u, w) + \eta \langle \xi, u \rangle\}.$$

The coordinatewise first-order condition is

$$-\frac{1}{\widetilde{w}_i} + \frac{1}{w_i} + \eta\xi_i = 0,$$

hence

$$\eta\xi_i w_i = \frac{w_i}{\widetilde{w}_i} - 1. \tag{37}$$

If $\eta \|\xi\|_{*,w} \le 1/4$, then

$$\eta|w_i\xi_i| \le \eta \left(\sum_{j=1}^K w_j^2 \xi_j^2\right)^{1/2} \le \frac{1}{4}.$$

Combining equation 37,

$$\left|\frac{w_i}{\widetilde{w}_i} - 1\right| \le \frac{1}{4},$$

Therefore

$$\frac{3}{4} \le \frac{w_i}{\widetilde{w}_i} \le \frac{5}{4}.$$

By the Bregman Pythagorean inequality,

$$D_\Phi(w, \widetilde{w}) \ge D_\Phi(w, w^+) + D_\Phi(w^+, \widetilde{w}) \ge D_\Phi(w, w^+).$$

On the other hand, when $r \ge \frac{1}{\sqrt{2}}$, by Taylor's theorem there exists $\theta \ge \frac{1}{\sqrt{2}}$ such that the function $h(r) = r - 1 - \log r = h(1) + h'(1)(r-1) + \frac{h''(\theta)}{2}(r-1)^2 \le (r-1)^2$, thus when $\frac{w_i}{\widetilde{w}_i} \ge \frac{3}{4} \ge \frac{1}{\sqrt{2}}$, by applying equation 37 we have

$$D_\Phi(w, w^+) \le D_\Phi(w, \widetilde{w}) = \sum_{i=1}^K h\left(\frac{w_i}{\widetilde{w}_i}\right) \le \sum_{i=1}^K \eta^2 \xi_i^2 w_i^2 = \eta^2 \|\xi\|_{*,w}^2 \le \frac{1}{16}. \tag{38}$$

Therefore $D_\Phi(w, w^+) \le \frac{1}{16}$, and for any $i \in [K]$, $h\left(\frac{w_i}{w_i^+}\right) \le \frac{1}{16}$ and so

$$0.686 \le \frac{w_i}{w_i^+} \le 1.376.$$

Then, since $\frac{h(r)}{(1/r-1)^2}$ is an increasing function for $r \ge 0.686$, there exists the numerical lower bound

$$\frac{h(w_i/w_i^+)}{(w_i^+/w_i - 1)^2} \ge 0.3.$$

Therefore

$$\left\|w^+ - w\right\|_w^2 = \sum_{i=1}^K \left(\frac{w_i^+}{w_i} - 1\right)^2 \le \frac{10}{3} \sum_{i=1}^K h\left(\frac{w_i}{w_i^+}\right) \le \frac{10}{3}\eta^2 \|\xi\|_{*,w}^2 \le 4\eta^2 \|\xi\|_{*,w}^2.$$

Taking square roots gives

$$\left\|w^+ - w\right\|_w \le 2\eta \|\xi\|_{*,w}. \qquad \square$$

Fix an epoch and denote for shortness

$$s = T_k, \qquad t \in [s, s + H_k].$$

Define $R_t := \max_{i \in [K]} \left|\frac{w_{t,i}}{w_{s,i}} - 1\right|$.

Proposition D.5 (Intra-Epoch Relative-Coordinate Variation)**.** *If* $\eta_s \le \frac{1}{6M(H_k+1)}$, *then*

$$R_t \le \frac{1}{2}, \qquad \forall t \in [s, s+H_k].$$

Proof. We use induction. At the initial time $t = s = T_k$,

$$R_s = 0 \le \frac{1}{2}.$$

Assume $R_t \le 1/2$. Then

$$\|\nabla\Phi(w_t) - \nabla\Phi(w_s)\|_{*,w_t}^2 = \sum_{i=1}^{K}\left(\frac{w_{t,i}}{w_{s,i}} - 1\right)^2 \le KR_t^2, \tag{39}$$

and

$$\|\nabla\Phi(w_s)\|_{*,w_t}^2 = \sum_{i=1}^{K}\left(\frac{w_{t,i}}{w_{s,i}}\right)^2 \le \frac{9}{4}K. \tag{40}$$

By the effective-update-direction decomposition of mirror descent,

$$\begin{aligned}\|\xi_t\|_{*,w_t} &\le \left\|\widehat{F}_t\right\|_{*,w_t} + \alpha_t \|\nabla\Phi(w_t)\|_{*,w_t} + |\lambda - \alpha_t| \|\nabla\Phi(w_t) - \nabla\Phi(w_s)\|_{*,w_t} \\ &\quad + |\alpha_s - \alpha_t| \|\nabla\Phi(w_s)\|_{*,w_t} \\ &\le 2 + \alpha_t\sqrt{K} + |\lambda - \alpha_t|\sqrt{K}\, R_t + |\alpha_s - \alpha_t|\frac{3}{2}\sqrt{K} \\ &\le 2 + 4a\sqrt{K} + \frac{\lambda\sqrt{K}}{2} = M. \end{aligned} \tag{41}$$

where

$$\alpha_t = a\log\left(\frac{t+T_{\text{off}}}{\delta}\right)(t+T_{\text{off}})^{-1/4}, \qquad \eta_t = \eta(t+T_{\text{off}})^{-3/4}.$$

Since

$$\eta_t M \le \eta_s M \le \frac{1}{4},$$

Lemma 3.7 gives

$$\|w_{t+1} - w_t\|_{w_t} \le 2\eta_t \|\xi_t\|_{*,w_t} \le 2\eta_t M. \tag{42}$$

hence for any $i \in [K]$,

$$\left|\frac{w_{t+1,i}}{w_{t,i}} - 1\right| \le 2\eta_t M. \tag{43}$$

Let

$$\gamma_{t,i} := \frac{w_{t,i}}{w_{s,i}}, \qquad \mu_{t,i} := \frac{w_{t+1,i}}{w_{t,i}} - 1.$$

then

$$\begin{aligned} R_{t+1} &= \max_i |\gamma_{t,i}(1+\mu_{t,i}) - 1| \\ &\le R_t + (R_t+1)\max_i |\mu_{t,i}| \\ &\le R_t + (R_t+1)2\eta_t M, \end{aligned} \tag{44}$$

where the last inequality follows from equation 43. When $R_t \le 1/2$,

$$\begin{aligned} R_{t+1} &\le R_t + (R_t+1)2\eta_t M \\ &\le R_t + 3\eta_s M. \end{aligned} \tag{45}$$

Iterating the above inequality, and using the definition $\eta_s \le 1/6M(H_k+1)$, we obtain

$$R_{t+1} \le R_s + 3\eta_s M(t+1-s) \le 3\eta_s M(H_k+1) \le \frac{1}{2}.$$

This closes the induction. □

Proposition D.6 (Refined Intra-Epoch Relative-Coordinate Variation Bound via Discrete Gronwall)**.** *If the step size satisfies $\eta_{T_k} \le \frac{1}{6M(H_k+1)}$ for any k, then $\forall t \in [T_k, T_k + H_k]$,*

$$R_t := \max_{i\in[K]} \left| \frac{w_{t,i}}{w_{T_k,i}} - 1 \right| \le 3M\eta_{T_k} H_k. \tag{46}$$

Furthermore, we have $\eta_t \|\xi_t\|_{*,w_t} \le 1/4$ and $\|w_{t+1} - w_t\|_{w_t} \le 2\eta_t \|\xi_t\|_{*,w_t}$.

Proof. By Proposition D.5 and equation 44, $R_t \le 1/2$, and

$$R_{t+1} \le R_t + (R_t + 1)2\eta_t M = (1 + 2\eta_t M)R_t + 2\eta_t M.$$

The discrete Gronwall inequality gives

$$\begin{aligned} R_t &\le R_s \prod_{j=s}^{t-1}(1 + 2\eta_j M) + \sum_{k=s}^{t-1} 2\eta_k M \prod_{j=k+1}^{t-1}(1 + 2\eta_j M) \\ &\le 2M \sum_{k=s}^{t-1} \eta_k \exp\left(2M \sum_{j=k+1}^{t-1} \eta_j \right). \end{aligned} \tag{47}$$

Since $\eta_j \le \eta_s$ and $t - s \le H_k$,

$$2M \sum_{j=k+1}^{t-1} \eta_j \le 2MH_k\eta_s \le \frac{1}{3}.$$

Therefore

$$R_t \le 2Me^{1/3}\eta_s H_k \le 3M\eta_s H_k.$$

The second part directly follows from Lemma 3.7 due to the definition of η_t. □

Theorem D.7. *If $\eta_{T_k} \le \frac{1}{6M(H_k+1)}$, then, for all $t \in [T_k, T_{k+1})$,*

$$\|e_t\|_{*,w_t} \le 3M\eta_{T_k} H_k \lambda\sqrt{K} + \frac{3}{2}\sqrt{K}|\alpha_{T_k} - \alpha_t| \tag{48}$$

$$\le 3M\eta_{T_k} H_k \lambda\sqrt{K} + \frac{3}{8}\sqrt{K}\, H_k \frac{\alpha_{T_k}}{T_k + T_{\text{off}}}. \tag{49}$$

Consequently

$$\|e_t\|_{*,w_t} \le O\left(\lambda H_k \eta_{T_k} + H_k \frac{\alpha_{T_k}}{T_k + T_{\text{off}}} \right).$$

Proof. The original manuscript uses

$$e_t = (\lambda - \alpha_t)\big(\nabla\Phi(w_t) - \nabla\Phi(w_{T_k})\big) + (\alpha_{T_k} - \alpha_t)\nabla\Phi(w_{T_k}). \tag{50}$$

therefore

$$\begin{aligned} \|e_t\|_{*,w_t} \le{}& |\lambda - \alpha_t| \, \|\nabla\Phi(w_t) - \nabla\Phi(w_{T_k})\|_{*,w_t} \\ &+ |\alpha_{T_k} - \alpha_t| \, \|\nabla\Phi(w_{T_k})\|_{*,w_t}. \end{aligned}$$

From equation 39, equation 40 and Proposition 3.8,

$$\|e_t\|_{*,w_t} \le \lambda\sqrt{K}\, R_t + \frac{3}{2}\sqrt{K}|\alpha_{T_k} - \alpha_t|,$$

which yields equation 48.

Let $f(s) = a\log(s/\delta)s^{-1/4}$. then we have

$$|f'(s)| = as^{-5/4}\left|1 - \frac{1}{4}\log(s/\delta)\right| \le \frac{1}{4}\frac{f(s)}{s}, \qquad \text{provided } \log(s/\delta) \ge 4.$$

By the mean value theorem, there exist $\theta \in (0,1)$ such that

$$\begin{aligned}|\alpha_{T_k} - \alpha_t| &= |f(T_k + T_{\text{off}}) - f(t + T_{\text{off}})| \\ &= |f'(\theta(t + T_{\text{off}}) + (1-\theta)(T_k + T_{\text{off}}))|\,|t - T_k| \\ &\le \frac{1}{4} H_k \frac{\alpha_{T_k}}{T_k + T_{\text{off}}},\end{aligned}$$

Thus we obtain equation 49. □

Proposition D.8. *Under the conditions of Theorem D.7, further assume that* $T_{\text{off}} \ge \max(1, \frac{\lambda^2\eta^2}{a^2} 32HM\sqrt{K}) * 32HM\sqrt{K}$*, then we can set* $c_e = \frac{1}{2}$ *and have* $E_t = \|e_t\|_{*,w_t} \le c_e \alpha_t$ *for all* t*.*

Proof. Let

$$s_t := t + T_{\text{off}}, \qquad s_k := s_{T_k} = T_k + T_{\text{off}}.$$

When

$$t \in [T_k, T_k + H),$$

we have

$$s_k \le s_t \le s_k + H.$$

If $T_{\text{off}} \ge 2H$, then

$$1 \le \frac{s_t}{s_k} \le 1 + \frac{H}{T_{\text{off}}} \le \frac{3}{2}.$$

By the structure of Theorem D.7,

$$E_t \le C_E H \left[\lambda\eta s_k^{-3/4} + a \log(s_k/\delta) s_k^{-5/4}\right],$$

where $C_E = 4M\sqrt{K}$. Because

$$s_k^{-3/4} \le \left(1 + \frac{H}{T_{\text{off}}}\right)^{3/4} s_t^{-3/4},$$

and

$$\log(s_k/\delta) s_k^{-5/4} \le \left(1 + \frac{H}{T_{\text{off}}}\right)^{5/4} \log(s_t/\delta) s_t^{-5/4},$$

hence we can define

$$D_E := C_E \left(1 + \frac{H}{T_{\text{off}}}\right)^{5/4},$$

and obtain

$$\boxed{E_t \le D_E H \left[\lambda\eta s_t^{-3/4} + a \log(s_t/\delta) s_t^{-5/4}\right].} \tag{51}$$

Since

$$\alpha_t = a \log(s_t/\delta) s_t^{-1/4},$$

so

$$\begin{aligned}\frac{E_t}{\alpha_t} &\le D_E H \left[\frac{\lambda\eta}{a \log(s_t/\delta)\sqrt{s_t}} + \frac{1}{s_t}\right] \\ &\le D_E H \left[\frac{\lambda\eta}{a \log(s_0/\delta)\sqrt{s_0}} + \frac{1}{s_0}\right] \le \frac{1}{2},\end{aligned}$$

where the last inequality holds due to the condition $T_{\text{off}} = s_0 \ge \max(1, \frac{\lambda^2\eta^2}{a^2} 32HM\sqrt{K}) * 32HM\sqrt{K}$, and the fact that $D_E \le 8M\sqrt{K}$. Therefore, there exits $c_e \ge \frac{1}{2}$ such that $E_t \le c_e \alpha_t$. □

Proof of Lemma 3.3. Denote

$$D_\mu := \mathrm{Diag}(\mu), \qquad D_\nu := \mathrm{Diag}(\nu),$$

then the diagonal matrix

$$W(w) = \begin{pmatrix} D_\mu & 0 \\ 0 & D_\nu \end{pmatrix}.$$

Therefore

$$\begin{aligned} W(w)JW(w) &= \begin{pmatrix} D_\mu & 0 \\ 0 & D_\nu \end{pmatrix} \begin{pmatrix} 0 & A \\ -A^\top & 0 \end{pmatrix} \begin{pmatrix} D_\mu & 0 \\ 0 & D_\nu \end{pmatrix} \\ &= \begin{pmatrix} 0 & D_\mu A D_\nu \\ -D_\nu A^\top D_\mu & 0 \end{pmatrix}. \end{aligned}$$

Let

$$B := D_\mu A D_\nu.$$

then

$$W(w)JW(w) = \begin{pmatrix} 0 & B \\ -B^\top & 0 \end{pmatrix}.$$

Note that

$$\begin{pmatrix} 0 & B \\ -B^\top & 0 \end{pmatrix}^\top \begin{pmatrix} 0 & B \\ -B^\top & 0 \end{pmatrix} = \begin{pmatrix} BB^\top & 0 \\ 0 & B^\top B \end{pmatrix}.$$

therefore

$$\begin{aligned} \left\|W(w)JW(w)\right\|_{\mathrm{op}}^2 &= \max\left\{\lambda_{\max}(BB^\top), \lambda_{\max}(B^\top B)\right\} \\ &= \|B\|_{\mathrm{op}}^2. \end{aligned}$$

Thus

$$\left\|W(w)JW(w)\right\|_{\mathrm{op}} = \|D_\mu A D_\nu\|_{\mathrm{op}}.$$

Next, use the fact that the operator norm is at most the Frobenius norm:

$$\begin{aligned} \|D_\mu A D_\nu\|_{\mathrm{op}}^2 &\le \|D_\mu A D_\nu\|_{\mathrm{F}}^2 \\ &= \sum_{i=1}^m \sum_{j=1}^n \mu_i^2 A_{ij}^2 \nu_j^2 \\ &\le \|A\|_{\max}^2 \sum_{i=1}^m \sum_{j=1}^n \mu_i^2 \nu_j^2 \\ &\le \|A\|_{\max}^2 \left(\sum_{i=1}^m \mu_i^2\right)\left(\sum_{j=1}^n \nu_j^2\right) \le \|A\|_{\max}^2. \end{aligned}$$

When $A_{ij} \in [0,1]$,

$$\|A\|_{\max} \le 1,$$

therefore

$$\boxed{\left\|W(w)JW(w)\right\|_{\mathrm{op}} \le 1.}$$ □

D.2 Dual Norm Bounds for the Update Residual Term $\left\|r_t^{\mathrm{md}}\right\|_{w_t}$

Lemma D.9. *Assume that* $\left\|\widehat{F}_t\right\|_{*,w_t} \le \sigma$, *define* $\bar{\sigma} \triangleq \sigma + \alpha_0\sqrt{K}$, *then we have* $d_t = \|p_t\|_{*,w_t}^2 \le \bar{\sigma}^2$.

Proof. By the definition of the dual projected residual,

$$g_t \in \arg\min_{g \in \mathcal{G}} \|F(w_t) + \alpha_t \nabla\Phi(w_t) - g\|_{*,w_t},$$

it follows that

$$p_t = \Pi_t\big(F(w_t) + \alpha_t \nabla\Phi(w_t)\big) = F(w_t) + \alpha_t \nabla\Phi(w_t) - g_t.$$

Therefore,

$$\begin{aligned} d_t &= \|p_t\|_{*,w_t}^2 \\ &= \big\|\Pi_t\big(F(w_t) + \alpha_t \nabla\Phi(w_t)\big)\big\|_{*,w_t}^2 \\ &\le \|F(w_t) + \alpha_t \nabla\Phi(w_t)\|_{*,w_t}^2 \\ &= \Big\|\mathbb{E}_t[\widehat{F}_t] + \alpha_t \nabla\Phi(w_t)\Big\|_{*,w_t}^2 \\ &\le \mathbb{E}_t\left[\Big\|\widehat{F}_t + \alpha_t \nabla\Phi(w_t)\Big\|_{*,w_t}^2\right] \\ &\le (\sigma + \alpha_t\sqrt{K})^2 \triangleq \bar{\sigma}^2. \end{aligned}$$

where the first inequality due to the non-expansiveness of the projection operator, the second inequality follows from the convexity of the function

$$x \longmapsto \|x\|_{*,w_t}^2$$

and the conditional Jensen inequality. □

Let

$$\varepsilon_t := \widehat{F}_t - F(w_t), \qquad \mathbb{E}_t[\varepsilon_t] = 0, \qquad \zeta_t := \Pi_t \varepsilon_t.$$

Further let

$$\widehat{p}_t := \Pi_t\big(\widehat{F}_t + \alpha_t \nabla\Phi(w_t) + e_t\big).$$

Thus

$$\widehat{p}_t = \Pi_t\big(\varepsilon_t + F(w_t) + \alpha_t \nabla\Phi(w_t) + e_t\big) = \zeta_t + p_t + \overline{e}_t,$$

Define the log-barrier Taylor remainder

$$r_t^{\Phi} := \nabla\Phi(w_{t+1}) - \nabla\Phi(w_t) - H_t\Delta_t.$$

Then we can write the update residual as $r_t^{\mathrm{md}} = -H_t^{-1}\Pi_t r_t^{\Phi}$ for shortness.

Proof of Proposition 3.9. The first-order optimality condition of the mirror update shows that there exists a blockwise-constant vector $q_t \in \mathcal{G}$ such that

$$\nabla\Phi(w_{t+1}) - \nabla\Phi(w_t) + \eta_t(\xi_t - q_t) = 0. \tag{52}$$

Apply the projection Π_t to equation 52,

$$\Pi_t\left[\nabla\Phi(w_{t+1}) - \nabla\Phi(w_t)\right] + \eta_t \Pi_t \xi_t = 0. \tag{53}$$

Since $\Delta_t \in \mathcal{W}_0$, we have for any $g \in \mathcal{G}$,

$$\langle H_t\Delta_t, g\rangle_{*,w_t} = (H_t\Delta_t)^{\top} H_t^{-1} g = \Delta_t^{\top} g = 0.$$

therefore

$$H_t\Delta_t \in \mathcal{G}^{\perp,*,w_t}, \qquad \Pi_t(H_t\Delta_t) = H_t\Delta_t.$$

Substitute $\nabla\Phi(w_{t+1}) - \nabla\Phi(w_t) = H_t\Delta_t + r_t^{\Phi}$ into equation 53 yields

$$0 = H_t\Delta_t + \Pi_t r_t^{\Phi} + \eta_t \widehat{p}_t.$$

Therefore

$$\Delta_t = -\eta_t H_t^{-1}\widehat{p}_t - H_t^{-1}\Pi_t r_t^{\Phi},$$

.

Moreover,

$$\left\|r_t^{\mathrm{md}}\right\|_{w_t} = \left\|-H_t^{-1}\Pi_t r_t^{\Phi}\right\|_{w_t} = \left\|\Pi_t r_t^{\Phi}\right\|_{*,w_t} \le \left\|r_t^{\Phi}\right\|_{*,w_t}.$$

Coordinatewise,

$$r_{t,i}^{\Phi} = -\frac{1}{w_{t+1,i}} + \frac{1}{w_{t,i}} - \frac{\Delta_{t,i}}{w_{t,i}^2} = -\frac{\Delta_{t,i}^2}{w_{t,i}^2 w_{t+1,i}}.$$

Using

$$\frac{1}{2} w_{t,i} \le w_{t+1,i} \le 2 w_{t,i},$$

the dual-norm of log-barrier Taylor remainder satisfies

$$\left\| r_t^{\Phi} \right\|_{*,w_t}^2 = \sum_i w_{t,i}^2 (r_{t,i}^{\Phi})^2 \le 4 \sum_i \frac{\Delta_{t,i}^4}{w_{t,i}^4} \le 4 \left(\sum_i \frac{\Delta_{t,i}^2}{w_{t,i}^2} \right)^2 = 4 \left\| w_{t+1} - w_t \right\|_{w_t}^4.$$

so

$$\left\| r_t^{\Phi} \right\|_{*,w_t} \le 2 \left\| w_{t+1} - w_t \right\|_{w_t}^2.$$

Recalling Proposition 3.8 finishes the whole proof. □

E Proof for Section 3.3: (I) Upper Bounds for the First-Order and Second-Order Taylor Approximations

Define

$$\mathcal{B}_t := 2\sqrt{d_t} + (2 + \alpha_t) \left\| \Delta_t \right\|_{w_t} + \sqrt{K} \left| \Delta_{\alpha,t} \right|. \tag{54}$$

It is not difficult to see that $\mathcal{B}_t \le 2\bar{\sigma} + \frac{2+\alpha_0}{2} + \sqrt{K}\alpha_0 \triangleq \bar{\mathcal{B}}$ is bounded.

Lemma E.1 (Control of Local Quantities along the Taylor Path)**.** *Under the conditions of Proposition D.5 (i.e. small step size $\eta_{T_k} \le \frac{1}{6M(H_k+1)}$,). Then, for all $s \in [0,1]$,*

$$\left\| \Delta_t \right\|_{w_{t,s}} \le 2 \left\| \Delta_t \right\|_{w_t}, \tag{55}$$

$$\left\| q_t(w_{t,s}, \alpha_{t,s}) \right\|_{*,w_{t,s}} \le \mathcal{B}_t. \tag{56}$$

Proof. From Lemma 3.7 and Proposition D.5, we have $\left\| \Delta_t \right\|_{w_t} \le \frac{1}{2}$, and hence, for every coordinate,

$$\left| \frac{\Delta_{t,i}}{w_{t,i}} \right| \le \frac{1}{2}.$$

Therefore

$$\frac{1}{2} w_{t,i} \le w_{t,s,i} = w_{t,i} + s\Delta_{t,i} \le \frac{3}{2} w_{t,i},$$

which yields equation 55.

Next, using $q_t(w_t, \alpha_t) = p_t$ and $\alpha_{t,s} = \alpha_t + s\Delta_{\alpha,t}$, we have the exact decomposition

$$q_t(w_{t,s}, \alpha_{t,s}) - p_t = sJ\Delta_t + \alpha_t \big(\nabla\Phi(w_{t,s}) - \nabla\Phi(w_t) \big) + s\Delta_{\alpha,t} \nabla\Phi(w_{t,s}).$$

Bounding each term in the above expression gives:

$$\left\| p_t \right\|_{*,w_{t,s}} \le \frac{3}{2} \left\| p_t \right\|_{*,w_t} \le 2\sqrt{d_t}.$$

Applying Lemma 3.3, we have

$$\left\| sJ\Delta_t \right\|_{*,w_{t,s}} = s \left\| W_{t,s} J W_{t,s} W_{t,s}^{-1} \Delta_t \right\|_2 \le \left\| W_{t,s} J W_{t,s} \right\|_{op} \left\| W_{t,s}^{-1} \Delta_t \right\|_2 \le \left\| \Delta_t \right\|_{w_{t,s}} \le 2 \left\| \Delta_t \right\|_{w_t}.$$

For the barrier-gradient difference, coordinatewise

$$w_{t,s,i} \left(-\frac{1}{w_{t,s,i}} + \frac{1}{w_{t,i}} \right) = s \frac{\Delta_{t,i}}{w_{t,i}},$$

so

$$\left\| \nabla\Phi(w_{t,s}) - \nabla\Phi(w_t) \right\|_{*,w_{t,s}} = s \left\| \Delta_t \right\|_{w_t} \le \left\| \Delta_t \right\|_{w_t}.$$

At the same time

$$\left\| s\Delta_{\alpha,t} \nabla\Phi(w_{t,s}) \right\|_{*,w_{t,s}} \le \sqrt{K} \left| \Delta_{\alpha,t} \right|.$$

Combining the above results, we obtain equation 56. □

At the current point, denote

$$u_t := H_t^{-1/2} p_t, \qquad v_t := H_t^{-1/2} \bar{e}_t, \qquad z_t := H_t^{-1/2} \zeta_t, \qquad r_t^{\mathrm{w}} := H_t^{1/2} r_t^{\mathrm{md}}.$$

By Proposition 3.9,

$$H_t^{1/2} \Delta_t = -\eta_t (u_t + v_t + z_t) + r_t^{\mathrm{w}}. \tag{57}$$

Define

$$C_t := H_t^{-1/2} J^\top H_t^{-1/2}, \qquad C_t^\top = -C_t, \qquad \|C_t\|_{\mathrm{op}} \le \max_{i,j} |A_{i,j}| = 1.$$

Proof of Proposition 3.5. In equation 32, take $(w, \alpha) = (w_t, \alpha_t)$, and use $q_t(w_t, \alpha_t) = p_t$, yielding the first-order term

$$\langle \nabla \Psi_t(w_t, \alpha_t), \delta_t \rangle = \underbrace{2 \langle (J + \alpha_t H_t) \Delta_t, p_t \rangle_{*, w_t}}_{\text{(I)}} + \underbrace{2 \Delta_{\alpha,t} \langle \nabla \Phi(w_t), p_t \rangle_{*, w_t}}_{\text{(II)}} + \underbrace{2 \sum_i \frac{\Delta_{t,i}}{w_{t,i}} u_{t,i}^2}_{\text{(III)}}. \tag{58}$$

From equation 57, the first term equals

$$\text{(I)} = -2\eta_t \langle u_t + v_t + z_t, (C_t + \alpha_t I) u_t \rangle + 2 \left\langle (J + \alpha_t H_t) r_t^{\mathrm{md}}, p_t \right\rangle_{*, w_t}.$$

Since C_t is skew-symmetric, $\langle u_t, C_t u_t \rangle = 0$. Applying the Cauchy–Schwarz inequality, $-\langle v_t, (C_t + \alpha_t I) u_t \rangle \le (1 + \alpha_t) \|v_t\| \|u_t\| \le (1 + \alpha_t) E_t \sqrt{d_t}$, and $\left\langle (J + \alpha_t H_t) r_t^{\mathrm{md}}, p_t \right\rangle_{*, w_t} \le \left\| (J + \alpha_t H_t) r_t^{\mathrm{md}} \right\|_{*, w_t} \|p_t\|_{*, w_t}$, we have

$$\boxed{\begin{aligned} \text{(I)} = 2 \langle (J + \alpha_t H_t) \Delta_t, p_t \rangle_{*, w_t} \le & -2\eta_t \alpha_t d_t + 2\eta_t E_t (L + \alpha_t) \sqrt{d_t} \\ & - 2\eta_t \langle z_t, (C_t + \alpha_t I) u_t \rangle + 2(1 + \alpha_t) \left\| r_t^{\mathrm{md}} \right\|_{w_t} \sqrt{d_t}, \end{aligned}} \tag{59}$$

where in the last inequality, we use Lemma 3.3 to obtain

$$\begin{aligned} \left\| (J + \alpha_t H_t) r_t^{\mathrm{md}} \right\|_{*, w_t} &= \left\| H_t^{-1/2} (J + \alpha_t H_t) H_t^{-1/2} H_t^{1/2} r_t^{\mathrm{md}} \right\|_2 \\ &\le \left\| H_t^{-1/2} J H_t^{-1/2} + \alpha_t I \right\|_{\mathrm{op}} \left\| H_t^{1/2} r_t^{\mathrm{md}} \right\|_2 \\ &\le (1 + \alpha_t) \left\| r_t^{\mathrm{md}} \right\|_{w_t}. \end{aligned}$$

Analogously, the second term in Eq.(58) satisfies

$$\boxed{\text{(II)} = 2 |\Delta_{\alpha,t}| \left| \langle \nabla \Phi(w_t), p_t \rangle_{*, w_t} \right| \le 2\sqrt{K} |\Delta_{\alpha,t}| \sqrt{d_t},} \tag{60}$$

Again by equation 57, the third term in Eq.(58) satisfies

$$\boxed{\begin{aligned} \text{(III)} = 2 \sum_i \frac{\Delta_{t,i}}{w_{t,i}} u_{t,i}^2 = & -2\eta_t \sum_i (u_{t,i} + v_{t,i} + z_{t,i}) u_{t,i}^2 + 2 \sum_i r_{t,i}^{\mathrm{w}} u_{t,i}^2 \\ & \le 2\eta_t d_t^{3/2} + 2\eta_t E_t d_t - 2\eta_t \sum_i z_{t,i} u_{t,i}^2 + 2 \left\| r_t^{\mathrm{md}} \right\|_{w_t} d_t. \end{aligned}} \tag{61}$$

Here we used

$$\left| \sum_i u_{t,i}^3 \right| \le \|u_t\|_2^3 = d_t^{3/2}, \qquad \|u_t \odot u_t\|_2 \le \|u_t\|_2^2 = d_t.$$

Substituting the above results back into equation 58, and then using equation 6, gives the first claim.

For the second claim, since z_t is projected noise with zero conditional mean, and u_t, C_t are both $\mathcal{F}_t$-measurable, hence $\mathbb{E}_t[\mathcal{M}_{t+1}^\Psi] = 0$. Finally,

$$\begin{aligned} |\mathcal{M}_{t+1}^\Psi| &\le 2\eta_t \|z_t\|_2 \left(\|(C_t + \alpha_t I) u_t\|_2 + \|u_t \odot u_t\|_2 \right) \\ &\le 2\eta_t \|z_t\|_2 \left[(1 + \alpha_t) \sqrt{d_t} + d_t \right]. \end{aligned}$$

Combining this with $\|z_t\|_2 = \|\zeta_t\|_{*, w_t} \le 3$ finishes the whole proof. □

Lemma E.2 (Spectral Upper Bound for the Whitened Hessian)**.** *Fix $s \in [0,1]$ and write, for short,*

$$w = w_{t,s}, \quad \alpha = \alpha_{t,s}, \quad W = \mathrm{Diag}(w), \quad H = H(w) = W^{-2},$$

and

$$q = q_t(w, \alpha), \quad Q = \mathrm{Diag}(q), \quad B = J + \alpha H.$$

Define

$$C := H^{-1/2} B H^{-1/2} = WBW, \qquad D := WQ = \mathrm{Diag}(w_i q_i). \tag{62}$$

then

$$\widehat{\mathcal{H}}_{ww} := H^{-1/2} \mathcal{H}_{ww}(w, \alpha) H^{-1/2} = 2C^\top C + 4(C^\top D^\top + DC) + 2D^2 - 4\alpha D. \tag{63}$$

If $\|D\|_{\mathrm{op}} \le b$, then

$$\lambda_{\max}(\widehat{\mathcal{H}}_{ww}) \le 2(1+\alpha^2) + 8b + 2b^2 + 4\alpha b. \tag{64}$$

Proof. Because $H^{-1/2} = W$, by equation 34, whitening term by term gives

$$\begin{aligned} W(2B^\top W^2 B)W &= 2C^\top C, \\ W(4B^\top WQ)W &= 4C^\top D^\top, \\ W(4WQB)W &= 4DC, \\ W(2Q^2)W &= 2D^2, \\ W(-4\alpha W^{-1} Q)W &= -4\alpha D, \end{aligned}$$

which yields equation 63.

To prove equation 64, notice that $C = WBW = S + \alpha I$,where $S = WJW$. From $J^\top = -J$ it follows that $S^\top = -S$, Therefore

$$CC^\top = -SS + \alpha^2 I, \qquad C^\top D^\top + DC = SD - DS + 2\alpha D.$$

Plug the above result into equation 63, we have $\widehat{\mathcal{H}}_{ww} = 2(-SS+\alpha^2 I) + 4(SD - DS) + 2D^2 + 4\alpha D$. Using Weyl's inequality, and the fact that $\|S\|_{\mathrm{op}} \le 1$ and $\|D\|_{\mathrm{op}} \le b$,we have

$$\begin{aligned} \lambda_{max}(\widehat{\mathcal{H}}_{ww}) \le& 2\lambda_{max}(-SS + \alpha^2 I) + 4\lambda_{max}(SD - DS) + 2\lambda_{max}(D^\top D) + 4\alpha\lambda_{max}(D) \\ \le& 2\left\|-SS + \alpha^2 I\right\|_{\mathrm{op}} + 4\left\|SD - DS\right\|_{\mathrm{op}} + 2\left\|D^\top D\right\|_{\mathrm{op}} + 4\alpha\left\|D\right\|_{\mathrm{op}} \\ \le& 2(1+\alpha^2) + 8b + 2b^2 + 4\alpha b. \end{aligned}$$

□

Proof of Proposition 3.6. Fix $s \in [0,1]$ and let

$$H_{t,s} := H(w_{t,s}), \qquad \widehat{\mathcal{H}}_{ww,t,s} := H_{t,s}^{-1/2} \mathcal{H}_{ww}(w_{t,s}, \alpha_{t,s}) H_{t,s}^{-1/2}.$$

Let $y_{t,s} := H_{t,s}^{1/2} \Delta_t$, we have

$$\Delta_t^\top \mathcal{H}_{ww}(w_{t,s}, \alpha_{t,s}) \Delta_t = y_{t,s}^\top \widehat{\mathcal{H}}_{ww,t,s} y_{t,s} \le \lambda_{\max}(\widehat{\mathcal{H}}_{ww,t,s}) \left\|\Delta_t\right\|_{w_{t,s}}^2. \tag{65}$$

Define the diagonal matrix

$$D_{t,s} := W_{t,s} Q_t(w_{t,s}, \alpha_{t,s}).$$

Then, by equation 56,

$$\left\|D_{t,s}\right\|_{\mathrm{op}} = \max_i |w_{t,s,i} q_{t,i}(w_{t,s}, \alpha_{t,s})| \le \mathcal{B}_t. \tag{66}$$

Applying Remark C.3, Lemma E.2 and equation 65-66,we have

$$\begin{aligned} \delta_t^\top \nabla^2 \Psi_t(w_{t,s}, \alpha_{t,s}) \delta_t =& \Delta_t^\top \mathcal{H}_{ww} \Delta_t + 2K(\Delta_{\alpha,t})^2 \\ \le& 4\lambda_{\max}(\widehat{\mathcal{H}}_{ww,t,s}) \left\|\Delta_t\right\|_{w_t}^2 + 2K(\Delta_{\alpha,t})^2 \qquad (67) \\ \le& 8\big[(1+\alpha_t^2) + 4\mathcal{B}_t + \mathcal{B}_t^2 + 2\alpha_t \mathcal{B}_t\big] \left\|\Delta_t\right\|_{w_t}^2 + 2K(\Delta_{\alpha,t})^2. \qquad (68) \end{aligned}$$

Plug the above result into the expression $\mathcal{R}_t^\Psi = \int_0^1 (1-s) \delta_t^\top \nabla^2 \Psi_t(w_{t,s}, \alpha_{t,s}) \delta_t \, \mathrm{d}s$ and use the fact that $\int_0^1 (1-s) \, \mathrm{d}s = 1/2$, we finish the whole proof. □

Combining Proposition 3.5 and Proposition 3.6 gives to the following upper bounds over the deviation $d_{t+1} - d_t$.

Proposition E.3 (Compact Path Recursion for the Frozen Potential)**.** *Assume the conditions of Proposition D.5 hold. Then, the martingale difference*

$$\mathcal{M}_{t+1}^{\Psi} := -2\eta_t \left[\langle z_t, (C_t + \alpha_t I) u_t \rangle + \sum_{i=1}^{K} z_{t,i} u_{t,i}^2 \right]. \tag{69}$$

satisfies $\mathbb{E}_t[\mathcal{M}_{t+1}^{\Psi}] = 0$*, and* $|\mathcal{M}_{t+1}^{\Psi}| \le 6\eta_t \Big[(1+\alpha_t)\sqrt{d_t} + d_t \Big]$*. Besides, almost surely*

$$\begin{aligned} d_{t+1} - d_t \le & -2\eta_t \alpha_t d_t + 2\eta_t d_t^{3/2} + 2\eta_t E_t d_t \\ & + 2\eta_t E_t (1+\alpha_t)\sqrt{d_t} + 2\sqrt{K} |\Delta_{\alpha,t}| \sqrt{d_t} \\ & + 2(1+\alpha_t) \left\| r_t^{\mathrm{md}} \right\|_{w_t} \sqrt{d_t} + 2 \left\| r_t^{\mathrm{md}} \right\|_{w_t} d_t \\ & + \mathcal{R}_t^{\Psi} + \mathcal{M}_{t+1}^{\Psi}, \end{aligned} \tag{70}$$

where $\mathcal{R}_t^{\Psi}$ *is controlled by Proposition 3.6.*

F Proof for Section 3.3: (II) From the Taylor Expansion to Residual Recursive Expression

Proposition F.1. *Suppose for any* $k \in \mathbb{N}_+$ *that* $\eta_{T_k} = \eta\,(T_k + T_{\mathrm{off}})^{-3/4} \le \frac{1}{6M(H_k+1)}$*, and* $T_{\mathrm{off}} \ge \max\Big\{ 32HM\sqrt{K}, \frac{\lambda^2\eta^2}{a^2}\, 32^2 H^2 M^2 K \Big\}$*. Define the stopping time* $\mathcal{T} := \inf\big\{ t \ge 0 : d_t > \vartheta^2 \alpha_t^2 \big\}$ *with* $\vartheta \in (0, 1/2)$*. Define* $\chi := 1 - \vartheta - c_e$ *with* $c_e \le 1/2$*. Then there exist constants* $C_0, C_1, C_2 > 0$ *which only depends on the value of* $K, M, c_e, \vartheta, \eta, \lambda, a$ *such that, for every* $t < \mathcal{T}$*,*

$$d_{t+1} \le \big(1 - \chi\eta_t\alpha_t\big) d_t + C_0\eta_t^2 + C_1 \frac{\alpha_t}{\eta_t s_t^2} + M_{t+1}, \tag{71}$$

where $\mathbb{E}_t[M_{t+1}] = 0$*, and* $|M_{t+1}|^2 \le C_2 \eta_t^2 d_t$*.*

Proof of Proposition F.1. Starting from the one-step estimate in Eq.(70),

$$\begin{aligned} d_{t+1} - d_t \le & \underbrace{-2\eta_t\alpha_t d_t + 2\eta_t d_t^{3/2} + 2\eta_t E_t d_t}_{\text{(I)}} \\ & + \underbrace{2\eta_t E_t (1+\alpha_t)\sqrt{d_t} + 2\sqrt{K}\,|\Delta_{\alpha,t}|\sqrt{d_t}}_{\text{(II)}} \\ & + \underbrace{2(1+\alpha_t) \left\| r_t^{\mathrm{md}} \right\|_{w_t} \sqrt{d_t} + 2 \left\| r_t^{\mathrm{md}} \right\|_{w_t} d_t}_{\text{(III)}} + \underbrace{\mathcal{R}_t^{\Psi}}_{\text{(IV)}} + \underbrace{\mathcal{M}_{t+1}^{\Psi}}_{\text{(V)}}. \end{aligned} \tag{72}$$

For term (I) , the conditions of Proposition D.8 are satisfied and hence $E_t \le c_e\alpha_t$ with $c_e \le 1/2$. Then for any $t < \mathcal{T}$,

$$\text{(I)} = -2\eta_t d_t \big(\alpha_t - \sqrt{d_t} - E_t\big) \le -2\eta_t d_t (1 - \vartheta - c_e)\alpha_t = -2\chi\eta_t\alpha_t d_t. \tag{73}$$

For term (II), using $|\Delta_{\alpha,t}| \le \alpha_t/(4s_t)$ and Young's inequality,

$$\begin{aligned} 2\sqrt{K}\,|\Delta_{\alpha,t}|\sqrt{d_t} &\le \frac{\sqrt{K}}{2}\frac{\alpha_t}{s_t}\sqrt{d_t} \\ &= \sqrt{\chi\eta_t\alpha_t d_t}\sqrt{\frac{K\alpha_t}{4\chi\eta_t s_t^2}} \\ &\le \frac{\chi\eta_t\alpha_t d_t}{2} + \frac{K}{8\chi}\frac{\alpha_t}{\eta_t s_t^2}. \end{aligned} \tag{74}$$

Similarly,

$$2\eta_t E_t(1+\alpha_t)\sqrt{d_t} = \sqrt{\chi\eta_t\alpha_t d_t}\frac{2\eta_t E_t(1+\alpha_t)}{\sqrt{\chi\eta_t\alpha_t}} \leq \frac{\chi\eta_t\alpha_t d_t}{2} + \frac{2(1+\alpha_t)^2}{\chi}\frac{\eta_t E_t^2}{\alpha_t}. \tag{75}$$

From Proposition D.8,

$$E_t \leq D_E H(\lambda\eta + a)s_t^{-3/4}. \tag{76}$$

Hence

$$\frac{\eta_t E_t^2}{\alpha_t} \leq D_E^2 H^2(\lambda\eta+a)^2\frac{\eta_t s_t^{-3/2}}{\alpha_t} \leq D_E^2 H^2(\lambda\eta+a)^2\frac{\eta_t^2}{a\eta} \tag{77}$$

Consequently, let $C_{0,1} \triangleq \frac{2(1+\alpha_t)^2}{\chi a\eta}D_E^2H^2(\lambda\eta+a)^2$, $C_{1,1} \triangleq \frac{K}{8\chi}$, combining the above results gives

$$\boxed{\begin{aligned}&\text{(I)} + \text{(II)}\\ \leq& -\chi\eta_t\alpha_t d_t + \frac{K}{8\chi}\frac{\alpha_t}{\eta_t s_t^2} + \frac{2(1+\alpha_t)^2}{\chi a\eta}D_E^2H^2(\lambda\eta+a)^2\eta_t^2\\ \triangleq& -\chi\eta_t\alpha_t d_t + C_{1,1}\frac{\alpha_t}{\eta_t s_t^2} + C_{0,1}\eta_t^2\end{aligned}}$$

For term (III), according to Proposition D.5,

$$\left\|r_t^{\mathrm{md}}\right\|_{w_t} \leq 2\left\|w_{t+1}-w_t\right\|_{w_t}^2 \leq 8\eta_t^2M^2. \tag{78}$$

Therefore,

$$\boxed{\begin{aligned}\text{(III)} &\leq 8M^2\eta_t^2\left[2(1+\alpha_t)\sqrt{d_t}+2d_t\right]\\ &\leq 8M^2\eta_t^2\left[2(1+\alpha_0)\bar{\sigma}+2\bar{\sigma}^2\right] \triangleq C_{0,2}\eta_t^2.\end{aligned}}$$

For term (IV) , using Proposition E.3,

$$\boxed{\begin{aligned}\text{(IV)} = \mathcal{R}_t^{\Psi} &\leq K(\Delta_{\alpha,t})^2 + 4\left[(1+\alpha_t^2)+4\mathcal{B}_t+\mathcal{B}_t^2+2\alpha_t\mathcal{B}_t\right]\left\|\Delta_t\right\|_{w_t}^2\\ &\leq \frac{K}{16}\frac{\alpha_t^2}{s_t^2} + 16M^2\eta_t^2\left[(1+\alpha_0^2)+\overline{\mathcal{B}}\big(2\overline{\mathcal{B}}+2\alpha_0\big)\right]\\ &\leq \frac{Ka\eta}{16}\frac{\alpha_t}{\eta^t s_t^2} + 16M^2\eta_t^2\left[(1+\alpha_0^2)+\overline{\mathcal{B}}\big(2\overline{\mathcal{B}}+2\alpha_0\big)\right]\\ &\triangleq C_{1,2}\frac{\alpha_t}{\eta^t s_t^2} + C_{0,3}\eta_t^2\end{aligned}}$$

Here the last inequality holds because

$$\frac{\alpha_t^2}{s_t^2} \leq \frac{a\eta\alpha_t}{\eta_t s_t^2} \quad\Longleftrightarrow\quad \alpha_t\eta_t \leq a\eta \quad\Longleftrightarrow\quad l_t s_t^{-1} \leq l_0 s_0^{-1} \leq 1. \tag{79}$$

Finally, for term (V) , applying Proposition E.3, we have

$$\begin{aligned}&\text{(V)} = \mathcal{M}_{t+1}^{\Psi},\\ &\left|\mathcal{M}_{t+1}^{\Psi}\right|^2 \leq 36\eta_t^2\left[(1+\alpha_t)\sqrt{d_t}+\vartheta\alpha_t\sqrt{d_t}\right]^2\\ &\qquad\leq 36\eta_t^2\left[1+2\alpha_0\right]^2 d_t \triangleq C_2\eta_t^2 d_t,\end{aligned} \tag{80}$$

which leads to the almost surely upper bound on M_{t+1}. Finally, combining the above upper bounds on different terms (I)-(IV), and letting $C_0 = C_{0,1}+C_{0,2}+C_{0,3}$, $C_1 = C_{1,1}+C_{1,2}$ gives equation 71. To

be precise, since $M = 2+4a\sqrt{K}+\frac{\lambda\sqrt{K}}{2} = \mathcal{O}((a+\lambda)\sqrt{K})$, $\bar{\sigma} = \sigma + a\sqrt{K}$, $\bar{\mathcal{B}} = 2\bar{\sigma}+\frac{2+a}{2}+\frac{a\sqrt{K}}{4} = \mathcal{O}(a\sqrt{K}+\sigma)$, we have

$$C_0 = \frac{2(1+a)^2}{\chi a\eta} D_E^2 H^2(\lambda\eta + a)^2 + 8M^2[2(1+\alpha_0)\bar{\sigma} + 2\bar{\sigma}^2] + 16M^2\left[(1+a^2) + \overline{\mathcal{B}}\big(2\overline{\mathcal{B}} + 2a\big)\right]$$
$$= \mathcal{O}\Big((a+\lambda)^2 K[\sigma^2 + \frac{K^2H^2a^3}{\eta}]\Big) \tag{81}$$
$$C_1 = \frac{K}{8\chi} + \frac{Ka\eta}{16} = \mathcal{O}(Ka\eta), \qquad C_2 = 36\left[1+2a\right]^2 = \mathcal{O}(a^2). \qquad \square$$

Proof of Proposition 3.10. Recall equation 71 that

$$d_{t+1} \le \big(1 - \chi\eta_t\alpha_t\big)d_t + C_0\eta_t^2 + C_1\frac{\alpha_t}{\eta_t s_t^2} + M_{t+1}, \tag{82}$$

Dividing both terms with $a^2 l_{t+1}$. For the first term in the RHS, since $l_t \le l_{t+1}$, we have

$$\boxed{\frac{(1-\chi\eta_t\alpha_t)d_t}{a^2 l_{t+1}} = \frac{l_t}{l_{t+1}}\left(1 - \frac{\chi\eta a l_t}{s_t}\right)U_t \le \left(1 - \frac{1}{s_t}\right)U_t,}$$

where the last inequality uses the fact that $\chi\eta a l_t \ge \chi\eta a l_0 \ge 1$. The non-negativity of the right-hand side is guaranteed by the assumption that $\chi\eta a l_t/s_t \le \chi\eta a l_0/s_0 \le 1/2$.

The second term in the RHS satisfies

$$\boxed{\frac{C_0\eta_t^2}{a^2 l_{t+1}} = \frac{C_0\eta^2 s_t^{-3/2}}{a^2 l_{t+1}} \le \frac{C_0\eta^2}{a^2} s_{t+1}^{-3/2},}$$

where the last inequality due to the fact that $(s_{t+1}/s_t)^{3/2} \le \ln(s_{t+1}/\delta)$ when $s_{t+1} \ge T_{\text{off}} \ge 4$.

The third term

$$\boxed{\begin{aligned}\frac{C_1}{a^2 l_{t+1}}\frac{\alpha_t}{\eta_t s_t^2} &= \frac{C_1}{a\eta}\frac{l_t}{l_{t+1}} s_t^{-3/2} \\ &\le \frac{2C_1}{a\eta} s_{t+1}^{-3/2},\end{aligned}}$$

where the last inequality holds since $(s_{t+1}/s_t)^{3/2} \le (3/2)^{3/2} \le 2$. Finally,

$$|W_{t+1}|^2 \le \frac{C_2\eta_t^2 d_t}{a^4 l_{t+1}^2} = \frac{C_2\eta^2}{a^2} s_t^{-3/2}\frac{l_t}{l_{t+1}^2}U_t \le \frac{2C_2\eta^2}{a^2 l_0} s_{t+1}^{-3/2} U_t.$$

Combining the above results finishes the whole proof. $\square$

Proof of Theorem 3.11. Under the conditions of Proposition 3.10, before the stopping time $\mathcal{T}$, U_t satisfies all recursion conditions of Lemma 3.2. The initial condition is equivalent to

$$U_0 \le \frac{\vartheta^2}{8}\frac{l_0}{\sqrt{s_0}}.$$

Therefore

$$\mathbb{P}(\mathcal{T} < \infty) \le \delta.$$

Thus, with probability at least $1-\delta$, for all $t \ge 0$,

$$d_t \le \vartheta^2\alpha_t^2.$$

Because $\vartheta < 1$,

$$\sqrt{d_t} \le \vartheta\alpha_t \le \alpha_t.$$

Applying Lemma 3.1,

$$\mathrm{EG}(w_t) \le 2K\alpha_t = 2Ka\log\left(\frac{t+T_{\mathrm{off}}}{\delta}\right)(t+T_{\mathrm{off}})^{-1/4}.$$

Besides, to satisfy the conditions $B_* = \frac{C_0\eta^2}{a^2} + \frac{2C_1}{a\eta} \le \frac{\vartheta^2 l_0}{16}$, $V_\star = \frac{2C_2\eta^2}{a^2 l_0} \le \frac{\vartheta^2}{128}$ in Proposition 3.2, we plug the values of factors $C_0 = \mathcal{O}\Big((a+\lambda)^2K[\sigma^2 + \frac{K^2H^2a^3}{\eta}]\Big), C_1 = \mathcal{O}(Ka\eta), C_2 = \mathcal{O}(a^2)$ from equation 81 into the constants $B_*, V_\star$, we have

$$\begin{aligned} l_0 \ge& \max\Big(\frac{16}{\vartheta^2}(\frac{C_0\eta^2}{a^2} + \frac{2C_1}{a\eta}), \quad \frac{256C_2\eta^2}{a^2\vartheta^2}\Big) \\ =& \mathcal{O}\Big(K(a+\lambda)^2(\frac{\sigma^2\eta^2}{a^2} + K^2H^2a\eta)\Big), \end{aligned}$$

which is equivalent to $T_{\mathrm{off}} \ge \mathcal{O}(\exp\{K(a+\lambda)^2(\frac{\sigma^2\eta^2}{a^2} + K^2H^2a\eta)\}) \ge 4$. Using another condition $1 \le \chi\eta a l_0 \le \frac{s_0}{2}$ of Proposition 3.10, we have

$$T_{\mathrm{off}} \ge \exp\{\max(2\chi\eta a, \frac{1}{\chi\eta a})\}.$$

Combining the above results, $T_{\mathrm{off}} \ge \max\{\mathcal{O}(\exp\{K(a+\lambda)^2(\frac{\sigma^2\eta^2}{a^2} + K^2H^2a\eta)\}), \mathcal{O}(\exp\{\frac{1}{\eta a}\})\}$. □

G Proof for Section 3.4: Constant-Probability Lower Bound for Uncoupled Algorithms in Matrix Games with Bandit Feedback

G.1 Information model for uncoupled bandit learning

Section 2 writes the expected loss and exploitability as $\ell(\mu,\nu)$ and $\mathrm{EG}(\mu,\nu)$ for the ambient matrix A. In the lower-bound argument we compare several matrices simultaneously, so we use the explicit subscripts

$$\ell_A(\mu,\nu) := \mu^\top A\nu, \qquad \mathrm{EG}_A(\mu,\nu) := \max_{\nu'\in\Delta_n} \ell_A(\mu,\nu') - \min_{\mu'\in\Delta_m} \ell_A(\mu',\nu).$$

Lemma G.1 (Minimax Value of Mean Matrix and One-Sided Exploitability Separation)**.** *For every $\varepsilon \in [-1/12, 1/12]$, the minimax value of A^ε is $1/2$, and the max-player strategy $\nu^\star = (1/2,1/2)^\top$ is minimax. Moreover, let $\mu = (p, 1-p)^\top$. For every $\varepsilon \in (0, 1/12]$,*

$$p \le \frac{1}{2} \Longrightarrow \mathrm{EG}_{A+\varepsilon}(\mu,\nu) \ge \frac{\varepsilon}{2} \qquad \forall \nu \in \Delta_2, \tag{83}$$

$$p \ge \frac{1}{2} \Longrightarrow \mathrm{EG}_{A-\varepsilon}(\mu,\nu) \ge \frac{\varepsilon}{2} \qquad \forall \nu \in \Delta_2. \tag{84}$$

Proof. Take $\nu^\star = (1/2,1/2)^\top$. Since the loss vector against row-player is $A^\epsilon\nu^* = \begin{pmatrix}1/2\\1/2\end{pmatrix}$, we have $\min_{\mu\in\Delta_2} \ell_{A^\varepsilon}(\mu,\nu^\star) = 1/2$, and so

$$\max_{\nu\in\Delta_2}\min_{\mu\in\Delta_2} \ell_{A^\varepsilon}(\mu,\nu) \ge \frac{1}{2}. \tag{85}$$

For $\mu := \mu(p) = (p, 1-p)^\top$, the loss vector against column-player is $\mu^\top A^\epsilon = \begin{pmatrix}1/3 + p(1/3-\varepsilon)\\ 2/3 - p(1/3-\varepsilon)\end{pmatrix}$, and obviously $\mu^\top A^\epsilon \mathbf{1} = 1$. Setting $p^\star_\varepsilon := \frac{1}{2-6\varepsilon}$ gives $\mu(p^\star_\epsilon)^\top A^\epsilon = \begin{pmatrix}1/2\\1/2\end{pmatrix}$, and $\mu(p^\star_\epsilon)^\top A^\epsilon\nu = 1/2$ for any $\nu \in \Delta_2$. Notice that since $|\varepsilon| \le 1/12$, one has $p^\star_\varepsilon \in [2/5, 2/3] \subset [0,1]$. Therefore we have $\max_{\nu\in\Delta_2} \ell_{A^\epsilon}(\mu(p^\star_\epsilon),\nu) = 1/2$, and so

$$\min_{\mu\in\Delta_2}\max_{\nu\in\Delta_2} \ell_{A^\varepsilon}(\mu,\nu) \le \frac{1}{2}. \tag{86}$$

Combining equation 85-equation 86 gives the minimax value $\min_{\mu\in\Delta_2}\max_{\nu\in\Delta_2}\ell_{A^\varepsilon}(\mu,\nu)=1/2$. Now assume $\varepsilon>0$ and $p\le 1/2$. By simple computation,

$$\ell_{A^{+\varepsilon}}(\mu,e_2)=\frac{2}{3}-p\Big(\frac{1}{3}-\varepsilon\Big)\ge\frac{2}{3}-\frac{1}{2}\Big(\frac{1}{3}-\varepsilon\Big)=\frac{1}{2}+\frac{\varepsilon}{2}.$$

Therefore

$$\max_{\nu'\in\Delta_2}\ell_{A^{+\varepsilon}}(\mu,\nu')\ge\frac{1}{2}+\frac{\varepsilon}{2}.$$

Since the minimax value is $1/2$, then for every ν, $\min_{\mu'\in\Delta_2}\ell_{A^{+\varepsilon}}(\mu',\nu)\le\frac{1}{2}$. Using the definition of duality gap EG_A, we have

$$\mathrm{EG}_{A^{+\varepsilon}}(\mu,\nu)=\max_{\nu'\in\Delta_2}\ell_{A^{+\varepsilon}}(\mu,\nu')-\min_{\mu'\in\Delta_2}\ell_{A^{+\varepsilon}}(\mu',\nu)\ge-\frac{1}{2}+\left(\frac{1}{2}+\frac{\varepsilon}{2}\right)=\frac{\varepsilon}{2}.$$

This proves equation 83. For $A^{-\varepsilon}$, the loss against the first pure column is $\ell_{A^{-\varepsilon}}(\mu,e_1)=\frac{1}{3}+p\Big(\frac{1}{3}+\varepsilon\Big)$. If $p\ge 1/2$, then

$$\ell_{A^{-\varepsilon}}(\mu,e_1)\ge\frac{1}{3}+\frac{1}{2}\Big(\frac{1}{3}+\varepsilon\Big)=\frac{1}{2}+\frac{\varepsilon}{2}.$$

The same value argument yields $\mathrm{EG}_{A^{-\varepsilon}}(\mu,\nu)\ge\varepsilon/2$. □

We next give a KL inequality for Bernoulli random variables.

Lemma G.2 (Elementary Bernoulli KL bound)**.** *For $p,q\in(0,1)$,*

$$\mathrm{kl}(\mathrm{Ber}(p),\mathrm{Ber}(q))\le\frac{(p-q)^2}{q(1-q)}. \tag{87}$$

Consequently, if $q\in[1/4,3/4]$, then

$$\mathrm{kl}(\mathrm{Ber}(p),\mathrm{Ber}(q))\le\frac{16}{3}(p-q)^2. \tag{88}$$

Proof. Using basic inequality $\log x\le x-1$,

$$\begin{aligned}\mathrm{kl}(\mathrm{Ber}(p),\mathrm{Ber}(q))&=p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q}\\&\le p\left(\frac{p}{q}-1\right)+(1-p)\left(\frac{1-p}{1-q}-1\right)=\frac{(p-q)^2}{q(1-q)}.\end{aligned}$$

If $q\in[1/4,3/4]$, then $q(1-q)\ge 3/16$, proving equation 88. □

KL bound on the full joint trajectory. Recall that, for each $\varepsilon\in[-1/12,1/12]$, $\mathbb{P}_\varepsilon$ denotes the probability measure induced jointly by running the fixed learning algorithm in the stochastic game G^ε. For $T\ge 2$, define the full trajectory up to round $T-1$ by

$$Z_{T-1}:=(\omega_{\min},\omega_{\max},a_1,b_1,\ell_1,\dots,a_{T-1},b_{T-1},\ell_{T-1}).$$

Its trajectory space is

$$\mathsf{Z}_{T-1}:=\mathsf{S}_{\min}\times\mathsf{S}_{\max}\times\big([m]\times[n]\times[0,1]\big)^{T-1},$$

equipped with the product σ-algebra

$$\mathcal{Z}_{T-1}:=\mathcal{S}_{\min}\otimes\mathcal{S}_{\max}\otimes\Big(2^{[m]}\otimes 2^{[n]}\otimes\mathcal{B}([0,1])\Big)^{\otimes(T-1)}.$$

Thus Z_{T-1} is a random vector taking values in the measurable space $(\mathsf{Z}_{T-1},\mathcal{Z}_{T-1})$. We define

$$\mathbb{P}_\varepsilon^{T-1}:=\mathrm{Law}_{\mathbb{P}_\varepsilon}(Z_{T-1})=\mathbb{P}_\varepsilon\circ Z_{T-1}^{-1}. \tag{89}$$

Equivalently, for every event $E\in\mathcal{Z}_{T-1}$,

$$\mathbb{P}_\varepsilon^{T-1}(E)=\mathbb{P}_\varepsilon(Z_{T-1}\in E)\,. \tag{90}$$

Thus $\mathbb{P}_\varepsilon^{T-1}$ is the *joint probability law of the entire trajectory* through round $T-1$ under G^ε. Consequently, the KL divergence below compares the distributions of the *entire interaction histories* generated by the same learning algorithm under two different game environments.

Lemma G.3 (Full-trajectory KL bound)**.** *For every uncoupled learning algorithm, every* $\theta \in \{-1, +1\}$*, every* $\varepsilon \in (0, 1/12]$*, and every* $T \geq 2$*,*

$$D_{\mathrm{KL}}\big(\mathbb{P}_0^{T-1} \,\big\|\, \mathbb{P}_{\theta\varepsilon}^{T-1}\big) \leq \frac{16}{3}(T-1)\varepsilon^2 \leq \frac{16}{3}T\varepsilon^2. \tag{91}$$

Proof. The joint law of random seeds $(\omega_{\min}, \omega_{\max})$ is independent of the game, so its relative-entropy contribution is zero. We now apply the chain rule of two joint distributions $P_{X,Y}$ and $Q_{X,Y}$

$$D_{\mathrm{KL}}(P_{X,Y}\|Q_{X,Y}) = D_{\mathrm{KL}}(P_X\|Q_X) + \mathbb{E}_{X\sim P_X}\left[D_{\mathrm{KL}}\big(P_{Y|X}\|Q_{Y|X}\big)\right]$$

to the chronological trajectory, by conditioning further on a sigma-algebra $\mathcal{F}_t$ and taking $X = (a_t, b_t), Y = \ell_t$, we obtain for each $t < T$,

$$\begin{aligned} &D_{\mathrm{KL}}(\mathbb{P}_0(a_t, b_t, \ell_t \mid \mathcal{F}_t) \,\|\, \mathbb{P}_{\theta\varepsilon}(a_t, b_t, \ell_t \mid \mathcal{F}_t)) \\ =&D_{\mathrm{KL}}(\mathbb{P}_0(a_t, b_t \mid \mathcal{F}_t) \,\|\, \mathbb{P}_{\theta\varepsilon}(a_t, b_t \mid \mathcal{F}_t)) + \mathbb{E}_0[D_{\mathrm{KL}}(\mathbb{P}_0(\ell_t \mid \mathcal{F}_t, a_t, b_t) \,\|\, \mathbb{P}_{\theta\varepsilon}(\ell_t \mid \mathcal{F}_t, a_t, b_t)) \,|\, \mathcal{F}_t]\,. \end{aligned} \tag{92}$$

The first term in the right-hand side of equation 92 is exactly zero. Indeed, for any realized global history h_t of the history random variable H_t generating $\mathcal{F}_t$, the two candidate environments induce the same realized local histories, so

$$\mathbb{P}_0(a_t = a, b_t = b \mid H_t = h_t) = \mu_t(a)\nu_t(b) = \mathbb{P}_{\theta\varepsilon}(a_t = a, b_t = b \mid H_t = h_t).$$

The only remaining difference between the two environments is the Bernoulli observation law in the second term of the right-hand side of equation 92. Condition on $\mathcal{F}_t$ and on $(a_t, b_t) = (a, b)$. If $a = 2$, the corresponding cell distribution is identical in G^0 and $\mathsf{G}^{\theta\varepsilon}$, so the conditional KL is zero. If $a = 1$, the two Bernoulli means differ in absolute value by exactly ε. Since $\varepsilon \leq 1/12$, every Bernoulli parameter used as the second argument of the KL divergence lies in $[1/4, 3/4]$. Lemma G.2 then yields

$$D_{\mathrm{KL}}(\mathbb{P}_0(\ell_t \mid \mathcal{F}_t, a_t, b_t) \,\|\, \mathbb{P}_{\theta\varepsilon}(\ell_t \mid \mathcal{F}_t, a_t, b_t)) \leq \frac{16}{3}\varepsilon^2 \mathbf{1}\{a_t = 1\}. \tag{93}$$

Applying the relative-entropy chain rule over $t = 1, \ldots, T-1$, using the zero KL contribution of the seeds and the action kernels, and then equation 93, gives

$$\begin{aligned} D_{\mathrm{KL}}\big(\mathbb{P}_0^{T-1} \,\big\|\, \mathbb{P}_{\theta\varepsilon}^{T-1}\big) &= \sum_{t=1}^{T-1} \mathbb{E}_0[D_{\mathrm{KL}}(\mathbb{P}_0(a_t, b_t, \ell_t \mid \mathcal{F}_t) \,\|\, \mathbb{P}_{\theta\varepsilon}(a_t, b_t, \ell_t \mid \mathcal{F}_t))] \\ &\leq \frac{16}{3}\varepsilon^2 \mathbb{E}_0\left[\sum_{t=1}^{T-1} \mathbf{1}\{a_t = 1\}\right] \\ &\leq \frac{16}{3}(T-1)\varepsilon^2. \end{aligned}$$

□

G.2 Fixed-horizon minimax lower bound and its sharpness

Proof of Theorem 3.12. Set $\varepsilon := \sqrt{\frac{3}{128T}}$. Since $T \geq 4$, we have $\varepsilon < 1/12$. Write the min-player's round-T mixed strategy as $\mu_T = (p_T, 1-p_T)^\top$, Since μ_T is chosen before round T and is $\mathcal{F}_T^{\min}$-measurable, its first coordinate p_T is also $\mathcal{F}_T^{\min}$-measurable. As $\mathcal{F}_T^{\min} \subseteq \mathcal{F}_T$, p_T is therefore $\mathcal{F}_T$-measurable as well.

The threshold $1/2$ separates the two perturbed hard instances. Indeed, by the equilibrium calculation in Lemma G.1, the min-player's equilibrium probability of the first row is

$$p^*_{+\varepsilon} = \frac{1}{2-6\varepsilon} > \frac{1}{2} \qquad \text{in } A^{+\varepsilon}. \qquad\qquad p^*_{-\varepsilon} = \frac{1}{2+6\varepsilon} < \frac{1}{2} \qquad \text{in } A^{-\varepsilon}.$$

Accordingly, define the two wrong-side events

$$E_+ := \{p_T \leq 1/2\}, \qquad E_- := \{p_T \geq 1/2\}. \tag{94}$$

By Lemma G.1, E_+ forces exploitability at least $\varepsilon/2$ under $A^{+\varepsilon}$, whereas E_- forces exploitability at least $\varepsilon/2$ under $A^{-\varepsilon}$. Their union is the whole sample space, so under G^0, $\mathbb{P}_0(E_+) + \mathbb{P}_0(E_-) \geq 1$. Hence there exists a deterministic sign $\theta_T \in \{+1, -1\}$ such that

$$\mathbb{P}_0(E_{\theta_T}) \geq \frac{1}{2}. \tag{95}$$

Using Lemma G.3 and $\varepsilon := \sqrt{\frac{3}{128T}}$,

$$D_{\mathrm{KL}}\big(\mathbb{P}_0^{T-1} \,\big\|\, \mathbb{P}_{\theta_T\varepsilon}^{T-1}\big) \leq \frac{16}{3} T\varepsilon^2 = \frac{1}{8}.$$

Pinsker's inequality therefore gives the upper bound on the total variation,

$$\mathrm{TV}\big(\mathbb{P}_0^{T-1}, \mathbb{P}_{\theta_T\varepsilon}^{T-1}\big) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}\big(\mathbb{P}_0^{T-1} \,\big\|\, \mathbb{P}_{\theta_T\varepsilon}^{T-1}\big)} \leq \sqrt{\frac{1}{2} \cdot \frac{1}{8}} = \frac{1}{4}. \tag{96}$$

Since E_{θ_T} is $\mathcal{F}_T$-measurable, by the definition of total variation $\mathrm{TV}(P, Q) = \max_A |P(A) - Q(A)|$,

$$\mathbb{P}_{\theta_T\varepsilon}(E_{\theta_T}) \geq \mathbb{P}_0(E_{\theta_T}) - \mathrm{TV}\big(\mathbb{P}_0^{T-1}, \mathbb{P}_{\theta_T\varepsilon}^{T-1}\big) \geq \frac{1}{2} - \frac{1}{4} = \frac{1}{4}. \tag{97}$$

If $\theta_T = +1$, then on E_+ one has $p_T \leq 1/2$, so Lemma G.1 implies

$$\mathrm{EG}_{A^{+\varepsilon}}(\mu_T, \nu_T) \geq \frac{\varepsilon}{2}.$$

If $\theta_T = -1$, then on E_- one has $p_T \geq 1/2$, so

$$\mathrm{EG}_{A^{-\varepsilon}}(\mu_T, \nu_T) \geq \frac{\varepsilon}{2}.$$

Thus in both cases

$$E_{\theta_T} \subseteq \left\{ \mathrm{EG}_{A^{\theta_T\varepsilon}}(\mu_T, \nu_T) \geq \frac{\varepsilon}{2} \right\}.$$

Together with equation 97, we achieve the final results,

$$\mathbb{P}_{\theta_T\varepsilon}\left(\mathrm{EG}_{A^{\theta_T\varepsilon}}(\mu_T, \nu_T) \geq \frac{\varepsilon}{2} \right) \geq \mathbb{P}_{\theta_T\varepsilon}(E_{\theta_T}) \geq \frac{1}{4}. \qquad \square$$

Proof of Theorem 3.13. Let the first $T-1$ rounds follow the deterministic four-round cycle

$$(a_t, b_t) \in \{(1,1), (1,2), (2,1), (2,2)\}, \tag{98}$$

repeated in this order. Equivalently, on each exploration round both players use the corresponding pure mixed strategies. This is a valid uncoupled strategy pair: each player's round-t strategy is a deterministic function of the public round index t and hence is measurable with respect to its own local pre-action filtration. No player needs to observe the opponent's realized action. Let

$$N_{ij} := \#\{t \leq T-1 : (a_t, b_t) = (i,j)\}, \qquad n := \min_{i,j} N_{ij} = \left\lfloor \frac{T-1}{4} \right\rfloor.$$

For every cell, define the empirical mean $\widehat{A}_{ij} := \frac{1}{N_{ij}} \sum_{\substack{t \leq T-1: \\ (a_t, b_t) = (i,j)}} \ell_t$. At round T, let $(\widehat{\mu}, \widehat{\nu})$ be a saddle point of the empirical zero-sum game $\widehat{A}$, chosen using any fixed deterministic tie-breaking rule, and set the mixed-strategy at round T as $\mu_T := \widehat{\mu}, \quad \nu_T := \widehat{\nu}$. These terminal mixed strategies are measurable with respect to the respective local histories up to round $T-1$, so the construction remains within the information model of Section G.1.

By Hoeffding's inequality, for each cell (i,j) and every $\eta > 0$,

$$\mathbb{P}_A\Big(|\widehat{A}_{ij} - A_{ij}| > \eta \Big) \leq 2\exp(-2N_{ij}\eta^2) \leq 2\exp(-2n\eta^2).$$

A union bound over the four cells gives

$$\mathbb{P}_A\Big(\|\widehat{A} - A\|_\infty > \eta \Big) \leq 8\exp(-2n\eta^2).$$

Hence, with probability at least $1-\delta$,

$$\|\widehat{A}-A\|_\infty \le \eta_T(\delta) := \sqrt{\frac{\log(8/\delta)}{2n}}. \tag{99}$$

It remains to translate entrywise estimation error into exploitability. On the event $\|\widehat{A}-A\|_\infty \le \eta$, for every pair of mixed strategies (μ,ν),

$$|\mu^\top (A-\widehat{A})\nu| \le \eta. \tag{100}$$

Since $(\widehat{\mu},\widehat{\nu})$ is a saddle point of $\widehat{A}$, if $v(\widehat{A})$ denotes the value of the empirical game, then

$$\max_{\nu\in\Delta_2} \widehat{\mu}^\top A\nu \le \max_{\nu\in\Delta_2} \widehat{\mu}^\top \widehat{A}\nu + \eta = v(\widehat{A}) + \eta,$$
$$\min_{\mu\in\Delta_2} \mu^\top A\widehat{\nu} \ge \min_{\mu\in\Delta_2} \mu^\top \widehat{A}\widehat{\nu} - \eta = v(\widehat{A}) - \eta.$$

Therefore

$$\mathrm{EG}_A(\widehat{\mu},\widehat{\nu}) = \max_{\nu\in\Delta_2} \widehat{\mu}^\top A\nu - \min_{\mu\in\Delta_2} \mu^\top A\widehat{\nu} \le 2\eta. \tag{101}$$

Combining equation 99 and equation 101 yields the following inequality,

$$\mathbb{P}_A\left(\mathrm{EG}_A(\mu_T,\nu_T) \le \sqrt{\frac{2\log(8/\delta)}{\lfloor (T-1)/4 \rfloor}}\right) \ge 1-\delta, \tag{102}$$

Finally, for $T \ge 8$, using $\left\lfloor \frac{T-1}{4} \right\rfloor \ge \frac{T}{8}$, completes the whole proof. □

H Proof for Section 4: Extension of Last-Iterate Convergence to the Extensive-Form Games

Lemma H.1 (Normal-space invariance on the sequence-form polytope)**.** *If $q \in \mathcal{G}_{\mathrm{E}}$, then $\langle q,u\rangle$ is constant over $u \in \mathcal{W}_{\mathrm{E}}$. Consequently, for every $w \in \mathrm{ri}(\mathcal{W}_{\mathrm{E}})$, every $\eta>0$, and every direction ξ,*

$$\arg\min_{u\in\mathcal{W}_{\mathrm{E}}}\{D_{\Phi_{\mathrm{E}}}(u,w)+\eta\langle \xi,u\rangle\} = \arg\min_{u\in\mathcal{W}_{\mathrm{E}}}\{D_{\Phi_{\mathrm{E}}}(u,w)+\eta\langle \xi+q,u\rangle\}.$$

Proof. Write $q = E^\top y$. For every feasible u,

$$\langle q,u\rangle = y^\top Eu = y^\top e,$$

which is independent of u. The two mirror objectives therefore differ by a constant. □

Definition H.2 (Flow-polytope reference map)**.** For $\sigma \in \mathrm{ri}(\mathcal{W}_{\mathrm{E}})$ and $\rho \in (0,1]$, define

$$\mathsf{R}^{\mathrm{E}}_\rho(\sigma) := \arg\min_{u\in\mathrm{ri}(\mathcal{W}_{\mathrm{E}})} \left\{\Phi_{\mathrm{E}}(u) - (1-\rho)\langle \nabla\Phi_{\mathrm{E}}(\sigma),u\rangle\right\}. \tag{103}$$

Lemma H.3 (Existence, uniqueness, and exact start matching)**.** *For every $\sigma \in \mathrm{ri}(\mathcal{W}_{\mathrm{E}})$ and $\rho \in (0,1]$, the minimizer in equation 103 exists, is unique, and lies in $\mathrm{ri}(\mathcal{W}_{\mathrm{E}})$. Let $\alpha \in (0,\lambda]$, $\rho = \alpha/\lambda$, and $\widehat{\sigma} = \mathsf{R}^{\mathrm{E}}_\rho(\sigma)$. Then there exists $q \in \mathcal{G}_{\mathrm{E}}$ such that*

$$\lambda\big(\nabla\Phi_{\mathrm{E}}(\sigma) - \nabla\Phi_{\mathrm{E}}(\widehat{\sigma})\big) = \alpha\nabla\Phi_{\mathrm{E}}(\sigma) + q. \tag{104}$$

Proof. The sequence-form polytope is compact and has a nonempty relative interior. The log-barrier tends to $+\infty$ at every relative-boundary point where a realization coordinate tends to zero. Hence the objective in equation 103 attains its minimum in the relative interior. Strict convexity of Φ_{E} on the affine hull yields uniqueness.

The KKT condition gives a multiplier y for the flow constraints such that

$$\nabla\Phi_{\mathrm{E}}(\widehat{\sigma}) - (1-\rho)\nabla\Phi_{\mathrm{E}}(\sigma) + E^\top y = 0.$$

Rearranging and multiplying by λ gives

$$\lambda(\nabla\Phi_{\mathrm{E}}(\sigma) - \nabla\Phi_{\mathrm{E}}(\widehat{\sigma})) = \lambda\rho\nabla\Phi_{\mathrm{E}}(\sigma) + \lambda E^\top y.$$

Since $\lambda\rho = \alpha$ and $\lambda E^\top y \in \mathcal{G}_{\mathrm{E}}$, equation 104 follows. □

Algorithm 2: Adaptively Regularized OMD with Log-Barrier in a Perfect-Recall EFG

Input: Sequence-form polytopes $\mathcal{Q}_{\min}, \mathcal{Q}_{\max}$; fixed $\lambda > 0$; refresh times T_k; $\eta, a, T_{\text{off}}, \delta$.

Initialize $w_0 = (r_0, s_0)$ to satisfy $\|\Pi_0 \left(F(w_0) + \alpha_0 \nabla\Phi(w_0)\right)\|_{*,w_0} \le \alpha_0/8$, and set $k \leftarrow 0$.
for $t = 0, 1, 2, \ldots$ **do**
  if $t = T_k$ **then**
    $\sigma^k \leftarrow w_t$, $\rho_k \leftarrow \alpha_{T_k}/\lambda$;
    Compute the reference: $\widehat{\sigma}^k \leftarrow \arg\min_{u \in \mathrm{ri}(\mathcal{W}_{\mathrm{E}})} \left\{\Phi_{\mathrm{E}}(u) - (1-\rho_k)\langle \nabla\Phi_{\mathrm{E}}(\sigma^k), u\rangle\right\}$.
  Set $\eta_t \leftarrow \eta(t + T_{\text{off}})^{-3/4}$, $\alpha_t \leftarrow a \log((t + T_{\text{off}})/\delta)(t + T_{\text{off}})^{-1/4}$;
  Sample one complete trajectory of depth at most D_{g} and observe the stage losses $\{\ell_{t,h}\}_{h=0}^{H-1}$;
  Construct $\widehat{F}_t$ from equation 19;
  Update the whole sequence-form profile by
  $$w_{t+1} \leftarrow \arg\min_{w \in \mathcal{W}_{\mathrm{E}}} \left\{ D_{\Phi_{\mathrm{E}}}(w, w_t) + \eta_t \left\langle \widehat{F}_t + \lambda\big(\nabla\Phi_{\mathrm{E}}(w_t) - \nabla\Phi_{\mathrm{E}}(\widehat{\sigma}^k)\big), w \right\rangle \right\}.$$
  if $t + 1 = T_{k+1}$ **then**
    $k \leftarrow k + 1$

Lemma H.4 (Conditional unbiasedness modulo the sequence-form normal space)**.** *For every interior sequence-form strategy profile $w_t = (r_t, s_t)$, there exists a matrix $\widetilde{A}_{\mathrm{E}} \in \mathbb{R}^{K_{\min} \times K_{\max}}$ such that the conditional mean of the signed importance-sampling estimator satisfies*

$$\mathbb{E}_t[\widehat{F}_t] = \widetilde{F}_{\mathrm{E}}(w_t), \qquad \widetilde{F}_{\mathrm{E}}(r, s) := \begin{pmatrix} \widetilde{A}_{\mathrm{E}} s \\ -\widetilde{A}_{\mathrm{E}}^{\top} r \end{pmatrix}. \tag{105}$$

Moreover, $\widetilde{A}_{\mathrm{E}}$ represents the same payoff as the canonical matrix A_{E} on any sequence-form polytopes $(r, s) \in \mathcal{Q}_{\min} \times \mathcal{Q}_{\max}$: $r^{\top} \widetilde{A}_{\mathrm{E}} s = r^{\top} A_{\mathrm{E}} s = \ell(r, s)$. Consequently, $\widetilde{F}_{\mathrm{E}}(w_t) - F_{\mathrm{E}}(w_t) \in \mathcal{G}_{\mathrm{E}}$.

Proof of Lemma H.4. We divide the whole proof into 5 steps.

Step 1: Atomic depth-h history–action events and their common coefficient. Fix a depth h. Let $\mathcal{H}_h$ denote the set of complete game histories immediately before the two depth-h actions are selected. For $\xi \in \mathcal{H}_h$, write $x_h(\xi) \in \mathcal{X}_h$ and $y_h(\xi) \in \mathcal{Y}_h$ for the two information sets induced by ξ. For actions $a \in \mathcal{A}(x_h(\xi))$ and $b \in \mathcal{B}(y_h(\xi))$, let

$$i_h(\xi, a) := (x_h(\xi), a) \in \Sigma_{\min,h}, \qquad j_h(\xi, b) := (y_h(\xi), b) \in \Sigma_{\max,h}$$

denote the corresponding sequence-form coordinates. Let $\kappa_h(\xi)$ denote the product of all chance and state-transition probabilities along ξ, then the conditional probability of the complete depth-h event factors as

$$\mathbb{P}_t\big(\xi_h^t = \xi,\ a_h^t = a,\ b_h^t = b\big) = \kappa_h(\xi)\, r_{t,i_h(\xi,a)}\, s_{t,j_h(\xi,b)}. \tag{106}$$

For every pair of sequence coordinates $(i, j) \in \Sigma_{\min} \times \Sigma_{\max}$, define one common coefficient

$$c_{ij} := \sum_{h=0}^{D_{\mathrm{g}}-1} \sum_{\substack{\xi \in \mathcal{H}_h,\ a, b: \\ i_h(\xi,a)=i,\ j_h(\xi,b)=j}} \kappa_h(\xi)\, \bar{\ell}_h(\xi, a, b), \tag{107}$$

and set $\widetilde{A}_{\mathrm{E}}[i, j] := c_{ij}$. If no atomic history–action event is compatible with (i, j), the corresponding sum is empty and $c_{ij} = 0$. Finally set

$$\bar{\ell}_h(\xi, a, b) := \mathbb{E}_t\big[\ell_{t,h} \mid \xi_h^t = \xi,\ a_h^t = a,\ b_h^t = b\big].$$

Step 2: The minimizing block equals $\widetilde{A}_{\mathrm{E}} s_t$. Fix a minimizing-player coordinate $i = (x_h, a) \in \Sigma_{\min,h}$. For the fixed coordinate $i = (x_h, a)$, for every compatible history $\xi \in \mathcal{H}_h$ and maximizing-player action b, the disjoint union holds as

$$\{(x_h^t, a_h^t) = i\} = \bigsqcup_{\substack{\xi \in \mathcal{H}_h,\ b: \\ i_h(\xi,a)=i}} \{\xi_h^t = \xi,\ a_h^t = a,\ b_h^t = b\}$$

Applying the elementary conditional-expectation identity

$$\mathbb{E}_t[X\mathbf{1}_{\mathcal{E}}] = \mathbb{P}_t(\mathcal{E})\,\mathbb{E}_t[X \mid \mathcal{E}]$$

with $X = l_{t,h}$ and the definition $\bar{l}_h(\xi, a, b)$, the expectation of Eq.equation 19 becomes

$$\begin{aligned}
\mathbb{E}_t\Big[\widehat{F}_t^{\min}(i)\Big] &= \frac{1}{r_{t,i}}\mathbb{E}_t\big[l_{t,h}\,\mathbf{1}\{(x_h^t, a_h^t) = i\}\big] \\
&= \frac{1}{r_{t,i}}\mathbb{P}_t\big((x_h^t, a_h^t) = i\big)\bar{l}_h(\xi, a, b) \\
&= \frac{1}{r_{t,i}} \sum_{\substack{\xi\in\mathcal{H}_h,\ b: \\ i_h(\xi,a)=i}} \mathbb{P}_t\big(\xi_h^t = \xi,\ a_h^t = a,\ b_h^t = b\big)\,\bar{l}_h(\xi, a, b) \\
&= \sum_{\substack{\xi\in\mathcal{H}_h,\ b: \\ i_h(\xi,a)=i}} \kappa_h(\xi)\, s_{t,j_h(\xi,b)}\bar{l}_h(\xi, a, b) \qquad (108)
\end{aligned}$$

where the factor $r_{t,i}$ cancels in the last inequality due to equation 106. Before going on with the proof, we further define

$$\mathcal{I}_i := \{(\xi, b) :\ \xi \in \mathcal{H}_h,\ i_h(\xi, a) = i\},$$

and for each maximizing-player sequence coordinate $j \in \Sigma_{\max}$, define

$$\mathcal{I}_{ij} := \{(\xi, b) :\ \xi \in \mathcal{H}_h,\ i_h(\xi, a) = i,\ j_h(\xi, b) = j\}.$$

By perfect recall, every compatible pair (ξ, b) determines a unique maximizing-player sequence coordinate $j_h(\xi, b)$. Consequently, $\mathcal{I}_i = \bigsqcup_{j\in\Sigma_{\max}} \mathcal{I}_{ij}$, and for any coefficients $q_{\xi b}$,

$$\sum_{(\xi,b)\in\mathcal{I}_i} s_{t,j_h(\xi,b)} q_{\xi b} = \sum_{j\in\Sigma_{\max}} s_{t,j} \sum_{(\xi,b)\in\mathcal{I}_{ij}} q_{\xi b}.$$

Using the above results, Eq.(108) becomes

$$\begin{aligned}
\mathbb{E}_t\Big[\widehat{F}_t^{\min}(i)\Big] &= \sum_{\substack{\xi\in\mathcal{H}_h,\ b: \\ i_h(\xi,a)=i}} \kappa_h(\xi)\, s_{t,j_h(\xi,b)}\,\bar{l}_h(\xi, a, b) \\
&= \sum_{j\in\Sigma_{\max}} s_{t,j} \sum_{\substack{\xi\in\mathcal{H}_h,\ b: \\ i_h(\xi,a)=i,\ j_h(\xi,b)=j}} \kappa_h(\xi)\,\bar{l}_h(\xi, a, b) \qquad (109) \\
&= \sum_{j\in\Sigma_{\max}} \widetilde{A}_{\mathrm{E}}[i, j] s_{t,j} = (\widetilde{A}_{\mathrm{E}} s_t)_i. \qquad (110)
\end{aligned}$$

Thus the entire minimizing block satisfies $\mathbb{E}_t[\widehat{F}_t^{\min}] = \widetilde{A}_{\mathrm{E}} s_t$.

Step 3: The maximizing block uses the same coefficients and equals $-\widetilde{A}_{\mathrm{E}}^{\top} r_t$. Analogously, for any maximizing-player coordinate $j = (y_h, b) \in \Sigma_{\max,h}$, the signed estimator in equation 19 can be reformulated as

$$\begin{aligned}
\mathbb{E}_t\Big[\widehat{F}_t^{\max}(j)\Big] &= -\frac{1}{s_{t,j}}\mathbb{E}_t\big[\ell_{t,h}\,\mathbf{1}\{(y_h^t, b_h^t) = j\}\big] \\
&= -\frac{1}{s_{t,j}} \sum_{\substack{\xi\in\mathcal{H}_h,\ a: \\ j_h(\xi,b)=j}} \mathbb{P}_t\big(\xi_h^t = \xi,\ a_h^t = a,\ b_h^t = b\big)\,\bar{\ell}_h(\xi, a, b) \\
&= -\sum_{\substack{\xi\in\mathcal{H}_h,\ a: \\ j_h(\xi,b)=j}} \kappa_h(\xi)\, r_{t,i_h(\xi,a)}\,\bar{\ell}_h(\xi, a, b) \\
&= -\sum_{i\in\Sigma_{\min}} r_{t,i} \sum_{\substack{\xi\in\mathcal{H}_h,\ a: \\ i_h(\xi,a)=i,\ j_h(\xi,b)=j}} \kappa_h(\xi)\,\bar{\ell}_h(\xi, a, b) \\
&= -\sum_{i\in\Sigma_{\min}} \widetilde{A}_{\mathrm{E}}[i, j] r_{t,i} \\
&= -(\widetilde{A}_{\mathrm{E}}^{\top} r_t)_j. \qquad (111)
\end{aligned}$$

Thus we have $\mathbb{E}_t[\widehat{F}_t^{\max}] = -\widetilde{A}_{\mathrm{E}}^\top r_t$. Combining the above results yields

$$\mathbb{E}_t[\widehat{F}_t] = \begin{pmatrix} \widetilde{A}_{\mathrm{E}} s_t \\ -\widetilde{A}_{\mathrm{E}}^\top r_t \end{pmatrix} = \widetilde{F}_{\mathrm{E}}(w_t),$$

which proves equation 105.

Step 4: $\widetilde{A}_{\mathrm{E}}$ represents the same feasible payoff as A_{E}. Take any feasible realization plans $(r, s) \in \mathcal{Q}_{\min} \times \mathcal{Q}_{\max}$. Using equation 107, equation 106 and the definition of $\bar{l}_h(\xi, a, b)$,

$$\begin{aligned} r^\top \widetilde{A}_{\mathrm{E}} s &= \sum_{i \in \Sigma_{\min}} \sum_{j \in \Sigma_{\max}} r_i s_j c_{ij} \\ &= \sum_{h=0}^{D_{\mathrm{g}}-1} \sum_{\xi \in \mathcal{H}_h} \sum_{a,b} \kappa_h(\xi)\, r_{i_h(\xi,a)}\, s_{j_h(\xi,b)}\, \bar{l}_h(\xi, a, b) \\ &= \sum_{h=0}^{D_{\mathrm{g}}-1} \sum_{\xi \in \mathcal{H}_h} \sum_{a,b} \mathbb{P}(\xi_h = \xi,\ a_h = a,\ b_h = b)\, \mathbb{E}[\ell_h \mid \xi_h = \xi,\ a_h = a,\ b_h = b] \qquad (112) \\ &= \sum_{h=0}^{D_{\mathrm{g}}-1} \mathbb{E}_{r,s}[\ell_h] = \mathbb{E}_{r,s}\left[\sum_{h=0}^{D_{\mathrm{g}}-1} \ell_h\right] = \ell(r, s) = r^\top A_{\mathrm{E}} s. \qquad (113) \end{aligned}$$

Therefore, $r^\top \widetilde{A}_{\mathrm{E}} s = r^\top A_{\mathrm{E}} s = \ell(r, s), \qquad \forall (r, s) \in \mathcal{Q}_{\min} \times \mathcal{Q}_{\max}$.

Step 5: Difference from the canonical representative is a normal-space vector. Let $D := \widetilde{A}_{\mathrm{E}} - A_{\mathrm{E}}$. Then we have

$$r^\top D s = 0 \qquad \forall (r, s) \in \mathcal{Q}_{\min} \times \mathcal{Q}_{\max}.$$

For any fixed feasible s and an interior feasible r°, then every tangent direction $\delta r \in \ker E_{\min}$ with a constant ε leads to

$$(r^\circ + \varepsilon \delta r)^\top D s = 0, \qquad (r^\circ)^\top D s = 0.$$

This gives

$$\delta r^\top D s = 0 \qquad \forall \delta r \in \ker E_{\min}.$$

Hence

$$D s \in (\ker E_{\min})^\perp = \mathrm{range}(E_{\min}^\top).$$

By the symmetric argument,

$$D^\top r \in \mathrm{range}(E_{\max}^\top).$$

Then at $w_t = (r_t, s_t)$,

$$\widetilde{F}_{\mathrm{E}}(w_t) - F_{\mathrm{E}}(w_t) = \begin{pmatrix} D s_t \\ -D^\top r_t \end{pmatrix} \in \mathrm{range}(E^\top) = \mathcal{G}_{\mathrm{E}},$$

which proves $\widetilde{F}_{\mathrm{E}}(w_t) - F_{\mathrm{E}}(w_t) \in \mathcal{G}_{\mathrm{E}}$.

Finally, $\Pi^{\mathrm{E}}_{w_t}$ annihilates every vector in $\mathcal{G}_{\mathrm{E}}$, so $\Pi^{\mathrm{E}}_{w_t} \mathbb{E}_t[\widehat{F}_t] = \Pi^{\mathrm{E}}_{w_t} \widetilde{F}_{\mathrm{E}}(w_t) = \Pi^{\mathrm{E}}_{w_t} F_{\mathrm{E}}(w_t)$. □

Lemma H.5 (Pathwise local-dual-norm bound)**.** *For every interior realization-plan profile $w_t = (r_t, s_t)$, almost surely,*

$$\|\widehat{F}_t\|^2_{*,w_t} \le 2D_{\mathrm{g}}, \qquad \|\widehat{F}_t\|_{*,w_t} \le \sigma_{\mathrm{E}} := \sqrt{2D_{\mathrm{g}}}. \qquad (114)$$

Proof. Using the definition of $F_t^{\min}$ in equation 19,

$$\begin{aligned}\|\widehat{F}_t^{\min}\|_{*,r_t}^2 &= \sum_{h=0}^{D_{\mathrm{g}}-1} \sum_{x_h\in\mathcal{X}_h} \sum_{a\in\mathcal{A}(x_h)} r_t(x_h,a)^2\big(\widehat{F}_t^{\min}(x_h,a)\big)^2 \\ &= \sum_{h=0}^{D_{\mathrm{g}}-1} \ell_{t,h}^2 \sum_{x_h\in\mathcal{X}_h} \sum_{a\in\mathcal{A}(x_h)} \mathbf{1}\big\{(x_h^t,a_h^t)=(x_h,a)\big\} \\ &\le \sum_{h=0}^{D_{\mathrm{g}}-1} \ell_{t,h}^2 \le D_{\mathrm{g}}.\end{aligned}$$

The same calculation for the maximizing player gives $\|\widehat{F}_t^{\max}\|_{*,s_t}^2 \le D_{\mathrm{g}}$. Finally, we have

$$\|\widehat{F}_t\|_{*,w_t}^2 = \|\widehat{F}_t^{\min}\|_{*,r_t}^2 + \|\widehat{F}_t^{\max}\|_{*,s_t}^2 \le 2D_{\mathrm{g}}. \qquad \square$$

Corollary H.6 (Conditional-mean and centered-noise bounds)**.** *Under the assumptions of Lemmas H.4–H.5,*

$$\|\widetilde{F}_{\mathrm{E}}(w_t)\|_{*,w_t} \le \sigma_{\mathrm{E}}, \qquad \|\widehat{F}_t - \widetilde{F}_{\mathrm{E}}(w_t)\|_{*,w_t} \le 2\sigma_{\mathrm{E}} \quad \textit{a.s.} \tag{115}$$

Moreover, the projected canonical mean satisfies

$$\|\Pi_{w_t}^{\mathrm{E}} F_{\mathrm{E}}(w_t)\|_{*,w_t} = \|\Pi_{w_t}^{\mathrm{E}} \widetilde{F}_{\mathrm{E}}(w_t)\|_{*,w_t} \le \sigma_{\mathrm{E}}. \tag{116}$$

Proof. Using Jensen's inequality, Lemma H.4, and Lemma H.5 give

$$\|\widetilde{F}_{\mathrm{E}}(w_t)\|_{*,w_t} = \|\mathbb{E}_t[\widehat{F}_t]\|_{*,w_t} \le \mathbb{E}_t\|\widehat{F}_t\|_{*,w_t} \le \sigma_{\mathrm{E}}. \qquad \square$$

Proof of Lemma 4.1. Set

$$B := D_r A_{\mathrm{E}} D_s.$$

Fix a minimizing-player sequence $i \in \Sigma_{\min}$. Since $r_i, s_j \ge 0$, using the definition equation 17 gives

$$\begin{aligned}\sum_j |B_{ij}| &= r_i \sum_j s_j |(A_{\mathrm{E}})_{ij}| && (117)\\ &\le r_i \sum_j s_j \sum_{\substack{z\in\mathcal{Z}:\\ \sigma_{\min}(z)=i,\ \sigma_{\max}(z)=j}} \pi_c(z)|L(z)| && (118)\\ &= \sum_{\substack{z\in\mathcal{Z}:\\ \sigma_{\min}(z)=i}} \mathbb{P}_{r,s}(z)|L(z)| && (119)\\ &\le U_{\mathrm{E}} \sum_{\substack{z\in\mathcal{Z}:\\ \sigma_{\min}(z)=i}} \mathbb{P}_{r,s}(z) \le U_{\mathrm{E}}. && (120)\end{aligned}$$

The third line uses the reaching probability factorization $\mathbb{P}_{r,s}(z) = r_{\sigma_{\min}(z)} s_{\sigma_{\max}(z)} \pi_c(z)$. The last inequality follows due to the fact that $\sum_{z\in\mathcal{Z}} \mathbb{P}_{r,s}(z) = 1$. Taking the maximum over rows yields

$$\|B\|_\infty \le U_{\mathrm{E}}.$$

The column bound is symmetric. For every $j \in \Sigma_{\max}$,

$$\sum_i |B_{ij}| \le \sum_{\substack{z\in\mathcal{Z}:\\ \sigma_{\max}(z)=j}} \mathbb{P}_{r,s}(z)|L(z)| \le U_{\mathrm{E}}, \tag{121}$$

so $\|B\|_1 \le U_{\mathrm{E}}$. The Schur inequality therefore gives

$$\|B\|_{\mathrm{op}} \le \sqrt{\|B\|_1\|B\|_\infty} \le U_{\mathrm{E}}. \tag{122}$$

Finally,

$$W(w)J_{\mathrm{E}}W(w) = \begin{pmatrix} 0 & B \\ -B^\top & 0 \end{pmatrix}.$$

Hence

$$\|W(w)J_{\mathrm{E}}W(w)\|_{\mathrm{op}} = \|B\|_{\mathrm{op}} \le U_{\mathrm{E}}. \qquad \square$$