FASTER LAST-ITERATE CONVERGENCE FOR ZERO-SUM GAMES WITH STOCHASTIC FEEDBACK VIA ANCHOR-REGULARIZED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study uncoupled learning in two-player zero-sum matrix games under stochastic feedback. Each player updates from only its own feedback, without observing the opponent's action, strategy, or payoffs. We propose anchor-regularized follow-the-regularized-leader (AR-FTRL), which repeatedly solves the game regularized toward an anchor strategy profile, updating the anchor to each solution. As a warm-up, under unbiased noisy gradients with bounded second moment, the L^2 norm of the gap of the last strategy converges at rate $\widetilde{\mathcal{O}}(T^{-1/2})$. This guarantee exploits the metric subregularity of the zero-sum game's bilinear structure, under which the regularized equilibrium contracts toward the Nash equilibrium set by a constant factor at each anchor update. Our main result extends this to bandit feedback, where each player observes only the scalar payoff of its sampled action. Under a unique fully mixed Nash equilibrium, the L^2 norm of the gap of the last strategy converges at the same $\widetilde{\mathcal{O}}(T^{-1/2})$ rate after T observations, even though the importance-weighted gradient estimator has unbounded second moment near the boundary of the simplex. Both guarantees hold only at the anchor updates, not at every round. Under bandit feedback, this yields a faster rate than the $\widetilde{\mathcal{O}}(t^{-1/4})$ of every-round guarantees, though the guarantee is no longer anytime.

1 INTRODUCTION

This paper studies learning equilibria in two-player zero-sum matrix games under stochastic feedback. When each player minimizes regret, the time-averaged strategies converge to a Nash equilibrium, which is referred to as *average-iterate convergence*. However, the actual sequence of strategies does not necessarily converge and can cycle around the equilibrium even in simple bilinear cases (Mertikopoulos et al., 2018; Bailey & Piliouras, 2018; Cheung & Piliouras, 2019). This motivates *last-iterate convergence*, a stronger notion than average-iterate convergence, in which the final strategy itself approaches the equilibrium set. This stronger guarantee is useful when the final strategy is deployed directly or when maintaining an average requires an additional model, as in function approximation (Lee et al., 2021).

Stochastic feedback is a natural information model for repeated games because the payoff matrix and exact gradient vectors may be unavailable to the players. Under noisy gradient feedback, each player observes an unbiased estimate of its full gradient, as may arise from sampling or mini-batch evaluation. Under the more restrictive bandit feedback model, each player observes only its sampled action and the realized scalar payoff (Cai et al., 2023; Fiegel et al., 2025). The player must then estimate its gradient from this scalar payoff, most commonly by importance weighting. Unlike the bounded-variance estimate available under noisy gradients, this estimator has unbounded variance, since it grows as a sampling probability approaches zero (Lattimore & Szepesvári, 2020). Controlling this variance is the central difficulty under bandit feedback.

Prior results provide either rates for strategies played at every round t or rates for a strategy returned at a prescribed horizon T. Under bandit feedback, for the strategy profile played at every round t, Cai et al. (2023) established a high-probability $\widetilde{\mathcal{O}}(t^{-1/8})$ rate,[1] and Fiegel et al. (2025) obtained an

[1]Throughout, $\widetilde{\mathcal{O}}(\cdot)$ suppresses constants and polylogarithmic factors in its time argument.

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.

Table 1: Finite-time last-iterate guarantees under bandit feedback. Unless stated otherwise, the games are two-player zero-sum matrix games. Rates in t hold for the strategy profile played at every round t, whereas the rate in T holds for the last strategy at a prescribed time T.

Work	Convergence rate of the gap
Cai et al. (2023)	High probability: $\widetilde{\mathcal{O}}(t^{-1/8})$ L^2: $\widetilde{\mathcal{O}}(t^{-1/6})$
Cai et al. (2025) (zero-sum polymatrix games)	High probability: $\widetilde{\mathcal{O}}(t^{-1/5})$
Fiegel et al. (2025)	L^2: $\widetilde{\mathcal{O}}(t^{-1/4})$
Fiegel et al. (2026)	High probability: $\widetilde{\mathcal{O}}(t^{-1/4})$
This work (unique fully mixed Nash equilibrium)	L^2: $\boldsymbol{\widetilde{\mathcal{O}}(T^{-1/2})}$

$\widetilde{\mathcal{O}}(t^{-1/4})$ rate for the L^2 norm of the gap. Under noisy gradient feedback, existing methods attain asymptotic convergence or slower polynomial rates (Abe et al., 2023; 2024; 2025; Chen & Mazumdar, 2024; Ito et al., 2026). It has therefore remained open whether the last iterate itself, rather than a time average, can attain the canonical $T^{-1/2}$ rate (Nemirovski et al., 2009)[2] at a prescribed sequence of times.

To overcome this barrier, we propose *anchor-regularized follow-the-regularized-leader* (AR-FTRL), which repeatedly approximately solves an anchor-regularized game over successive epochs. Within each epoch, the players hold a reference profile, called the anchor, fixed and use log-barrier FTRL to track the equilibrium of the corresponding regularized game. At the end of each epoch, they reset the anchor to the last iterate and proceed to the next regularized game. The log-barrier regularizer keeps every played strategy in the interior of the simplex, controlling the variance of the gradient estimates, while repeated anchor resets move the anchor toward the Nash equilibrium rather than relying on a vanishing regularizer. AR-FTRL adopts the periodic anchor resets used by payoff-perturbation methods to recover an exact equilibrium (Perolat et al., 2021; Abe et al., 2023; 2024). Under a unique fully mixed Nash equilibrium, AR-FTRL achieves an $\widetilde{\mathcal{O}}(T^{-1/2})$ rate for the L^2 norm of the last strategy's gap only at the epoch boundaries, without requiring the terminal horizon in advance.

Our contributions are twofold. We first analyze noisy gradient feedback as a warm-up that isolates the contraction mechanism of the anchor resets. Under unbiased noisy gradients with bounded second moment, we prove a rate of $\widetilde{\mathcal{O}}(T^{-1/2})$ for the last strategy at the epoch boundaries. To the best of our knowledge, this is the first last-iterate rate of this order in a constrained bilinear game under noisy gradient feedback. We then extend the mechanism to bandit feedback, which is the main focus of the paper. Under bounded scalar feedback whose conditional mean equals the corresponding game payoff and a unique fully mixed Nash equilibrium, we establish the same rate for the L^2 norm of the gap from payoff observations alone. To the best of our knowledge, this is the first such guarantee from bandit feedback under these assumptions. In both settings, the epoch schedule is problem-independent and sub-geometric. The algorithm requires neither the metric-subregularity constant nor the minimum equilibrium coordinate.

2 Related Work

Learning with bandit feedback. Learning in games from limited feedback has been studied through regret and asymptotic equilibrium convergence, from payoff-value (zeroth-order) feedback in concave and monotone games (Bravo et al., 2018; Tatarenko & Kamgarpour, 2019) and from gradient feedback (Mertikopoulos & Zhou, 2019; Giannou et al., 2021). Finite-time last-iterate guarantees under this feedback model are more recent. For the strategy profile played at every round t, Cai et al. (2023) establish a high-probability $\widetilde{\mathcal{O}}(t^{-1/8})$ rate. Cai et al. (2025) obtain a high-probability $\widetilde{\mathcal{O}}(t^{-1/5})$ rate for the broader class of zero-sum polymatrix games. Fiegel et al. (2025) obtain an $\widetilde{\mathcal{O}}(t^{-1/4})$ rate for the L^2 norm of the gap. Fiegel et al. (2026) strengthen this to a high-probability guarantee at the same $\widetilde{\mathcal{O}}(t^{-1/4})$ rate on the strategy played at every round. Ito et al. (2025) establish

[2]Unless otherwise stated, rates for our stochastic-feedback guarantees refer to the L^2 norm of the gap function defined in Section 3.

an instance-dependent last-iterate rate for games with a unique pure-strategy Nash equilibrium. Under a different feedback model, Maiti et al. (2026) observe the expected payoff $x^\top Ay$ of the mixed strategies rather than the realized payoff of a sampled action. The authors note that the two settings are therefore not directly comparable. In contrast, AR-FTRL attains an $\widetilde{\mathcal{O}}(T^{-1/2})$ rate for the L^2 norm of the gap of the last strategy at a prescribed time T. This improves on the fastest every-round rate, $\widetilde{\mathcal{O}}(t^{-1/4})$, at the cost of holding only at such prescribed times rather than at every round. Table 1 summarizes these guarantees. Last-iterate guarantees under full and noisy gradient feedback, together with a detailed comparison against our noisy-gradient result, are deferred to Appendix D.

Payoff perturbation and periodic reference resets. Payoff perturbation regularizes the players' payoff functions with strongly convex terms to stabilize learning, a classical technique (Facchinei & Pang, 2003) whose roots trace to quantal response equilibria (McKelvey & Palfrey, 1995). This idea has since been developed through entropy and strongly convex regularization (Cen et al., 2021; 2023; Pattathil et al., 2023). Under a fixed perturbation, these methods converge only to an approximate equilibrium of the original game (Sokota et al., 2023; Coucheney et al., 2015; Leslie & Collins, 2005; Hussain et al., 2023). To recover an exact equilibrium, later studies periodically re-initialize the reference strategy so that the perturbed objective stays strongly convex (Perolat et al., 2021; Abe et al., 2023; 2024). In particular, adaptively perturbed mirror descent repeatedly updates its slingshot strategy to remove the bias of a fixed perturbation and converge to an equilibrium of the original game (Abe et al., 2024). AR-FTRL adopts the same idea of a periodically reset reference, but it holds the anchor fixed within each epoch and runs log-barrier FTRL on the game regularized around the anchor. This separation yields a contractive map on the anchors under metric subregularity. It also enables the global and local inner-loop bounds required under importance-weighted bandit feedback.

3 Preliminaries

This section formalizes the two-player zero-sum normal-form game that we study and the online learning protocol with gradient feedback.

Two-player zero-sum normal-form game. We study a two-player zero-sum *normal-form game*, or *matrix game*, with payoff matrix $A \in \mathbb{R}^{m\times n}$. Its min-max formulation is the *bilinear saddle-point problem*

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} x^\top Ay, \tag{1}$$

over the probability simplices $\mathcal{X} = \Delta_m$ and $\mathcal{Y} = \Delta_n$, where $\Delta_d := \{v \in \mathbb{R}^d_{\geq 0} : \mathbf{1}_d^\top v = 1\}$. We refer to $x^\top Ay$ as the *payoff function* and to $z := (x, y) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ as a *strategy profile*. Let $D := \max_{z,z'\in\mathcal{Z}} \|z - z'\|$ denote the diameter of $\mathcal{Z}$, and let $F(z) := (F_x(z), F_y(z)) = (Ay, -A^\top x)$ denote the gradient operator induced by the payoff function.

Nash equilibrium and gap function. A strategy profile $(x^*, y^*) \in \mathcal{Z}$ is a *Nash equilibrium* if

$$(x^*)^\top Ay \leq (x^*)^\top Ay^* \leq x^\top Ay^*, \qquad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}.$$

Let $\mathcal{Z}^* \subseteq \mathcal{Z}$ denote the set of Nash equilibria. To quantify the proximity of a strategy profile $z = (x, y)$ to equilibrium, we use the *gap function*

$$\mathrm{GAP}(z) := \max_{\tilde{y}\in\mathcal{Y}} x^\top A\tilde{y} - \min_{\tilde{x}\in\mathcal{X}} \tilde{x}^\top Ay,$$

which is non-negative for all $z \in \mathcal{Z}$ and equals zero exactly if and only if $z \in \mathcal{Z}^*$.

Online learning protocol. We consider the online learning protocol in which, at every iteration $t = 1, \ldots, T$, (i) each player selects a strategy $x^t \in \mathcal{X}$ and $y^t \in \mathcal{Y}$ based on previously observed feedback, and (ii) each player observes gradient feedback $\widehat{F}_x^t$ and $\widehat{F}_y^t$, respectively. We write $\widehat{F}^t := ((\widehat{F}_x^t)^\top, (\widehat{F}_y^t)^\top)^\top$, and let $\mathcal{H}_t$ denote the history available after z^t is chosen and just before the feedback at iteration t is observed.

Additional notation. For a closed convex set $\mathcal{S} \subseteq \mathbb{R}^d$, let $\Pi_{\mathcal{S}}$ denote the Euclidean projection onto $\mathcal{S}$. We write $\|A\|$ for the largest singular value of A, $R_x := \max_{x \in \mathcal{X}} \|x\|$ and $R_y := \max_{y \in \mathcal{Y}} \|y\|$ for the maximum Euclidean norms over $\mathcal{X}$ and $\mathcal{Y}$, and $\operatorname{dist}(z, \mathcal{Z}^*) := \|z - \Pi_{\mathcal{Z}^*}(z)\|$ for the Euclidean distance from $z \in \mathcal{Z}$ to the equilibrium set $\mathcal{Z}^*$.

4 Anchor-Regularized FTRL

This section introduces *anchor-regularized follow-the-regularized-leader* (AR-FTRL), a stochastic first-order method for the bilinear saddle-point problem (1). To obtain last-iterate convergence to a Nash equilibrium, AR-FTRL runs in epochs $k = 1, \dots, K$, and in each epoch it fixes an *anchor* $r = (r_x, r_y) \in \mathcal{Z}$ and approximately solves the *anchor-regularized game*

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \; x^\top A y + (\mu/2)\|x - r_x\|^2 - (\mu/2)\|y - r_y\|^2 \tag{2}$$

for a fixed regularization strength $\mu > 0$. The quadratic terms make the objective strongly convex in x and strongly concave in y. Hence, (2) has a unique saddle point. In epoch k, AR-FTRL runs the epoch for $N_k \in \mathbb{N}$ iterations of log-barrier FTRL to approximately solve (2) with the anchor $r^k = (r_x^k, r_y^k)$. The first anchor is the uniform profile $r^1 := u$, where $u_x := \mathbf{1}_m/m$, $u_y := \mathbf{1}_n/n$, and $u := (u_x, u_y)$.

Within epoch k, we initialize at the anchor, $z_{k,1} := r^k$, and write $\widehat{F}^{k,s}$ for the feedback observed at $z_{k,s}$. For a fixed learning rate $\eta > 0$, the players update independently at each $s \in \{1, \dots, N_k\}$ as

$$\begin{aligned} x_{k,s+1} &:= \arg\min_{x \in \mathcal{X}} \left\{ \eta \left\langle \sum_{v=1}^{s} \widehat{F}_x^{k,v}, x \right\rangle + \frac{s\eta\mu}{2}\|x - r_x^k\|^2 - \sum_{i=1}^{m} \log x_i \right\}, \\ y_{k,s+1} &:= \arg\min_{y \in \mathcal{Y}} \left\{ \eta \left\langle \sum_{v=1}^{s} \widehat{F}_y^{k,v}, y \right\rangle + \frac{s\eta\mu}{2}\|y - r_y^k\|^2 - \sum_{j=1}^{n} \log y_j \right\}. \end{aligned} \tag{3}$$

The two playerwise updates in (3) are equivalently written as the joint update

$$z_{k,s+1} = \arg\min_{z \in \mathcal{Z}} \left\{ \eta \left\langle \sum_{v=1}^{s} \widehat{F}^{k,v}, z \right\rangle + \frac{s\eta\mu}{2}\|z - r^k\|^2 - \sum_{i=1}^{m} \log x_i - \sum_{j=1}^{n} \log y_j \right\}, \tag{4}$$

where $z = (x, y)$. The logarithmic terms keep every iterate strictly positive. Hence, (4) is well defined.

At the end of epoch k, AR-FTRL resets the anchor to the last iterate, setting $r_x^{k+1} := x_{k,N_k+1}$, $r_y^{k+1} := y_{k,N_k+1}$, and $r^{k+1} := (r_x^{k+1}, r_y^{k+1})$. Epoch $k+1$ then starts from this anchor, and its cumulative feedback in (4) restarts from zero. Each epoch thus runs the inner loop afresh and is coupled to the previous epoch only through its anchor.

Epoch k occupies iterations $T_{k-1} + 1, \dots, T_k$, where $T_0 := 0$ and $T_k := \sum_{j=1}^{k} N_j$. Indexing by the global iteration $t = T_{k-1} + s$, we can write $z^t = z_{k,s}$ and $\widehat{F}^t = \widehat{F}^{k,s}$ for $s = 1, \dots, N_k$. Under this indexing, the reset makes the last iterate of epoch k the first iterate of epoch $k+1$:

$$z^{T_k+1} = z_{k,N_k+1} = r^{k+1} = z_{k+1,1}. \tag{5}$$

The epoch lengths $(N_k)_{k=1}^K$ are set by the convergence theorems, and Algorithm 1 summarizes the procedure.

5 Warm-up: Noisy Gradient Feedback

As a warm-up, we analyze AR-FTRL under noisy gradient feedback. This setting isolates the contraction mechanism that also drives the result under bandit feedback in Section 6. The feedback available to AR-FTRL is a stochastic estimate $\widehat{F}^t$ of the gradient $F(z^t)$. We impose the following assumption on this estimate.

Algorithm 1 Anchor-Regularized FTRL (AR-FTRL)

Input: regularization strength $\mu > 0$, learning rate $\eta > 0$, and epoch lengths $(N_k)_{k=1}^K$
Initialize $T_0 \leftarrow 0$, $r_x^1 \leftarrow u_x$, and $r_y^1 \leftarrow u_y$
for $k = 1, \dots, K$ **do**
  $T_k \leftarrow T_{k-1} + N_k$, $x_{k,1} \leftarrow r_x^k$, and $y_{k,1} \leftarrow r_y^k$
  for $s = 1, \dots, N_k$ **do**
    Each player observes its feedback component $\widehat{F}_x^{k,s}$ or $\widehat{F}_y^{k,s}$
    Update $x_{k,s+1}$ and $y_{k,s+1}$ via (3)
  end for
  $r_x^{k+1} \leftarrow x_{k,N_k+1}$ and $r_y^{k+1} \leftarrow y_{k,N_k+1}$
end for
Play $z^{T_K+1} \leftarrow r^{K+1}$

Assumption 5.1. *The noisy feedback sequence $\{\widehat{F}^t\}_{t\geq 1}$ satisfies, for every $t \geq 1$: (a)* Unbiasedness*: $\mathbb{E}[\widehat{F}^t \mid \mathcal{H}_t] = F(z^t)$; and (b)* Bounded second moment*: $\mathbb{E}[\|\widehat{F}^t\|^2 \mid \mathcal{H}_t] \leq G^2$ for some constant $G > 0$.*

This section establishes that AR-FTRL attains an $\widetilde{\mathcal{O}}(T^{-1/2})$ last-iterate convergence rate in expectation under noisy gradient feedback. Section 5.1 states the rate (Theorem 5.3) under a problem-dependent geometric epoch schedule. Section 5.2 outlines the proof. The argument has two ingredients: a per-anchor contraction of the anchor-regularized equilibrium toward $\mathcal{Z}^*$, and an inner-loop bound showing that each epoch tracks its anchor-regularized equilibrium. The per-anchor contraction rests on the metric subregularity of F on polytopes (Lemma 5.2). The inner-loop bound instead uses the strong monotonicity of the anchor-regularized operator and the bounded second moment of the feedback. Section 5.3 presents a variant (Theorem 5.4) that uses a problem-independent schedule and recovers the same $\widetilde{\mathcal{O}}(T^{-1/2})$ rate. All proofs are deferred to Appendix A.

Before stating the rate, we record the structural property of bilinear saddle-point problems over polytopes on which the analysis relies. The gradient operator F is metrically subregular at every Nash equilibrium.

Lemma 5.2 (Wei et al., 2021, Theorem 5)**.** *There exists a constant $\tau > 0$, depending only on A, $\mathcal{X}$, and $\mathcal{Y}$, such that for every $z \in \mathcal{Z} \setminus \mathcal{Z}^*$,*

$$\sup_{\substack{z' \in \mathcal{Z} \\ z' \neq z}} \frac{\langle F(z),\, z - z' \rangle}{\|z' - z\|} \;\geq\; \tau \cdot \operatorname{dist}(z, \mathcal{Z}^*).$$

5.1 Geometric epoch schedule

Theorem 5.3 (Last-iterate rate under noisy gradient feedback)**.** *Suppose Assumption 5.1 holds. Run AR-FTRL (Algorithm 1) with regularization strength $\mu > 0$, fixed learning rate $\eta > 0$, any number of epochs $K \geq 1$, and epoch lengths $N_k = \lceil N_0 \beta^{k-1} \rceil$ for $k = 1, \dots, K$, where $N_0 \geq 1$ and $1 < \beta < (\mu^2 + \tau^2)/\mu^2$. Then, for every $k = 1, \dots, K$,*

$$\left(\mathbb{E}\left[\mathrm{GAP}(z^{T_k+1})^2\right]\right)^{1/2} \;\leq\; \widetilde{\mathcal{O}}\left(T_k^{-1/2}\right).$$

The last strategy $z^{T_k+1} = r^{k+1}$ is computed from the first T_k noisy gradient observations and is played at iteration $T_k + 1$.

The suppressed constant is given explicitly in Appendix A. It depends only on R_x, R_y, $\|A\|$, N_0, β, μ, τ, $\operatorname{dist}(r^1, \mathcal{Z}^*)$, and the inner-loop constant C_{inner} of Lemma A.4.

Since $T_K = \sum_{k=1}^K N_k = T$ is the total number of observations after K epochs, Theorem 5.3 with $k = K$ yields the $\widetilde{\mathcal{O}}(T^{-1/2})$ rate for the last strategy z^{T+1}. To the best of our knowledge, this is the first last-iterate rate of this order in a constrained bilinear game under noisy gradient feedback. The rate matches the canonical $T^{-1/2}$ rate for the average iterate in stochastic saddle-point optimization (Nemirovski et al., 2009), now for the last strategy.

5.2 Proof sketch

For an anchor $r \in \mathcal{Z}$, let $z^{\mu,r}$ denote the unique saddle point of the anchor-regularized game (2), the *anchor-regularized equilibrium* for r. Each epoch runs the log-barrier FTRL inner loop on a fixed anchor and returns its last iterate as the next anchor. The analysis shows that these anchors contract toward $\mathcal{Z}^*$ across epochs. The argument has three steps.

First, combining the first-order optimality condition of (2) for $z^{\mu,r}$ with Lemma 5.2 shows that $z^{\mu,r}$ lies strictly closer to $\mathcal{Z}^*$ than the anchor r. This yields the per-anchor contraction

$$\operatorname{dist}(z^{\mu,r},\ \mathcal{Z}^*)\ \le\ (\mu/\sqrt{\mu^2+\tau^2})\operatorname{dist}(r,\mathcal{Z}^*),$$

whose factor depends only on μ and τ.

Second, within epoch k, the log-barrier FTRL updates run on the fixed anchor r^k. Their last iterate $r^{k+1} = z^{T_k+1}$ tracks the anchor-regularized equilibrium z^{μ,r^k}. An analysis of these updates gives

$$\mathbb{E}\left[\left\|r^{k+1} - z^{\mu,r^k}\right\|^2\right]\ \le\ \widetilde{\mathcal{O}}(1/N_k).$$

Finally, combining this inner-loop bound with the per-anchor contraction through the triangle and Minkowski inequalities yields the outer-loop recursion

$$\sqrt{\mathbb{E}[\operatorname{dist}(r^{k+1},\mathcal{Z}^*)^2]}\ \le\ \frac{\mu}{\sqrt{\mu^2+\tau^2}}\sqrt{\mathbb{E}[\operatorname{dist}(r^k,\mathcal{Z}^*)^2]} + \widetilde{\mathcal{O}}\Big(N_k^{-1/2}\Big).$$

The factor $\mu/\sqrt{\mu^2+\tau^2} < 1$ exposes a trade-off in μ. A smaller μ tightens the contraction but slows the inner loop. Under the geometric schedule $N_k = \lceil N_0\beta^{k-1}\rceil$, the condition $\beta < (\mu^2+\tau^2)/\mu^2$ makes the unrolled recursion a convergent series. This series yields the $\widetilde{\mathcal{O}}(T^{-1/2})$ distance rate. Converting this distance rate to a gap then yields Theorem 5.3. The complete argument is given in Appendix A.

5.3 Problem-independent epoch schedule

The admissible range for β in Theorem 5.3 depends on the subregularity constant τ, which is typically unknown to the algorithm. We remove this dependence by replacing the geometric schedule with the problem-independent *sub-geometric* schedule $N_k = \lceil \exp(\sqrt{k+1})\rceil$.

Theorem 5.4 (Last-iterate rate with a problem-independent schedule)**.** *Suppose Assumption 5.1 holds. Run AR-FTRL (Algorithm 1) with regularization strength $\mu > 0$, fixed learning rate $\eta > 0$, any number of epochs $K \ge 1$, and epoch lengths $N_k = \lceil \exp(\sqrt{k+1})\rceil$. Then, for every $k = 1, \dots, K$,*

$$\left(\mathbb{E}\left[\operatorname{GAP}(z^{T_k+1})^2\right]\right)^{1/2}\ \le\ \widetilde{\mathcal{O}}\Big(T_k^{-1/2}\Big).$$

The last strategy $z^{T_k+1} = r^{k+1}$ is computed from the first T_k noisy gradient observations and is played at iteration $T_k + 1$.

The suppressed constant is given explicitly in Appendix B. It depends only on R_x, R_y, $\|A\|$, μ, τ, $\operatorname{dist}(r^1,\mathcal{Z}^*)$, and the inner-loop constant C_{inner} of Lemma A.4.

The schedule in Theorem 5.4 contains no problem-dependent quantity, so AR-FTRL can be instantiated without any knowledge of τ. Theorem 5.4 therefore recovers the $\widetilde{\mathcal{O}}(T^{-1/2})$ last-iterate convergence rate of Theorem 5.3 without any problem-dependent tuning. The proof, deferred to Appendix B, parallels the argument above with a refined accounting of the sub-geometric epoch lengths.

6 Main Result: Bandit Feedback

Section 5 analyzed AR-FTRL under noisy gradient feedback. We now establish our main result under the more restrictive *bandit feedback* model for a two-player zero-sum matrix game. At each round, the two players sample one action each from their current strategies and observe only the resulting

scalar payoff. Neither player observes the payoff matrix A or a full gradient vector. From this scalar payoff, each player forms an importance-weighted estimate of its gradient. AR-FTRL then runs on these estimates.

In contrast to methods that restrict the iterates to a simplex bounded away from the boundary, the method below does not require a known lower bound on the equilibrium coordinates. The log-barrier regularizer keeps every sampled strategy strictly positive, while a combination of global and local bounds handles the possibility that an exact anchor-regularized equilibrium lies on the boundary.

6.1 Feedback model

At inner time s of epoch k, let $\mathcal{H}_{k,s-1}$ denote the history up to and including the choice of the strategy $z_{k,s} = (x_{k,s}, y_{k,s})$, which is the global history $\mathcal{H}_{T_{k-1}+s}$ of Section 3 at the corresponding overall iteration. We impose the following assumptions.

Assumption 6.1 (Bounded bandit feedback)**.** *Conditionally on $\mathcal{H}_{k,s-1}$, the players independently sample actions according to*

$$\Pr(I_{k,s} = i \mid \mathcal{H}_{k,s-1}) = x_{k,s,i}, \qquad \Pr(J_{k,s} = j \mid \mathcal{H}_{k,s-1}) = y_{k,s,j}.$$

After sampling the actions, the players observe a scalar payoff $R_{k,s}$. There is a known constant $B > 0$ such that $|R_{k,s}| \le B$ almost surely and

$$\mathbb{E}[R_{k,s} \mid \mathcal{H}_{k,s-1}, I_{k,s} = i, J_{k,s} = j] = A_{ij}.$$

Assumption 6.2 (Unique fully mixed equilibrium)**.** *The game has a unique Nash equilibrium $z^* = (x^*, y^*)$ with $\rho_* := \min_\ell z^*_\ell > 0$. The margin ρ_* is unknown to the algorithm.*

The analysis under bandit feedback uses the same metric subregularity constant τ as Lemma 5.2. The uniqueness part of Assumption 6.2 identifies distance to the equilibrium set with distance to z^*. Its full-mixedness part supplies the minimum equilibrium coordinate ρ_*.

The importance-weighted estimator is

$$\widehat{G}^x_{k,s} := \frac{R_{k,s}}{x_{k,s,I_{k,s}}} e_{I_{k,s}}, \qquad \widehat{G}^y_{k,s} := -\frac{R_{k,s}}{y_{k,s,J_{k,s}}} e_{J_{k,s}}. \tag{6}$$

The barrier ensures positive sampling probabilities. We write $\widehat{G}_{k,s} := (\widehat{G}^x_{k,s}, \widehat{G}^y_{k,s})$ for the joint estimator. Direct conditioning gives $\mathbb{E}[\widehat{G}_{k,s} \mid \mathcal{H}_{k,s-1}] = F(z_{k,s})$. No uniform Euclidean second-moment bound is assumed for $\widehat{G}_{k,s}$.

6.2 Instantiation and convergence guarantee

Under bandit feedback, AR-FTRL in Algorithm 1 uses

$$\widehat{F}^{k,s}_x := \widehat{G}^x_{k,s}, \qquad \widehat{F}^{k,s}_y := \widehat{G}^y_{k,s}, \qquad N_k := \left\lceil \exp(\sqrt{k+1}) \right\rceil. \tag{7}$$

The cumulative horizon and global strategy sequence are those of Section 4. The boundary identity $z^{T_k+1} = r^{k+1} = z_{k+1,1}$ is (5). Thus, each player updates using only its own sampled action, sampling probability, scalar payoff, and anchor. The joint history is used only in the analysis.

Perhaps surprisingly, bandit feedback suffices for the last strategy to attain the same $\widetilde{\mathcal{O}}(T^{-1/2})$ rate as under noisy gradient feedback. This holds even though the importance-weighted estimator has unbounded variance and the algorithm is given neither the subregularity constant τ nor the minimum equilibrium coordinate ρ_*.

Theorem 6.3 (Last-iterate rate under bandit feedback)**.** *Suppose Assumptions 6.1 and 6.2 hold. Fix $\mu > 0$ and $\eta > 0$ such that $\eta B < 1$. For any $K \ge 1$, run the instantiation (7) of Algorithm 1 through the first round of epoch $K+1$. Then, for every $k = 1, \dots, K$,*

$$\left(\mathbb{E}[\mathrm{GAP}(z^{T_k+1})^2]\right)^{1/2} \le \widetilde{\mathcal{O}}\left(T_k^{-1/2}\right).$$

Each strategy $z^{T_k+1} = r^{k+1}$ is computed from the first T_k scalar payoff observations and is played at round $T_k + 1$. Its feedback is the first observation used in epoch $k+1$. Neither τ nor ρ_* is supplied to the algorithm.

The suppressed constant is given explicitly in Appendix C. It depends only on R_x, R_y, m, n, A, μ, B, η, τ, and ρ_*.

In particular, setting $T := T_K$ gives an $\widetilde{\mathcal{O}}(T^{-1/2})$ rate for the L^2 norm of the gap of the last strategy z^{T+1} under bandit feedback. To the best of our knowledge, this is the first last-iterate rate of this order from bandit feedback in this setting. For the strategy played at every round t, the analysis of Cai et al. (2023) implies an $\widetilde{\mathcal{O}}(t^{-1/6})$ L^2-gap rate. Fiegel et al. (2025) establish an $\widetilde{\mathcal{O}}(t^{-1/4})$ L^2-gap rate for the same guarantee scope. Our guarantee instead controls the L^2 norm of the gap of the last strategy at a prescribed sequence of times, and attains the faster $\widetilde{\mathcal{O}}(T^{-1/2})$ rate at those times. It gives no rate for the intermediate strategies within an epoch.

6.3 Proof sketch

The main difficulty is that the importance-weighted estimator (6) has unbounded variance, so the bounded-second-moment inner-loop bound of Section 5 no longer applies. The outer argument still follows the warm-up of Section 5.2, in which the per-anchor contraction drives a recursion on the anchors. The new work is the inner-loop analysis, which replaces that single bound by two, namely a global bound that holds even when the anchor-regularized equilibrium lies on the boundary of $\mathcal{Z}$ and a sharper local bound for an interior target. The argument has three steps. Steps (1) and (2) fix an anchor r and bound the inner iterates $z_1, \dots, z_{N+1}$ started at $z_1 = r$.

(1) Global inner-loop bound. The log-barrier keeps every sampling probability positive. The importance-weighted increment itself is not uniformly bounded. When $\eta B < 1$, however, the log-barrier bounds its effect on the FTRL step in both directions. Comparing the iterate with the interior point $p^\gamma = (1-\gamma)z^{\mu,r} + \gamma u$ for the uniform profile u and $\gamma = 1/(N+1)$, which keeps the barrier finite even for a boundary target, gives

$$\mathbb{E}\big[\|z_{N+1} - z^{\mu,r}\|^2\big] \le \widetilde{\mathcal{O}}(1/(\mu\sqrt{N})).$$

(2) Local inner-loop bound. When $\min_\ell [z^{\mu,r}]_\ell \ge \rho$, the barrier gradient at the target is bounded. Splitting the coordinates by their proximity to $z^{\mu,r}$ lets the strongly monotone drift absorb the deviating coordinates, which sharpens the bound to

$$\mathbb{E}\big[\|z_{N+1} - z^{\mu,r}\|^2\big] \le \widetilde{\mathcal{O}}(C_\ell(\rho)/(\mu N)).$$

Here, the constant $C_\ell(\rho)$ depends on the lower bound ρ on the target coordinates and increases as ρ decreases.

(3) Contraction and the anchor recursion. The per-anchor contraction gives $\|z^{\mu,r} - z^*\| \le q\|r - z^*\|$ with $q = \mu/\sqrt{\mu^2+\tau^2} < 1$. With $\delta_* = \rho_*/2$, the local bound of Step (2) applies on $\{\|r^k - z^*\| \le \delta_*\}$, and the global bound of Step (1) applies on its complement. At $r = r^k$ and $N = N_k$, the inner loop ends at $z_{N+1} = r^{k+1}$. Combining the two cases with the contraction and taking expectation yields the anchor recursion

$$\mathbb{E}\|r^{k+1} - z^*\|^2 \le \Big(\bar{q}^2 + \widetilde{\mathcal{O}}(1/(\delta_*^2 \mu\sqrt{N_k}))\Big)\, \mathbb{E}\|r^k - z^*\|^2 + \widetilde{\mathcal{O}}(1/(\mu N_k)),$$

where $\bar{q} \in (q, 1)$. The coefficient inflation combines the global bound of Step (1) with Markov's inequality on the complementary event. The additive term is the local bound of Step (2) at $N = N_k$. Both terms vanish as k grows, since $N_k \to \infty$. Solving the recursion and converting the epoch index to the sample count give

$$\mathbb{E}\|r^{k+1} - z^*\|^2 \le \widetilde{\mathcal{O}}(1/T_k).$$

The boundary identity (5) and a gap-to-distance conversion then give Theorem 6.3. The complete argument is given in Appendix C.

7 Numerical Experiments

In this section, we empirically evaluate AR-FTRL on two-player zero-sum matrix games under the bandit feedback model of Section 6. We compare it against the method of Cai et al. (2023), the

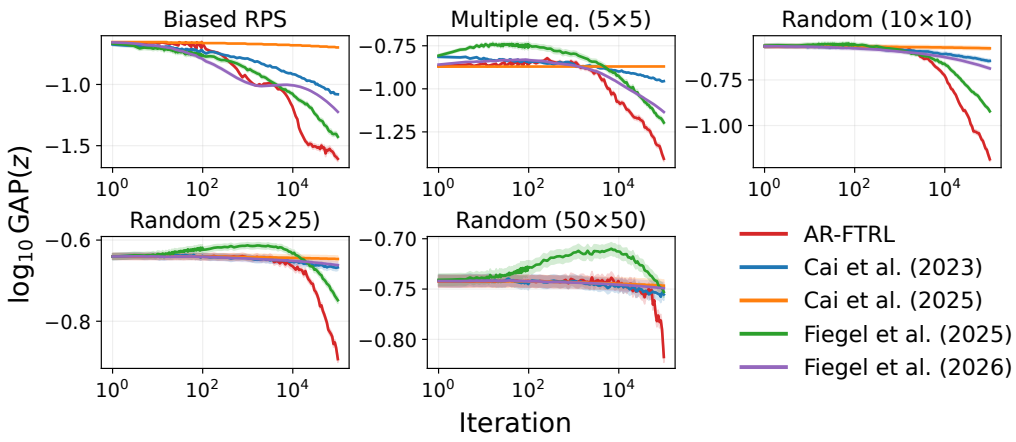


Figure 1: Gap under bandit feedback at horizon $T = 10^5$, on a base-10 logarithmic scale against the iteration count. Each curve is the mean over 100 evaluation seeds, and the shaded area represents the standard error.

average-to-last-iterate reduction applied to optimistic multiplicative weights (A2LOMWU) of Cai et al. (2025), the simultaneous explore-or-exploit method (SEE) of Fiegel et al. (2025), and the log-barrier regularized mirror descent (RMD-LB) of Fiegel et al. (2026). The warm-up experiments under noisy gradient feedback, which corroborate Theorems 5.3 and 5.4, are deferred to Appendix F.

We conduct experiments on five games whose payoff entries lie in $[0, 1]$. The first is the biased Rock–Paper–Scissors game, whose unique Nash equilibrium is the interior profile $((1/5, 3/5, 1/5), (1/5, 3/5, 1/5))$. This game satisfies the unique fully mixed equilibrium condition of Assumption 6.2. The second is the 5×5 multiple-equilibria game of Wei et al. (2021). The remaining three are dense random games of dimension $d \in \{10, 25, 50\}$, whose entries are drawn independently from the uniform distribution on $[0, 1]$. The multiple-equilibria game and the three random games need not have a unique fully mixed equilibrium, and we include them to examine AR-FTRL beyond the hypotheses of Theorem 6.3. All runs start from uniform initial strategies, and the payoff matrices of the two structured games are given in Appendix E.1.

Each run has a budget of $T = 10^5$ payoff observations. For every method, we evaluate the played strategy through the gap function $\mathrm{GAP}(z)$ of Section 3. For AR-FTRL, we use the learning rate $\eta = 1/10$, the regularization strength $\mu = 1$, and the problem-independent sub-geometric schedule $N_k = \lceil \exp(\sqrt{k+1}) \rceil$ of Theorem 5.4. These three settings are fixed across all games and both feedback models. The baseline hyperparameters and the tuning and evaluation protocol are given in Appendix E.2.

Figure 1 illustrates the gap under bandit feedback. At the horizon $T = 10^5$, AR-FTRL attains the lowest gap on all five games, with SEE the closest competitor. The advantage holds even on the multiple-equilibria and random games, which lie outside the hypotheses of Theorem 6.3.

8 Conclusion and Limitations

We introduced AR-FTRL for stochastic two-player zero-sum matrix games, which runs log-barrier FTRL on an anchor-regularized game over each epoch and resets the anchor to the last iterate at every epoch boundary. Under noisy gradient feedback, the anchor resets contract the iterate toward the set of Nash equilibria and attain an $\widetilde{\mathcal{O}}(T^{-1/2})$ last-iterate rate. Our main result attains the same rate from bandit feedback in games with a unique fully mixed Nash equilibrium, despite the unbounded variance of the importance-weighted gradient estimate. In both settings, the epoch schedule is problem-independent and sub-geometric, and the algorithm requires neither the metric-subregularity constant nor the minimum equilibrium coordinate. Its main limitation is the unique fully mixed equilibrium assumption of the bandit result, whose removal we leave to future work.

AI USE STATEMENT

We used generative AI tools for three purposes. First, we used them to aid and polish the writing, improving the prose and the clarity of exposition in the statements and proofs. Second, we used them to assist in developing and checking the proofs of our theoretical results. Third, we used them to assist in implementing the experiment code. We did not use generative AI tools for research ideation or to draft sections of the paper. The authors verified all definitions, theorems, and proofs, and take full responsibility for their correctness. We have reviewed all AI-assisted text, proofs, and code. We take responsibility for the final content of this work, including all text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

The problem setting and the assumptions are stated in Section 3, and AR-FTRL is defined in full in Section 4. The last-iterate guarantees under noisy gradient feedback and under bandit feedback are stated in Sections 5 and 6, respectively. Complete proofs, including the conditioning, the constants, and the conversion from epoch lengths to the number of feedback observations, are given in Appendices A, B, and C. The experimental setup, including the benchmark games, the hyperparameters, and the initialization, is described in Section 7, and the payoff matrices of the structured games are listed in Appendix E.1, with the tuning and evaluation protocol given in Appendix E.2. Source code to reproduce all experiments is included in the 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 *AISTATS*, pp. 7999–8028, 2023.

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. Adaptively perturbed mirror descent for learning in games. In *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 *ICLR*, 2025.

Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities. In *COLT*, pp. 326–358, 2021.

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

Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave N-person games. In *NeurIPS*, pp. 5666–5676, 2018.

Yang Cai and Weiqiang Zheng. Doubly optimal no-regret learning in monotone games. In *ICML*, pp. 3507–3524, 2023.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *NeurIPS*, pp. 33904–33919, 2022a.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Tight last-iterate convergence of the extragradient method for constrained monotone variational inequalities. *arXiv preprint arXiv:2204.09228*, 2022b.

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 *NeurIPS*, pp. 36364–36406, 2023.

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 *NeurIPS*, 2025.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. In *NeurIPS*, pp. 27952–27964, 2021.

Shicong Cen, Yuejie Chi, Simon S Du, and Lin Xiao. Faster last-iterate convergence of policy optimization in zero-sum Markov games. In *ICLR*, 2023.

Volkan Cevher, Georgios Piliouras, Ryann Sim, and Stratis Skoulakis. Min-max optimization made simple: Approximating the proximal point method via contraction maps. In *Symposium on Simplicity in Algorithms (SOSA)*, pp. 192–206, 2023.

Zaiwei Chen and Eric Mazumdar. Last-iterate convergence for generalized Frank–Wolfe in monotone variational inequalities. In *NeurIPS*, 2024.

Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. In *COLT*, volume 99, pp. 807–834, 2019.

Pierre Coucheney, Bruno Gaujal, and Panayotis Mertikopoulos. Penalty-regulated dynamics and robust learning procedures in games. *Mathematics of Operations Research*, 40(3):611–633, 2015.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *ITCS*, pp. 27:1–27:18, 2019.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *ICLR*, 2018.

Étienne de Montbrun and Jérôme Renault. Convergence of optimistic gradient descent ascent in bilinear games. *arXiv preprint arXiv:2208.03085*, 2022.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

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 *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. *arXiv preprint arXiv:2604.15242*, 2026.

Angeliki Giannou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. In *NeurIPS*, pp. 22655–22666, 2021.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR*, 2019.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *NeurIPS*, pp. 20766–20778, 2020a.

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

Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In *NeurIPS*, pp. 21858–21870, 2022.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS*, pp. 6938–6948, 2019.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *NeurIPS*, pp. 16223–16234, 2020.

Aamal Abbas Hussain, Francesco Belardinelli, and Georgios Piliouras. Asymptotic convergence and performance of multi-agent q-learning dynamics. In *AAMAS*, pp. 1578–1586, 2023.

Shinji Ito, Haipeng Luo, Taira Tsuchiya, and Yue Wu. Instance-dependent regret bounds for learning two-player zero-sum games with bandit feedback. In *Proceedings of the Thirty-Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2858–2892, 2025. URL https://proceedings.mlr.press/v291/ito25a.html.

Shinji Ito, Taira Tsuchiya, Kaito Ariu, and Kenshi Abe. Last-iterate convergence of regularized gradient methods for stochastic monotone variational inequalities. In *ICML*, 2026.

Jayash Koshal, Angelia Nedić, and Uday V Shanbhag. Single timescale regularized stochastic approximation schemes for monotone nash games under uncertainty. In *CDC*, pp. 231–236. IEEE, 2010.

Jayash Koshal, Angelia Nedic, and Uday V. Shanbhag. Regularized iterative stochastic approximation methods for stochastic variational inequality problems. *IEEE Transactions on Automatic Control*, 58(3):594–609, 2013.

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. *NeurIPS*, pp. 14293–14305, 2021.

Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *NeurIPS*, pp. 22588–22600, 2021.

Qi Lei, Sai Ganesh Nagarajan, Ioannis Panageas, et al. Last iterate convergence in no-regret learning: constrained min-max optimization for convex-concave landscapes. In *AISTATS*, pp. 1441–1449, 2021.

David S Leslie and Edmund J Collins. Individual q-learning in normal form games. *SIAM Journal on Control and Optimization*, 44(2):495–514, 2005.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *AISTATS*, pp. 907–915, 2019.

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

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1):465–507, 2019.

Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *SODA*, pp. 2703–2717, 2018.

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

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *AISTATS*, pp. 1497–1507, 2020.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Sarath Pattathil, Kaiqing Zhang, and Asuman Ozdaglar. Symmetric (optimistic) natural policy gradient for multi-agent learning with parameter convergence. In *AISTATS*, pp. 5641–5685, 2023.

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 *ICML*, pp. 8525–8535, 2021.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *COLT*, pp. 993–1019, 2013a.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *NeurIPS*, pp. 3066–3074, 2013b.

Samuel Sokota, Ryan D'Orazio, J Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *ICLR*, 2023.

Tatiana Tatarenko and Maryam Kamgarpour. Learning Nash equilibria in monotone games. In *CDC*, pp. 3104–3109. IEEE, 2019.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *ICLR*, 2021.

TaeHo Yoon and Ernest K Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. In *ICML*, pp. 12098–12109, 2021.

Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. *Mathematical Programming*, 165:391–431, 2017.

A Proofs for Theorem 5.3

This appendix proves Theorem 5.3. Each auxiliary lemma is stated inline within the proof that first requires it, and its own proof is given in a dedicated subsection that follows. The anchor-regularized equilibrium $z^{\mu,r}$ is the unique saddle point of the anchor-regularized game (2). It satisfies the first-order optimality condition

$$\langle F(z^{\mu,r}) + \mu(z^{\mu,r} - r),\, z - z^{\mu,r}\rangle \geq 0, \qquad z \in \mathcal{Z}. \tag{8}$$

Throughout this appendix, we write $\|v\|_\mu := \sqrt{\mu}\,\|v\|$, so that $\|v\|_\mu^2 = \mu\,\|v\|^2$, and $D_\mu := \max_{z,w\in\mathcal{Z}} \|z - w\|_\mu$ for the diameter of $\mathcal{Z}$ in this norm. The uniform point $u := (\mathbf{1}_m/m, \mathbf{1}_n/n)$ minimizes the log-barrier $h(z) := -\sum_{i=1}^m \log x_i - \sum_{j=1}^n \log y_j$. For an anchor $r \in \mathcal{Z}$, we write $F_r(z) := F(z) + \mu(z - r)$ for the perturbed operator and $q_r(z) := \|z - r\|_\mu^2/2$ for the anchor term. We write $D_h(a, b) := h(a) - h(b) - \langle \nabla h(b), a - b\rangle$ for the Bregman divergence of the log-barrier and $D_q(a, b) := \|a - b\|_\mu^2/2$ for the squared distance.

A.1 Proof of Theorem 5.3

Proof of Theorem 5.3. We first establish a recursion on the squared distance from the anchor to $\mathcal{Z}^*$ across epochs.

Lemma A.1. *Suppose Assumption 5.1 holds, and run AR-FTRL (Algorithm 1) with regularization strength $\mu > 0$ and fixed learning rate $\eta > 0$. Then, for every $k = 1, \ldots, K$,*

$$\sqrt{\mathbb{E}\left[\operatorname{dist}(r^{k+1}, \mathcal{Z}^*)^2\right]} \leq \frac{\mu}{\sqrt{\mu^2 + \tau^2}} \sqrt{\mathbb{E}\left[\operatorname{dist}(r^k, \mathcal{Z}^*)^2\right]} + \sqrt{C_{\text{inner}} \frac{\log(eN_k)}{N_k}},$$

where C_{inner} is the constant of Lemma A.4.

Write $q := \mu/\sqrt{\mu^2 + \tau^2}$. Iterating Lemma A.1 gives, for any epoch schedule,

$$\sqrt{\mathbb{E}[\operatorname{dist}(r^{k+1}, \mathcal{Z}^*)^2]} \leq q^k \operatorname{dist}(r^1, \mathcal{Z}^*) + \sqrt{C_{\text{inner}}} \sum_{j=1}^k q^{k-j} \sqrt{\frac{\log(eN_j)}{N_j}}. \tag{9}$$

Here $r^1 = u$ is deterministic by Algorithm 1.

We next bound the convolution term in (9). Since $N_j = \lceil N_0 \beta^{j-1} \rceil$, we have $N_0 \beta^{j-1} \le N_j \le (N_0+1)\beta^{j-1}$, so

$$\sqrt{\frac{\log(eN_j)}{N_j}} \le \frac{\beta^{-(j-1)/2}}{\sqrt{N_0}} \sqrt{\log(e(N_0+1)) + (j-1)\log\beta}$$
$$\le \frac{\beta^{-(j-1)/2}}{\sqrt{N_0}} \left(\sqrt{\log(e(N_0+1))} + \sqrt{\log\beta}\sqrt{j-1} \right),$$

where the second inequality uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. Substituting $i = k - j$, we obtain

$$\sum_{j=1}^{k} q^{k-j} \sqrt{\frac{\log(eN_j)}{N_j}} \le \frac{\beta^{-(k-1)/2}}{\sqrt{N_0}} \sum_{i=0}^{k-1} (q\sqrt{\beta})^i \left(\sqrt{\log(e(N_0+1))} + \sqrt{\log\beta}\sqrt{k-1-i} \right)$$
$$\le \frac{\beta^{-(k-1)/2}}{\sqrt{N_0}} \frac{\sqrt{\log(e(N_0+1))} + \sqrt{\log\beta}\sqrt{k-1}}{1 - q\sqrt{\beta}}, \tag{10}$$

where $q\sqrt{\beta} < 1$ follows from $\beta < (\mu^2 + \tau^2)/\mu^2$.

Since $\beta > 1$, the upper bound on N_j also gives

$$T_k \le (N_0+1) \sum_{j=1}^{k} \beta^{j-1} = \frac{(N_0+1)(\beta^k - 1)}{\beta - 1} \le \frac{(N_0+1)\beta^k}{\beta - 1},$$
$$\beta^{-(k-1)/2} \le \sqrt{\frac{\beta(N_0+1)}{\beta-1}} \frac{1}{\sqrt{T_k}}.$$

Moreover, $T_k \ge N_k \ge N_0 \beta^{k-1}$ and $N_0 \ge 1$ imply $\sqrt{\log\beta}\sqrt{k-1} \le \sqrt{\log(eT_k)}$. Substituting these bounds into (10) and using $\log(eT_k) \ge 1$ yields

$$\sqrt{C_{\text{inner}}} \sum_{j=1}^{k} q^{k-j} \sqrt{\frac{\log(eN_j)}{N_j}}$$
$$\le \frac{\sqrt{C_{\text{inner}}/N_0} \left(\sqrt{\log(e(N_0+1))} + 1 \right)}{1 - q\sqrt{\beta}} \sqrt{\frac{\beta(N_0+1)}{\beta-1}} \sqrt{\frac{\log(eT_k)}{T_k}}. \tag{11}$$

For the first term in (9), $q < \beta^{-1/2}$ gives

$$q^k < \beta^{-k/2} \le \sqrt{\frac{N_0+1}{\beta-1}} \frac{1}{\sqrt{T_k}} \le \sqrt{\frac{N_0+1}{\beta-1}} \sqrt{\frac{\log(eT_k)}{T_k}}.$$

Combining this with (9) and (11), we have

$$\sqrt{\mathbb{E}[\operatorname{dist}(r^{k+1}, \mathcal{Z}^*)^2]} \le C_{\text{dist}} \sqrt{\frac{\log(eT_k)}{T_k}},$$

where

$$C_{\text{dist}} := \sqrt{\frac{N_0+1}{\beta-1}} \operatorname{dist}(r^1, \mathcal{Z}^*)$$
$$+ \frac{\sqrt{C_{\text{inner}}/N_0} \left(\sqrt{\log(e(N_0+1))} + 1 \right)}{1 - q\sqrt{\beta}} \sqrt{\frac{\beta(N_0+1)}{\beta-1}}.$$

Finally, we convert the distance bound to a gap bound via the following deterministic estimate.

Lemma A.2. *For all $z \in \mathcal{Z}$,*

$$\operatorname{GAP}(z) \le \left(\sqrt{R_x^2 + R_y^2} \|A\| \right) \operatorname{dist}(z, \mathcal{Z}^*).$$

By Lemma A.2 and (5),

$$\begin{aligned}\left(\mathbb{E}[\mathrm{GAP}(z^{T_k+1})^2]\right)^{1/2} &= \left(\mathbb{E}[\mathrm{GAP}(r^{k+1})^2]\right)^{1/2} \\ &\le \sqrt{R_x^2+R_y^2}\|A\|\sqrt{\mathbb{E}[\mathrm{dist}(r^{k+1},\mathcal{Z}^*)^2]} \\ &\le \sqrt{R_x^2+R_y^2}\|A\|\, C_{\mathrm{dist}}\sqrt{\frac{\log(eT_k)}{T_k}}.\end{aligned}$$

□

A.2 Proof of Lemma A.1

Proof of Lemma A.1. By the triangle inequality and Minkowski's inequality,

$$\sqrt{\mathbb{E}\left[\mathrm{dist}(r^{k+1},\mathcal{Z}^*)^2\right]} \le \sqrt{\mathbb{E}\left[\left\|r^{k+1}-z^{\mu,r^k}\right\|^2\right]} + \sqrt{\mathbb{E}\left[\mathrm{dist}(z^{\mu,r^k},\mathcal{Z}^*)^2\right]}, \tag{12}$$

where the first step uses $\mathrm{dist}(r^{k+1},\mathcal{Z}^*) \le \|r^{k+1}-\Pi_{\mathcal{Z}^*}(z^{\mu,r^k})\| \le \|r^{k+1}-z^{\mu,r^k}\|+\mathrm{dist}(z^{\mu,r^k},\mathcal{Z}^*)$. We bound the two terms in (12) by a per-anchor contraction and an inner-loop tracking bound, both stated below. The contraction shows that the anchor-regularized equilibrium lies closer to $\mathcal{Z}^*$ than its anchor.

Lemma A.3. *For every anchor $r \in \mathcal{Z}$ and every $\mu > 0$,*

$$\mathrm{dist}(z^{\mu,r},\,\mathcal{Z}^*) \;\le\; \frac{\mu}{\sqrt{\mu^2+\tau^2}}\,\mathrm{dist}(r,\mathcal{Z}^*).$$

The inner-loop tracking bound controls how closely the last iterate of an epoch approximates its anchor-regularized equilibrium.

Lemma A.4. *Suppose Assumption 5.1 holds, and run AR-FTRL (Algorithm 1) with regularization strength $\mu > 0$ and fixed learning rate $\eta > 0$. Then, for every $k = 1,\dots,K$,*

$$\mathbb{E}\left[\left\|r^{k+1}-z^{\mu,r^k}\right\|^2 \,\middle|\, \mathcal{H}_{T_{k-1}+1}\right] \le C_{\mathrm{inner}}\frac{\log(eN_k)}{N_k},$$

where

$$\begin{aligned}C_{\mathrm{init}} &:= \eta DG + \frac{\eta\mu}{2}D^2, \\ C_E &:= \frac{3\mu}{2}D^2 + \left(\|A\|\sqrt{R_x^2+R_y^2}+\mu D\right)D, \\ C_{\mathrm{inner}} &:= \frac{4(m+n)}{\eta\mu} + \frac{4C_{\mathrm{init}}}{\eta\mu} + \frac{4C_E}{\mu} + \frac{4(G^2+\mu^2D^2)}{\mu^2} + 2D^2,\end{aligned}$$

which depends only on m, n, η, μ, G, D, $\|A\|$, R_x, and R_y.

Since $r^{k+1} = z_{k,N_k+1}$ is the last update of epoch k initialized at r^k, Lemma A.4 gives $\mathbb{E}[\|r^{k+1}-z^{\mu,r^k}\|^2] \le C_{\mathrm{inner}}\log(eN_k)/N_k$ after taking total expectation. Combining this with Lemma A.3 applied to the anchor r^k, (12) becomes

$$\sqrt{\mathbb{E}\left[\mathrm{dist}(r^{k+1},\mathcal{Z}^*)^2\right]} \le \frac{\mu}{\sqrt{\mu^2+\tau^2}}\sqrt{\mathbb{E}\left[\mathrm{dist}(r^k,\mathcal{Z}^*)^2\right]} + \sqrt{C_{\mathrm{inner}}\frac{\log(eN_k)}{N_k}}.$$

□

A.3 Proof of Lemma A.3

Proof of Lemma A.3. Fix an anchor $r \in \mathcal{Z}$ and abbreviate $p := z^{\mu,r}$. We first record a Pythagorean bound. For every $z^* \in \mathcal{Z}^*$, the Nash equilibrium property gives

$$\langle F(z^*), z - z^* \rangle = \langle Ay^*, x - x^* \rangle + \langle -A^\top x^*, y - y^* \rangle \geq 0, \qquad z \in \mathcal{Z}.$$

Taking $z = z^*$ in (8) and $z = p$ in the preceding inequality, and adding,

$$\begin{aligned} 0 &\leq \langle F(p) + \mu(p - r), z^* - p \rangle + \langle F(z^*), p - z^* \rangle \\ &= \langle F(p) - F(z^*), z^* - p \rangle + \mu \langle p - r, z^* - p \rangle \\ &= \mu \langle p - r, z^* - p \rangle, \end{aligned}$$

where the last equality uses $\langle F(p) - F(z^*), z^* - p \rangle = -\langle F(p) - F(z^*), p - z^* \rangle = 0$ by skew-symmetry of F. Taking $z^* = \Pi_{\mathcal{Z}^*}(r)$, the inequality $\langle p - r, \Pi_{\mathcal{Z}^*}(r) - p \rangle \geq 0$ gives

$$\begin{aligned} \operatorname{dist}(r, \mathcal{Z}^*)^2 &= \|r - p\|^2 + \|p - \Pi_{\mathcal{Z}^*}(r)\|^2 + 2\langle r - p, p - \Pi_{\mathcal{Z}^*}(r) \rangle \\ &\geq \|r - p\|^2 + \|p - \Pi_{\mathcal{Z}^*}(r)\|^2 . \end{aligned} \tag{13}$$

We next lower-bound $\|r - p\|$. From (8), for every $z \in \mathcal{Z}$,

$$\langle F(p), p - z \rangle \leq \mu \langle p - r, z - p \rangle \leq \mu \|p - r\| \|z - p\|,$$

where the last step is the Cauchy–Schwarz inequality. Taking the supremum over $z \in \mathcal{Z}$ with $z \neq p$,

$$\sup_{\substack{z \in \mathcal{Z} \\ z \neq p}} \frac{\langle F(p), p - z \rangle}{\|z - p\|} \leq \mu \|p - r\|.$$

If $p \in \mathcal{Z} \setminus \mathcal{Z}^*$, Lemma 5.2 gives $\mu \|p - r\| \geq \tau \operatorname{dist}(p, \mathcal{Z}^*)$. If $p \in \mathcal{Z}^*$, the same inequality is trivial because $\operatorname{dist}(p, \mathcal{Z}^*) = 0$. Hence, in all cases,

$$\|r - p\| \geq \frac{\tau}{\mu} \operatorname{dist}(p, \mathcal{Z}^*). \tag{14}$$

Substituting (14) into (13) and using $\|p - \Pi_{\mathcal{Z}^*}(r)\| \geq \operatorname{dist}(p, \mathcal{Z}^*)$,

$$\operatorname{dist}(r, \mathcal{Z}^*)^2 \geq \frac{\tau^2}{\mu^2} \operatorname{dist}(p, \mathcal{Z}^*)^2 + \operatorname{dist}(p, \mathcal{Z}^*)^2 = \frac{\mu^2 + \tau^2}{\mu^2} \operatorname{dist}(p, \mathcal{Z}^*)^2.$$

Rearranging gives the claim. □

A.4 Proof of Lemma A.4

Proof of Lemma A.4. We fix one epoch and suppress the epoch index. Let $\mathcal{H}_0 := \mathcal{H}_{T_{k-1}+1}$ and let $\mathcal{H}_{s-1}$ denote the history just before the feedback at inner step s is observed, including the choice of z_s. Write $\mathbb{E}_s[\cdot] := \mathbb{E}[\cdot \mid \mathcal{H}_{s-1}]$. We write r for the anchor, $z_1 = r, \ldots, z_{N+1}$ for the iterates, $\widehat{F}_s$ for the feedback at z_s, and $p := z^{\mu,r}$. The anchor r is determined by $\mathcal{H}_0$. Each iterate z_s is determined by $\mathcal{H}_{s-1}$ through the update rule. The map $r \mapsto z^{\mu,r}$ is nonexpansive by the firm-nonexpansiveness assertion of Lemma C.4, which does not rely on the assumptions specific to bandit feedback. Hence, p is determined by $\mathcal{H}_0$.

Initialization. For a comparator $w \in \operatorname{ri} \mathcal{Z}$ determined by $\mathcal{H}_0$, define

$$V_s(w) := D_h(w, z_s) + (s - 1)\eta D_q(w, z_s). \tag{15}$$

The first cumulative update is

$$z_2 = \arg\min_{z \in \operatorname{ri} \mathcal{Z}} \left\{ h(z) + \eta \langle \widehat{F}_1, z \rangle + \eta q_r(z) \right\}.$$

Its first-order condition along the affine simplex constraints gives

$$\begin{aligned} V_2(w) &= h(w) - h(z_2) + \eta \langle \widehat{F}_1, w - z_2 \rangle + \eta \big(q_r(w) - q_r(z_2) \big) \\ &\leq D_h(w, u) + \eta D \|\widehat{F}_1\| + \frac{\eta \mu D^2}{2}. \end{aligned}$$

Here u minimizes h, $\langle \nabla h(u), w-u\rangle = 0$, $\|w - z_2\| \le D$, and $q_r(w) - q_r(z_2) \le q_r(w) \le \mu D^2/2$. Although z_2 depends on the first observation, this pathwise bound and conditional Cauchy–Schwarz give

$$\begin{aligned}\mathbb{E}[V_2(w) \mid \mathcal{H}_0] &\le D_h(w,u) + \eta D\sqrt{\mathbb{E}[\|\widehat{F}_1\|^2 \mid \mathcal{H}_0]} + \frac{\eta\mu D^2}{2} \\ &\le D_h(w,u) + C_{\text{init}},\end{aligned} \tag{16}$$

where $C_{\text{init}} = \eta DG + \eta\mu D^2/2$ by Assumption 5.1(b).

One-step inequality. For $s = 2, \dots, N$, both z_s and z_{s+1} are cumulative minimizers in (4). They lie in $\operatorname{ri}\mathcal{Z}$ because the barrier diverges at the relative boundary. Subtracting their first-order conditions along the affine constraints gives, for every $w \in \operatorname{ri}\mathcal{Z}$,

$$0 \le \big\langle \eta\big(\widehat{F}_s + \mu(z_s - r)\big) + \nabla h(z_{s+1}) - \nabla h(z_s) + s\eta\mu(z_{s+1} - z_s),\ w - z_{s+1}\big\rangle.$$

Using the three-point identities

$$\begin{aligned}\langle \nabla h(z_{s+1}) - \nabla h(z_s), w - z_{s+1}\rangle &= D_h(w, z_s) - D_h(w, z_{s+1}) - D_h(z_{s+1}, z_s), \\ \mu\langle z_{s+1} - z_s, w - z_{s+1}\rangle &= D_q(w, z_s) - D_q(w, z_{s+1}) - D_q(z_{s+1}, z_s),\end{aligned}$$

and $z_{s+1} - w = (z_s - w) - (z_s - z_{s+1})$, we obtain

$$V_{s+1}(w) \le V_s(w) + \eta D_q(w, z_s) - \eta\langle \widehat{F}_s + \mu(z_s - r), z_s - w\rangle + S_s,$$

where

$$S_s := \eta\langle \widehat{F}_s + \mu(z_s - r), z_s - z_{s+1}\rangle - D_h(z_{s+1}, z_s) - s\eta D_q(z_{s+1}, z_s).$$

Since w, r, z_s are determined by $\mathcal{H}_{s-1}$, Assumption 5.1(a) yields

$$\mathbb{E}_s[V_{s+1}(w)] \le V_s(w) + \eta\big[D_q(w, z_s) - \langle F_r(z_s), z_s - w\rangle\big] + \mathbb{E}_s[S_s]. \tag{17}$$

To bound the stability term, use $D_h \ge 0$ and complete the square:

$$\begin{aligned}S_s &\le \eta\langle \widehat{F}_s + \mu(z_s - r), z_s - z_{s+1}\rangle - \frac{s\eta\mu}{2}\|z_s - z_{s+1}\|^2 \\ &\le \frac{\eta}{2s\mu}\|\widehat{F}_s + \mu(z_s - r)\|^2.\end{aligned}$$

Thus, $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, $\|z_s - r\| \le D$, and Assumption 5.1(b) give

$$\mathbb{E}_s[S_s] \le \frac{\eta}{s\mu}(G^2 + \mu^2 D^2). \tag{18}$$

Interior comparator. Set $\gamma := 1/(N+1)$ and $p^\gamma := (1-\gamma)p + \gamma u \in \operatorname{ri}\mathcal{Z}$. This comparator is determined by $\mathcal{H}_0$. By skew-symmetry of F and (8),

$$\begin{aligned}\langle F_r(z_s), z_s - p\rangle &= \langle F_r(p), z_s - p\rangle + \mu\|z_s - p\|^2 \ge 2D_q(p, z_s), \\ \langle F_r(z_s), z_s - p^\gamma\rangle &= \langle F_r(z_s), z_s - p\rangle + \gamma\langle F_r(z_s), p - u\rangle \\ &\ge 2D_q(p, z_s) - \gamma\left(\|A\|\sqrt{R_x^2 + R_y^2} + \mu D\right) D.\end{aligned}$$

The second bound also uses $\|F_r(z_s)\| \le \|A\|\sqrt{R_x^2 + R_y^2} + \mu D$ and $\|p - u\| \le D$. Also, expanding the square and using $\|p - u\|, \|p - z_s\| \le D$ and $\gamma \le 1$, we have

$$\begin{aligned}\|p^\gamma - z_s\|^2 &= \|p - z_s\|^2 + 2\gamma\langle p - z_s, u - p\rangle + \gamma^2\|u - p\|^2 \\ &\le \|p - z_s\|^2 + 3\gamma D^2.\end{aligned}$$

Combining these inequalities gives

$$D_q(p^\gamma, z_s) - \langle F_r(z_s), z_s - p^\gamma\rangle \le -D_q(p, z_s) + \gamma C_E, \tag{19}$$

where $C_E = 3\mu D^2/2 + (\|A\|\sqrt{R_x^2 + R_y^2} + \mu D)D$.

Terminal bound. Substitute (19) into (17) with $w = p^\gamma$ and condition on $\mathcal{H}_0$. The tower property gives, for $s = 2, \dots, N$,

$$\begin{aligned}\mathbb{E}[V_{s+1}(p^\gamma) \mid \mathcal{H}_0] &= \mathbb{E}[\mathbb{E}_s[V_{s+1}(p^\gamma)] \mid \mathcal{H}_0] \\ &\le \mathbb{E}[V_s(p^\gamma) \mid \mathcal{H}_0] - \eta\mathbb{E}[D_q(p, z_s) \mid \mathcal{H}_0] + \eta\gamma C_E + \mathbb{E}[S_s \mid \mathcal{H}_0].\end{aligned}$$

Summing, discarding the nonpositive drift, and applying (16) and (18), we obtain

$$\begin{aligned}\mathbb{E}[V_{N+1}(p^\gamma) \mid \mathcal{H}_0] &\le D_h(p^\gamma, u) + C_{\text{init}} + N\eta\gamma C_E + \sum_{s=2}^{N} \mathbb{E}[S_s \mid \mathcal{H}_0] \\ &\le D_h(p^\gamma, u) + C_{\text{init}} + \eta C_E + \frac{\eta}{\mu}(G^2 + \mu^2 D^2)\log(eN).\end{aligned}$$

Here the sums are empty when $N = 1$, and we used $N\gamma \le 1$ and $\sum_{s=2}^{N} 1/s \le \log(eN)$. By (15), $V_{N+1}(p^\gamma) \ge N\eta\mu\|z_{N+1} - p^\gamma\|^2/2$, so

$$\begin{aligned}&\mathbb{E}[\|z_{N+1} - p^\gamma\|^2 \mid \mathcal{H}_0] \\ &\quad\le \frac{2}{N\eta\mu}\left(D_h(p^\gamma, u) + C_{\text{init}} + \eta C_E + \frac{\eta}{\mu}(G^2 + \mu^2 D^2)\log(eN)\right). \qquad (20)\end{aligned}$$

To bound the barrier term, write $p^{\gamma,x}$ and $p^{\gamma,y}$ for the two blocks of p^γ. Since $\sum_i (mp_i^{\gamma,x} - 1) = 0$ and $\sum_j (np_j^{\gamma,y} - 1) = 0$, the linear terms in the divergence cancel exactly. Moreover, $mp_i^{\gamma,x} \ge \gamma$ and $np_j^{\gamma,y} \ge \gamma$, giving

$$\begin{aligned}D_h(p^\gamma, u) &= -\sum_{i=1}^{m} \log(mp_i^{\gamma,x}) - \sum_{j=1}^{n} \log(np_j^{\gamma,y}) \\ &\le (m+n)\log(1/\gamma) = (m+n)\log(N+1). \qquad (21)\end{aligned}$$

Finally, $\|p^\gamma - p\| \le \gamma D$ implies $\|z_{N+1} - p\|^2 \le 2\|z_{N+1} - p^\gamma\|^2 + 2\gamma^2 D^2$. Combining this with (20) and (21), and using $\log(N+1) \le \log(eN)$, $\log(eN) \ge 1$, and $\gamma^2 \le \log(eN)/N$, yields

$$\begin{aligned}\mathbb{E}[\|z_{N+1} - p\|^2 \mid \mathcal{H}_0] &\le \left(\frac{4(m+n)}{\eta\mu} + \frac{4C_{\text{init}}}{\eta\mu} + \frac{4C_E}{\mu} + \frac{4(G^2 + \mu^2 D^2)}{\mu^2} + 2D^2\right)\frac{\log(eN)}{N} \\ &= C_{\text{inner}}\frac{\log(eN)}{N}.\end{aligned}$$

Identifying $z_{N+1} = r^{k+1}$, $p = z^{\mu,r^k}$, $N = N_k$, and $\mathcal{H}_0 = \mathcal{H}_{T_{k-1}+1}$ proves the stated conditional bound. □

A.5 Proof of Lemma A.2

Proof of Lemma A.2. From the definition of the gap function, for any $z \in \mathcal{Z}$ and $z^* \in \mathcal{Z}^*$,

$$\begin{aligned}\text{GAP}(z) &= \max_{\tilde{y}\in\mathcal{Y}} x^\top A\tilde{y} - (x^*)^\top Ay^* + (x^*)^\top Ay^* - \min_{\tilde{x}\in\mathcal{X}} (\tilde{x})^\top Ay \\ &= \max_{\tilde{y}\in\mathcal{Y}} x^\top A\tilde{y} - \max_{\tilde{y}\in\mathcal{Y}} (x^*)^\top A\tilde{y} + \min_{\tilde{x}\in\mathcal{X}} (\tilde{x})^\top Ay^* - \min_{\tilde{x}\in\mathcal{X}} (\tilde{x})^\top Ay \\ &\le \max_{\tilde{y}\in\mathcal{Y}} (x - x^*)^\top A\tilde{y} - \min_{\tilde{x}\in\mathcal{X}} (\tilde{x})^\top A(y - y^*) \\ &\le \|x - x^*\|\|A\|R_y + \|y - y^*\|\|A\|R_x \\ &\le \sqrt{R_x^2 + R_y^2}\|A\|\|z - z^*\|.\end{aligned}$$

Taking $z^* = \Pi_{\mathcal{Z}^*}(z)$ yields the claim. □

B Proofs for Theorem 5.4

This appendix proves Theorem 5.4, the sub-geometric schedule variant introduced in Section 5.3.

B.1 Proof of Theorem 5.4

Proof of Theorem 5.4. Write $q := \mu/\sqrt{\mu^2+\tau^2}$. The schedule-independent expansion (9) applies here as well. For every $j \geq 1$, $e^{\sqrt{j+1}} \leq N_j \leq 2e^{\sqrt{j+1}}$, and hence

$$\begin{aligned}\sqrt{\frac{\log(eN_j)}{N_j}} &\leq \sqrt{\log(2e)+\sqrt{j+1}}\, e^{-\sqrt{j+1}/2} \\ &\leq \left(\sqrt{\log(2e)}+(j+1)^{1/4}\right) e^{-\sqrt{j+1}/2}.\end{aligned}$$

We control the resulting geometric convolution by the following estimate.

Lemma B.1. *Let $q \in (0,1)$ and $\gamma > 0$. Set $a := -\ln q > 0$ and define*

$$C_{q,\gamma} := \frac{\exp\left(\frac{\gamma^2}{2a}\right)}{1-\exp\left(-\frac{a}{2}\right)} + \exp\left(-a+\frac{\gamma^2}{4a}\right).$$

Then, for every $k \geq 1$,

$$q^k + \sum_{j=1}^{k} q^{k-j} \exp\left(-\gamma\sqrt{j-1}\right) \leq C_{q,\gamma} \exp\left(-\gamma\sqrt{k-1}\right).$$

Since $\sqrt{k+1}-\sqrt{k-1} \leq \sqrt{2}$ for $k \geq 1$ and $e^{1/\sqrt{2}} \leq \sqrt{2e}$, Lemma B.1 with $\gamma = 1/2$ gives

$$\begin{aligned}q^k + \sum_{j=1}^{k} q^{k-j} e^{-\sqrt{j+1}/2} &\leq q^k + \sum_{j=1}^{k} q^{k-j} e^{-\sqrt{j-1}/2} \\ &\leq C_{q,1/2} e^{-\sqrt{k-1}/2} \\ &\leq \sqrt{2e}\, C_{q,1/2} e^{-\sqrt{k+1}/2}. \end{aligned} \tag{22}$$

Using $(j+1)^{1/4} \leq (k+1)^{1/4}$ for $j \leq k$, the preceding feedback bound and (22) in (9) yield

$$\begin{aligned}\sqrt{\mathbb{E}[\operatorname{dist}(r^{k+1}, \mathcal{Z}^*)^2]} \leq \sqrt{2e}\, C_{q,1/2} \Big(\operatorname{dist}(r^1, \mathcal{Z}^*) + \sqrt{C_{\mathrm{inner}}} \left(\sqrt{\log(2e)} + (k+1)^{1/4}\right)\Big) \\ \times\, e^{-\sqrt{k+1}/2}.\end{aligned} \tag{23}$$

It remains to convert the epoch index into the sample count via the following deterministic comparison.

Lemma B.2. *Let*

$$N_k := \left\lceil \exp(\sqrt{k+1}) \right\rceil, \qquad T_k := \sum_{j=1}^{k} N_j, \qquad k = 1, 2, \ldots,$$

Then, for every $k \geq 1$,

$$\begin{aligned}\ln(T_k+e) &\geq \sqrt{k+1}, \\ \exp\left(-\frac{\sqrt{k+1}}{2}\right) &\leq 2\sqrt{\frac{\ln(T_k+e)}{T_k}}.\end{aligned}$$

By Lemma B.2, $\ln(T_k+e) \geq \sqrt{k+1}$. Hence, both $(k+1)^{1/4}$ and 1 are at most $\sqrt{\ln(T_k+e)}$. The same lemma gives $e^{-\sqrt{k+1}/2} \leq 2\sqrt{\ln(T_k+e)/T_k}$. Substituting these bounds into (23) gives

$$\begin{aligned}\sqrt{\mathbb{E}[\operatorname{dist}(r^{k+1}, \mathcal{Z}^*)^2]} &\leq C_{\mathrm{dist}}^{\mathrm{sub}} \frac{\ln(T_k+e)}{\sqrt{T_k}}, \\ C_{\mathrm{dist}}^{\mathrm{sub}} &:= 2\sqrt{2e}\, C_{q,1/2} \left(\operatorname{dist}(r^1, \mathcal{Z}^*) + \sqrt{C_{\mathrm{inner}}} \left(\sqrt{1+\log 2}+1\right)\right).\end{aligned}$$

Expanding $C_{q,1/2}$ with $q = \mu/\sqrt{\mu^2+\tau^2}$ gives $C_{\mathrm{dist}}^{\mathrm{sub}}$ in closed form. Finally, Lemma A.2 and (5) give

$$\begin{aligned}\left(\mathbb{E}[\mathrm{GAP}(z^{T_k+1})^2]\right)^{1/2} &= \left(\mathbb{E}[\mathrm{GAP}(r^{k+1})^2]\right)^{1/2} \\ &\le \sqrt{R_x^2+R_y^2}\|A\|\, C_{\mathrm{dist}}^{\mathrm{sub}} \frac{\ln(T_k+e)}{\sqrt{T_k}}.\end{aligned}$$

□

B.2 Proof of Lemma B.1

Proof of Lemma B.1. Changing variables $r = k-j$ and using $\sqrt{k-1}-\sqrt{k-1-r} \le \sqrt{r}$, we have

$$\begin{aligned}\sum_{j=1}^{k} q^{k-j} e^{-\gamma\sqrt{j-1}} &= e^{-\gamma\sqrt{k-1}} \sum_{r=0}^{k-1} q^r e^{\gamma(\sqrt{k-1}-\sqrt{k-1-r})} \\ &\le e^{-\gamma\sqrt{k-1}} \sum_{r=0}^{\infty} q^r e^{\gamma\sqrt{r}}.\end{aligned}$$

Since $q = e^{-a}$ and $\gamma\sqrt{r} \le ar/2 + \gamma^2/(2a)$,

$$\sum_{r=0}^{\infty} q^r e^{\gamma\sqrt{r}} \le e^{\gamma^2/(2a)} \sum_{r=0}^{\infty} e^{-ar/2} = \frac{e^{\gamma^2/(2a)}}{1-e^{-a/2}}. \tag{24}$$

For the remaining term, completing the square gives

$$\begin{aligned}-ak+\gamma\sqrt{k-1} &= -a + \frac{\gamma^2}{4a} - a\left(\sqrt{k-1}-\frac{\gamma}{2a}\right)^2 \le -a+\frac{\gamma^2}{4a}, \\ q^k &\le e^{-a+\gamma^2/(4a)} e^{-\gamma\sqrt{k-1}}.\end{aligned}$$

Adding this to the convolution bound obtained from (24) proves the claim, including $k=1$. □

B.3 Proof of Lemma B.2

Proof of Lemma B.2. Fix $k \ge 1$ and write $s := \sqrt{k+1}$. Since $T_k \ge N_k \ge e^s$, we have $\ln(T_k+e) \ge \ln T_k \ge s$. Using $N_j \le 2e^{\sqrt{j+1}}$ and monotonicity, we obtain

$$\begin{aligned}T_k &\le 2\sum_{j=1}^{k} e^{\sqrt{j+1}} \le 2\left(e^s + \int_1^k e^{\sqrt{x+1}}\,dx\right) \\ &= 2\left((2s-1)e^s - 2(\sqrt{2}-1)e^{\sqrt{2}}\right) \\ &\le 4se^s.\end{aligned} \tag{25}$$

The integral is zero when $k=1$, so this comparison also covers the first epoch. Combining (25) with $\ln(T_k+e) \ge s$ yields

$$e^{-s/2} \le 2\sqrt{\frac{s}{T_k}} \le 2\sqrt{\frac{\ln(T_k+e)}{T_k}}.$$

□

C Proofs for Theorem 6.3

Variational-inequality formulation. Recall from Section 3 the matrix game (1) on $\mathcal{Z} = \Delta_m \times \Delta_n$, with operator $F(z) = (Ay, -A^\top x)$. This operator is skew-symmetric, so $\langle F(v), v\rangle = 0$ for every v. With the ambient normal cone

$$\mathcal{N}_{\mathcal{Z}}(z) := \{\xi \in \mathbb{R}^{m+n} : \langle \xi, w-z\rangle \le 0 \text{ for every } w \in \mathcal{Z}\}$$

and $\mathcal{A}(z) := F(z) + \mathcal{N}_{\mathcal{Z}}(z)$, a point z^* is a Nash equilibrium if and only if $\langle F(z^*), z - z^* \rangle \geq 0$ for every $z \in \mathcal{Z}$, or equivalently $0 \in \mathcal{A}(z^*)$.

Each simplex has squared Euclidean diameter at most 2, so $D_\mu^2 \leq 4\mu$ and $\|z\| \leq \sqrt{2}$ for every $z \in \mathcal{Z}$. By Section 4, the anchor-regularized game (2) has a unique saddle point $z^{\mu,r} = P_\mu(r)$, the point $p \in \mathcal{Z}$ with $0 \in F_r(p) + \mathcal{N}_{\mathcal{Z}}(p)$, or equivalently

$$\langle F_r(p), z - p \rangle \geq 0, \qquad z \in \mathcal{Z}. \tag{26}$$

The perturbed operator F_r is strongly monotone,

$$\langle F_r(z) - F_r(w), z - w \rangle = \|z - w\|_\mu^2, \tag{27}$$

since the skew term contributes zero.

Under Assumption 6.2, $\mathcal{Z}^* = \{z^*\}$ and $\rho_* := \min_{1 \leq \ell \leq m+n} z_\ell^* > 0$. Uniqueness identifies $\operatorname{dist}(z, \mathcal{Z}^*)$ with $\|z - z^*\|$, while full mixedness supplies the positive local margin ρ_*.

Because the log-barrier keeps every inner iterate in $\operatorname{ri} \mathcal{Z}$, all sampling probabilities are positive. Hence, the joint estimator $\widehat{G}_{k,s} := (\widehat{G}^x_{k,s}, \widehat{G}^y_{k,s})$ in (6) is well defined. Suppressing the epoch index, every epoch starts at $z_1 = r$.

Barrier divergence. For scalars $a, b > 0$, the scalar log-barrier $-\log t$ has Bregman divergence $-\log a + \log b + (a - b)/b = a/b - 1 - \log(a/b)$. Summing over both blocks gives

$$\begin{aligned} D_h(a, b) = &\sum_{i=1}^{m} \left[\frac{a_i^x}{b_i^x} - 1 - \log\left(\frac{a_i^x}{b_i^x} \right) \right] \\ &+ \sum_{j=1}^{n} \left[\frac{a_j^y}{b_j^y} - 1 - \log\left(\frac{a_j^y}{b_j^y} \right) \right]. \end{aligned} \tag{28}$$

The orientation matters: if the first argument has a zero coordinate while the second is positive, then $D_h(a, b) = +\infty$.

Fix one epoch, suppress its index, and write r for its anchor, which is determined by $\mathcal{H}_0$, and $L_s := \sum_{v=1}^{s} \widehat{G}_v$. The inner history is $\mathcal{H}_s$. The strategy z_s is determined by $\mathcal{H}_{s-1}$, and we abbreviate $\mathbb{E}_s[\cdot] := \mathbb{E}[\cdot \mid \mathcal{H}_{s-1}]$. Define

$$C_{\text{bias}} := \left(\frac{\sqrt{2}\,\|A\|}{\sqrt{\mu}} + D_\mu \right) D_\mu + \left(1 + \frac{\|A\|}{\mu} \right) D_\mu^2.$$

C.1 Proof of Theorem 6.3

Proof. Fix any $k \in \{1, \ldots, K\}$. By (5), the strategy at the next epoch boundary satisfies $z^{T_k+1} = r^{k+1}$. The anchor error contracts each epoch, up to an additive inner-loop term.

Lemma C.1 (Global-to-local anchor recursion)**.** *For every epoch $k \geq 1$, let $\ell_k := \frac{\log(eN_k)}{N_k}$. Let*

$$q := \frac{\mu}{\sqrt{\mu^2 + \tau^2}}, \qquad \bar{q}^2 := \frac{1 + q^2}{2}, \qquad C_Y := \frac{1 + q^2}{1 - q^2}.$$

Then, $q < 1$, and the anchor errors satisfy

$$\mathbb{E}[\|r^{k+1} - z^*\|^2] \leq \left(\bar{q}^2 + \frac{4 C_Y C_g \sqrt{\ell_k}}{\mu \rho_*^2} \right) \mathbb{E}[\|r^k - z^*\|^2] + \frac{C_Y C_\ell(\rho_*/2)}{\mu} \ell_k,$$

where C_g is the constant of Lemma C.5 and $C_\ell(\cdot)$ that of Lemma C.6.

Solving this recursion bounds the anchor error at every epoch.

Lemma C.2 (Finite-time solution of the anchor recursion)**.** *Let the anchor errors* $(\mathbb{E}[\|r^k - z^*\|^2])_{k\ge 1}$ *satisfy the recursion of Lemma C.1. Let*

$$k_0 := \min\left\{k \ge 1 : \frac{4C_Y C_g}{\mu\rho_*^2}\sqrt{\ell_k} \le \frac{1-\bar{q}^2}{2}\right\}, \qquad a := \frac{1+\bar{q}^2}{2} \in (0,1), \qquad \lambda := \log\frac{1}{a} > 0.$$

Define

$$C_{\mathrm{rec}} := \max\left\{\frac{D_\mu^2}{\mu\ell_{k_0}}, 2(1+\log 2)\left[\frac{aD_\mu^2 e^{1/(4\lambda)}}{\mu\ell_{k_0}} + \frac{C_Y C_\ell(\rho_*/2)e^{1/(2\lambda)}}{\mu(1-\sqrt{a})}\right]\right\}.$$

Then, k_0 *is finite. Moreover, for every* $k \ge 1$*,*

$$\mathbb{E}[\|r^{k+1} - z^*\|^2] \le C_{\mathrm{rec}}\ell_k = C_{\mathrm{rec}}\frac{\log(eN_k)}{N_k}.$$

The constant C_{rec} *depends only on* m*,* n*,* A*,* μ*,* B*,* η*,* τ*, and* ρ_**.*

Lemma C.1 and Lemma C.2 give

$$\mathbb{E}[\|r^{k+1} - z^*\|^2] \le C_{\mathrm{rec}}\frac{\log(eN_k)}{N_k}.$$

It remains to convert epoch lengths into the number of feedback observations.

Lemma C.3 (Epoch-to-sample conversion)**.** *For every* $k \ge 1$*, let*

$$N_k = \left\lceil e^{\sqrt{k+1}} \right\rceil, \qquad T_k = \sum_{j=1}^{k} N_j.$$

Then, for every $k \ge 1$*,*

$$T_k \le (2e+1)\log(eT_k)N_k.$$

Consequently,

$$\ell_k = \frac{\log(eN_k)}{N_k} \le (2e+1)\frac{\log^2(eT_k)}{T_k}.$$

By Lemma C.3, this becomes

$$\mathbb{E}[\|r^{k+1} - z^*\|^2] \le (2e+1)C_{\mathrm{rec}}\frac{\log^2(eT_k)}{T_k}. \tag{29}$$

By $z^{T_k+1} = r^{k+1}$, Lemma A.2, Assumption 6.2, and (29),

$$\begin{aligned}\left(\mathbb{E}[\mathrm{GAP}(z^{T_k+1})^2]\right)^{1/2} &= \left(\mathbb{E}[\mathrm{GAP}(r^{k+1})^2]\right)^{1/2} \\ &\le \sqrt{R_x^2 + R_y^2}\,\|A\|\sqrt{\mathbb{E}[\|r^{k+1} - z^*\|^2]} \\ &\le \sqrt{R_x^2 + R_y^2}\,\|A\|\sqrt{(2e+1)C_{\mathrm{rec}}}\frac{\log(eT_k)}{\sqrt{T_k}},\end{aligned}$$

which proves the claim for every $k = 1, \dots, K$. □

C.2 Proof of Lemma C.1

Proof. The proximal map P_μ is firmly nonexpansive and fixes the equilibrium.

Lemma C.4 (Firm nonexpansiveness of the map P_μ)**.** *For all* $r, r' \in \mathcal{Z}$*,*

$$\|P_\mu(r) - P_\mu(r')\|^2 \le \langle r - r', P_\mu(r) - P_\mu(r')\rangle,$$

Consequently, P_μ *is* 1*-Lipschitz. Under Assumption 6.2,*

$$P_\mu(z^*) = z^*, \qquad \|P_\mu(r) - z^*\| \le \|r - z^*\|.$$

By Lemma C.4, P_μ is continuous. Since r^k is determined by the epoch-start history $\mathcal{H}_{k,0}$, so are $P_\mu(r^k)$ and $\|r^k - z^*\|$. By Lemma C.4, on $\{\|r^k - z^*\| \le \rho_*/2\}$, every coordinate satisfies

$$\left|[P_\mu(r^k)]_\ell - z^*_\ell\right| \le \left\|P_\mu(r^k) - z^*\right\| \le \left\|r^k - z^*\right\| \le \frac{\rho_*}{2}.$$

Since $z^*_\ell \ge \rho_*$, it follows that

$$\{\|r^k - z^*\| \le \rho_*/2\} \subseteq \left\{\min_\ell [P_\mu(r^k)]_\ell \ge \frac{\rho_*}{2}\right\}. \tag{30}$$

We use two inner-loop bounds on the epoch's terminal error, a global one valid for every anchor and a sharper local one for an interior proximal target.

Lemma C.5 (Global inner MSE, including boundary proximal targets)**.** *Fix one epoch of deterministic length $N \ge 1$. Let its anchor r be determined by $\mathcal{H}_0$, and let*

$$p = P_\mu(r),$$

which may lie on the relative boundary of $\mathcal{Z}$. Then

$$\mathbb{E}\left[\|z_{N+1} - p\|_\mu^2 \mid \mathcal{H}_0\right] \le \frac{C_g}{\sqrt{N}},$$

where one valid anchor-independent constant is

$$C_g := \frac{8(m+n)}{\eta} + \frac{4C_{\text{seed}}}{\eta} + 4C_{\text{bias}} + 10D_\mu^2 + \frac{4C_{\eta B}\sqrt{\eta}\,B^2(m+n)}{\sqrt{\mu}}.$$

Here C_{seed} and $C_{\eta B}$ are the constants of Lemma C.8 and Proposition C.11, respectively. The constant depends only on the fixed dimensions, A, μ, B, and η, and not on r, p, or an interior margin.

Lemma C.6 (Local inner MSE for an interior proximal target)**.** *Fix one epoch of length $N \ge 1$, let r be determined by $\mathcal{H}_0$, and let*

$$p = P_\mu(r).$$

Suppose that, for a deterministic analysis constant $\rho > 0$,

$$\min_{1 \le \ell \le m+n} p_\ell \ge \rho.$$

Define the analysis-only split parameter

$$\theta_\rho := \min\left\{\frac{1}{2}, \frac{\mu\rho}{16C_{\eta B}\eta B^2}\right\},$$

the barrier constant

$$H_\rho := m\log\frac{1}{m\rho} + n\log\frac{1}{n\rho},$$

and

$$B_\rho := \eta D_\mu^2 + C_{\eta B}\eta B^2\left[\frac{m}{\mu\theta_\rho\rho} + \frac{n}{\mu\theta_\rho\rho}\right].$$

Then

$$\mathbb{E}\left[\|z_{N+1} - p\|_\mu^2 \mid \mathcal{H}_0\right] \le C_\ell(\rho)\frac{\log(eN)}{N},$$

where

$$C_\ell(\rho) := \frac{2(H_\rho + C_{\text{seed}} + B_\rho)}{\eta}.$$

Neither ρ nor θ_ρ is an algorithm input.

Using $\|v\|_\mu^2 = \mu\|v\|^2$ and $\ell_k \geq 1/N_k$, Lemma C.5 gives a conditional Euclidean MSE at most $\frac{C_g}{\mu}\sqrt{\ell_k}$ for every anchor. On the event $\{\|r^k - z^*\| \leq \rho_*/2\}$, which is determined by $\mathcal{H}_{k,0}$, (30) and Lemma C.6 give the sharper bound $\frac{C_\ell(\rho_*/2)}{\mu}\ell_k$. Thus, conditioning on the epoch-start history,

$$\mathbb{E}[\|r^{k+1} - P_\mu(r^k)\|^2 \mid \mathcal{H}_{k,0}] \leq \frac{C_\ell(\rho_*/2)}{\mu}\ell_k + \frac{C_g}{\mu}\sqrt{\ell_k}\,\mathbf{1}_{\{\|r^k - z^*\| > \rho_*/2\}}. \tag{31}$$

The added $\frac{C_\ell(\rho_*/2)}{\mu}\ell_k$ on $\{\|r^k - z^*\| > \rho_*/2\}$ only increases the upper bound.

Lemma A.3 gives $\|P_\mu(r^k) - z^*\| \leq q\|r^k - z^*\|$. For every $\alpha > 0$, Young's inequality yields

$$\begin{aligned}\|r^{k+1} - z^*\|^2 &= \|(P_\mu(r^k) - z^*) + (r^{k+1} - P_\mu(r^k))\|^2 \\ &\leq (1+\alpha)\|P_\mu(r^k) - z^*\|^2 + (1+\alpha^{-1})\|r^{k+1} - P_\mu(r^k)\|^2.\end{aligned}$$

Choose $\alpha = (1-q^2)/(2q^2)$, so $(1+\alpha)q^2 = \bar{q}^2$ and $1+\alpha^{-1} = C_Y$. Taking conditional expectations and applying (31) gives

$$\mathbb{E}[\|r^{k+1} - z^*\|^2 \mid \mathcal{H}_{k,0}] \leq \bar{q}^2\|r^k - z^*\|^2 + \frac{C_Y C_\ell(\rho_*/2)}{\mu}\ell_k + \frac{C_Y C_g}{\mu}\sqrt{\ell_k}\,\mathbf{1}_{\{\|r^k - z^*\| > \rho_*/2\}}.$$

After taking full expectations, Markov's inequality $\Pr(\|r^k - z^*\| > \rho_*/2) \leq 4\mathbb{E}[\|r^k - z^*\|^2]/\rho_*^2$ implies

$$\mathbb{E}[\|r^{k+1} - z^*\|^2] \leq \left(\bar{q}^2 + \frac{4C_Y C_g\sqrt{\ell_k}}{\mu\rho_*^2}\right)\mathbb{E}[\|r^k - z^*\|^2] + \frac{C_Y C_\ell(\rho_*/2)}{\mu}\ell_k,$$

which is the claimed recursion. □

C.3 Proof of Lemma C.4

Proof. Let $p = P_\mu(r)$ and $p' = P_\mu(r')$. By (26) and the definition of the normal cone,

$$\mu(r - p) \in \mathcal{A}(p), \qquad \mu(r' - p') \in \mathcal{A}(p').$$

The operator $\mathcal{A} = F + \mathcal{N}_\mathcal{Z}$ is monotone because F is skew-symmetric and $\mathcal{N}_\mathcal{Z}$ is monotone. Applying monotonicity to these inclusions gives

$$0 \leq \mu\langle(r - p) - (r' - p'), p - p'\rangle = \mu\langle r - r', p - p'\rangle - \mu\|p - p'\|^2,$$

This proves the firm-nonexpansiveness inequality in Lemma C.4. Cauchy–Schwarz then yields $\|P_\mu(r) - P_\mu(r')\| \leq \|r - r'\|$, so P_μ is 1-Lipschitz. Since $0 \in \mathcal{A}(z^*)$, uniqueness of the solution to (26) gives $P_\mu(z^*) = z^*$. Taking $r' = z^*$ in this inequality proves the final inequality in Lemma C.4. □

C.4 Proof of Lemma C.5

Proof. Let $d := m + n$, choose $\gamma \in (0, 1)$, and define $p^\gamma := (1-\gamma)p + \gamma u$. Since u is strictly positive, $p^\gamma \in \operatorname{ri}\mathcal{Z}$ and $\|p^\gamma - p\|_\mu = \gamma\|u - p\|_\mu \leq \gamma D_\mu$. By Lemma C.4, p and p^γ are determined by $\mathcal{H}_0$.

We first bound the barrier divergence. In (28), the linear terms cancel because $\sum_i(mp_i^{\gamma,x} - 1) = \sum_j(np_j^{\gamma,y} - 1) = 0$. Thus,

$$D_h(p^\gamma, u) = -\sum_{i=1}^m \log(mp_i^{\gamma,x}) - \sum_{j=1}^n \log(np_j^{\gamma,y}). \tag{32}$$

Since $mp_i^{\gamma,x} = (1-\gamma)mp_i^x + \gamma \geq \gamma$ and $np_j^{\gamma,y} \geq \gamma$, (32) gives $D_h(p^\gamma, u) \leq d\log(1/\gamma)$.

Next, set $b := p^\gamma - p$. Since $F(z) = (Ay, -A^\top x)$ is linear with $\|F(z') - F(z)\| \leq \|A\|\,\|z' - z\|$, we have $F_r(p^\gamma) - F_r(p) = F(p^\gamma) - F(p) + \mu b$ and $\|F_r(p^\gamma) - F_r(p)\| \leq (\|A\| + \mu)\|b\|$. Using

(27) and (26), we obtain, for every $z \in \mathcal{Z}$,

$$\begin{aligned} D_q(p^\gamma, z) - \langle F_r(z), z - p^\gamma \rangle &= -D_q(p^\gamma, z) - \langle F_r(p), z - p \rangle + \langle F_r(p), b \rangle \\ &\quad - \langle F_r(p^\gamma) - F_r(p), z - p^\gamma \rangle \\ &\le \left(\frac{\sqrt{2}\,\|A\|}{\sqrt{\mu}} + D_\mu \right) \|b\|_\mu + \left(1 + \frac{\|A\|}{\mu} \right) \|b\|_\mu \, \|z - p^\gamma\|_\mu \\ &\le C_{\mathrm{bias}} \gamma. \end{aligned} \tag{33}$$

The first inequality discards the nonpositive terms $-D_q(p^\gamma, z)$ and $-\langle F_r(p), z - p \rangle$, uses Cauchy–Schwarz together with $\|F_r(p)\| \le \sqrt{2}\,\|A\| + D_\mu \sqrt{\mu}$ and the linearization bound above, and rewrites norms via $\|v\|_\mu = \sqrt{\mu}\,\|v\|$. The second inequality uses $\|b\|_\mu \le \gamma D_\mu$ and $\|z - p^\gamma\|_\mu \le D_\mu$.

The inner loop rests on three estimates, namely a one-step potential inequality, its value after the first feedback, and a bound on the accumulated stability terms.

Proposition C.7 (One-step potential inequality)**.** *Let $w \in \mathrm{ri}\,\mathcal{Z}$ be determined by $\mathcal{H}_0$, and define*

$$V_s(w) := D_h(w, z_s) + (s-1)\eta D_q(w, z_s).$$

Set

$$S_s := \eta \left\langle \widehat{G}_s + \mu(z_s - r), z_s - z_{s+1} \right\rangle - D_h(z_{s+1}, z_s) - s\eta D_q(z_{s+1}, z_s).$$

Then, for every $s \ge 2$,

$$\mathbb{E}_s[V_{s+1}(w)] \le V_s(w) + \eta \left[D_q(w, z_s) - \langle F_r(z_s), z_s - w \rangle \right] + \mathbb{E}_s[S_s].$$

Lemma C.8 (Potential after the first feedback)**.** *Suppose that $z_1 = r \in \mathrm{ri}\,\mathcal{Z}$. For every $w \in \mathrm{ri}\,\mathcal{Z}$ determined by $\mathcal{H}_0$,*

$$\mathbb{E}[V_2(w) \mid \mathcal{H}_0] \le D_h(w, u) + C_{\mathrm{seed}},$$

where

$$C_{\mathrm{seed}} := \eta B(m+n) + \frac{\eta}{2} D_\mu^2.$$

Proposition C.9 (Pathwise and conditional stability)**.** *For every inner step $s \ge 2$, the term S_s of Proposition C.7 satisfies, pathwise,*

$$\begin{aligned} S_s &\le \frac{\eta D_\mu^2}{s} \\ &\quad + C_{\eta B} \eta^2 R_s^2 \left[\frac{1}{1 + s\eta\mu x_{s,I_s}^2} + \frac{1}{1 + s\eta\mu y_{s,J_s}^2} \right]. \end{aligned}$$

Consequently,

$$\begin{aligned} \mathbb{E}_s[S_s] &\le \frac{\eta D_\mu^2}{s} \\ &\quad + C_{\eta B} \eta^2 B^2 \left[\sum_{i=1}^m \frac{x_{s,i}}{1 + s\eta\mu x_{s,i}^2} + \sum_{j=1}^n \frac{y_{s,j}}{1 + s\eta\mu y_{s,j}^2} \right]. \end{aligned}$$

In particular, we have

$$\mathbb{E}_s[S_s] \le \frac{\eta D_\mu^2}{s} + \frac{C_{\eta B} \eta^{3/2} B^2 (m+n)}{2\sqrt{\mu s}},$$

and hence

$$\sum_{s=2}^N \mathbb{E}[S_s \mid \mathcal{H}_0] \le \eta D_\mu^2 \log(eN) + \frac{C_{\eta B} \eta^{3/2} B^2 (m+n)\sqrt{N}}{\sqrt{\mu}}.$$

The sum is understood to be zero when $N = 1$.

Apply Proposition C.7 with $w = p^\gamma$ and sum from $s = 2$ to N while conditioning on $\mathcal{H}_0$. The tower property, Lemma C.8, the barrier bound above, (33), and the summed bound in Proposition C.9 give

$$\mathbb{E}[V_{N+1}(p^\gamma) \mid \mathcal{H}_0] \le d\log(1/\gamma) + C_{\text{seed}} + N\eta C_{\text{bias}}\gamma + \eta D_\mu^2 \log(eN) + \frac{C_{\eta B}\eta^{3/2}B^2(m+n)\sqrt{N}}{\sqrt{\mu}}. \tag{34}$$

For $N = 1$, the sums are empty and the same bound follows directly from Lemma C.8. By nonnegativity of the barrier divergence and the squared triangle inequality,

$$V_{N+1}(p^\gamma) \ge \frac{N\eta}{2}\|z_{N+1} - p^\gamma\|_\mu^2,$$
$$\|z_{N+1} - p\|_\mu^2 \le 2\|z_{N+1} - p^\gamma\|_\mu^2 + 2\gamma^2 D_\mu^2.$$

Combining these inequalities with (34) yields

$$\mathbb{E}[\|z_{N+1} - p\|_\mu^2 \mid \mathcal{H}_0] \le \frac{4d\log(1/\gamma)}{\eta N} + \frac{4C_{\text{seed}}}{\eta N} + 4C_{\text{bias}}\gamma + 2D_\mu^2\gamma^2 + \frac{4D_\mu^2\log(eN)}{N} + \frac{4C_{\eta B}\sqrt{\eta}B^2(m+n)}{\sqrt{\mu N}}. \tag{35}$$

Choose $\gamma = 1/(N+1)$. For $N \ge 1$, $\log(N+1) \le \log(eN) \le 2\sqrt{N}$; the second inequality follows because $2\sqrt{N} - 1 - \log N$ is positive at $N = 1$ and has nonnegative derivative. Using also $(N+1)^{-1} \le N^{-1} \le N^{-1/2}$ and $(N+1)^{-2} \le N^{-1/2}$ in (35), we obtain

$$\mathbb{E}[\|z_{N+1} - p\|_\mu^2 \mid \mathcal{H}_0] \le \left(\frac{8d}{\eta} + \frac{4C_{\text{seed}}}{\eta} + 4C_{\text{bias}} + 10D_\mu^2 + \frac{4C_{\eta B}\sqrt{\eta}B^2(m+n)}{\sqrt{\mu}}\right)\frac{1}{\sqrt{N}} = \frac{C_g}{\sqrt{N}}.$$

□

C.5 Proof of Proposition C.7

Proof. The cumulative update has an equivalent sequential form.

Proposition C.10 (Exact equivalence of cumulative and sequential composite FTRL)**.** *For every $s \ge 2$, the cumulative update (4) is exactly equivalent to*

$$z_{s+1} = \arg\min_{z \in \operatorname{ri}\mathcal{Z}} \left\{\eta\left\langle \widehat{G}_s + \mu(z_s - r), z\right\rangle + D_h(z, z_s) + s\eta D_q(z, z_s)\right\}.$$

No approximation or linearization of the accumulated quadratic regularizer is made.

For $s \ge 2$, the first-order condition in Proposition C.10 gives, for every $w \in \operatorname{ri}\mathcal{Z}$,

$$0 \le \left\langle \eta\big(\widehat{G}_s + \mu(z_s - r)\big) + \nabla h(z_{s+1}) - \nabla h(z_s) + s\eta\mu(z_{s+1} - z_s),\ w - z_{s+1}\right\rangle.$$

The three-point identities are

$$\langle \nabla h(z_{s+1}) - \nabla h(z_s), w - z_{s+1}\rangle = D_h(w, z_s) - D_h(w, z_{s+1}) - D_h(z_{s+1}, z_s),$$
$$\langle \mu(z_{s+1} - z_s), w - z_{s+1}\rangle = D_q(w, z_s) - D_q(w, z_{s+1}) - D_q(z_{s+1}, z_s).$$

Substitution and rearrangement give

$$V_{s+1}(w) \le V_s(w) + \eta D_q(w, z_s) - \eta\left\langle \widehat{G}_s + \mu(z_s - r), z_{s+1} - w\right\rangle - D_h(z_{s+1}, z_s) - s\eta D_q(z_{s+1}, z_s). \tag{36}$$

Since $z_{s+1} - w = (z_s - w) - (z_s - z_{s+1})$, (36) becomes

$$V_{s+1}(w) \le V_s(w) + \eta D_q(w, z_s) - \eta\left\langle \widehat{G}_s + \mu(z_s - r), z_s - w\right\rangle + S_s.$$

We take $\mathbb{E}_s$. Since z_s, w, r are determined by $\mathcal{H}_{s-1}$ and $\mathbb{E}_s[\widehat{G}_s] = F(z_s)$ by unbiasedness, the claim follows. □

C.6 Proof of Proposition C.10

Proof. Fix $s \geq 2$ and let $C \in \mathbb{R}^{2\times(m+n)}$ encode the affine constraints $Cz = (\mathbf{1}_m^\top x, \mathbf{1}_n^\top y)^\top = (1,1)^\top$. The objective defining z_s is $J_{s-1}(z) := \eta \langle L_{s-1}, z\rangle + (s-1)\eta q_r(z) + h(z)$. Since the barrier diverges at the relative boundary, $z_s \in \operatorname{ri}\mathcal{Z}$ and its nonnegativity constraints are inactive. Hence, there is $\lambda_s \in \mathbb{R}^2$ such that

$$\eta L_{s-1} + (s-1)\eta\mu(z_s - r) + \nabla h(z_s) + C^\top \lambda_s = 0. \tag{37}$$

For every feasible z, $C(z - z_s) = 0$. The affine term in the Bregman expansion of J_{s-1} therefore vanishes by (37), giving

$$J_{s-1}(z) = J_{s-1}(z_s) + D_h(z, z_s) + (s-1)\eta D_q(z, z_s),$$
$$q_r(z) = q_r(z_s) + \langle \mu(z_s - r), z - z_s\rangle + D_q(z, z_s).$$

Using this exact quadratic identity, we have

$$\begin{aligned} J_s(z) &= J_{s-1}(z) + \eta\left\langle \widehat{G}_s, z\right\rangle + \eta q_r(z) \\ &= \text{constant in } z + D_h(z, z_s) + s\eta D_q(z, z_s) + \eta\left\langle \widehat{G}_s + \mu(z_s - r), z\right\rangle. \end{aligned}$$

Taking the minimizer proves the claimed equivalence without linearizing the accumulated quadratic regularizer. □

C.7 Proof of Lemma C.8

Proof. Recall the potential $V_2(w) = D_h(w, z_2) + \eta D_q(w, z_2)$ from Proposition C.7. The first cumulative update is

$$z_2 = \arg\min_{z\in\operatorname{ri}\mathcal{Z}}\{h(z) + \eta\left\langle \widehat{G}_1, z\right\rangle + \eta q_r(z)\}.$$

Let $C \in \mathbb{R}^{2\times(m+n)}$ encode the affine simplex constraints. The first-order condition gives a multiplier $\lambda \in \mathbb{R}^2$ such that

$$\nabla h(z_2) + \eta\widehat{G}_1 + \eta\mu(z_2 - r) + C^\top\lambda = 0.$$

Taking its inner product with $w - z_2$, using $C(w - z_2) = 0$, and expanding the quadratic terms gives

$$V_2(w) = h(w) - h(z_2) + \eta\left\langle \widehat{G}_1, w - z_2\right\rangle + \eta\big(q_r(w) - q_r(z_2)\big). \tag{38}$$

The uniform point minimizes h, and $\nabla h(u)$ is constant within each simplex block, so

$$h(w) - h(z_2) \leq h(w) - h(u) = D_h(w, u),$$
$$q_r(w) - q_r(z_2) \leq q_r(w) \leq \frac{1}{2}D_\mu^2.$$

It remains to control the stochastic inner product even though z_2 depends on the first observation. Conditional on $\mathcal{H}_0$, $z_1 = r = (x_1, y_1)$ is fixed. Each coordinate of $w - z_2$ has absolute value at most one, so bounded feedback and conditional action sampling imply

$$\mathbb{E}[\left|\left\langle \widehat{G}_1^x, w^x - z_2^x\right\rangle\right| \mid \mathcal{H}_0] \leq \sum_{i=1}^m\sum_{j=1}^n x_{1,i}y_{1,j}\frac{B}{x_{1,i}} = Bm,$$
$$\mathbb{E}[\left|\left\langle \widehat{G}_1^y, w^y - z_2^y\right\rangle\right| \mid \mathcal{H}_0] \leq \sum_{i=1}^m\sum_{j=1}^n x_{1,i}y_{1,j}\frac{B}{y_{1,j}} = Bn.$$

Combining these inequalities with (38) yields

$$\mathbb{E}[V_2(w) \mid \mathcal{H}_0] \leq D_h(w, u) + \eta B(m+n) + \frac{\eta}{2}D_\mu^2 = D_h(w, u) + C_{\text{seed}}.$$

□

C.8 Proof of Proposition C.9

Proof. Split S_s exactly as

$$S_s = \left[\eta \left\langle \widehat{G}_s, z_s - z_{s+1} \right\rangle - D_h(z_{s+1}, z_s) - \frac{s\eta}{2} D_q(z_{s+1}, z_s)\right] + \left[\eta \langle \mu(z_s - r), z_s - z_{s+1}\rangle - \frac{s\eta}{2} D_q(z_{s+1}, z_s)\right]. \tag{39}$$

For each coordinate, $z_{s,\ell} > 0$, so put $t_\ell := z_{s+1,\ell}/z_{s,\ell} > 0$. Then, $z_{s,\ell} - z_{s+1,\ell} = z_{s,\ell}(1 - t_\ell)$, and (28) gives $D_h(z_{s+1}, z_s) = \sum_\ell \phi(t_\ell)$. In the first bracket of (39), the quadratic term assigned to coordinate ℓ is $s\eta\mu z_{s,\ell}^2(1 - t_\ell)^2/4$. Writing superscripts x and y for the two blocks of $z_s - z_{s+1}$ and t, the importance weights cancel on the sampled coordinates:

$$\eta \widehat{G}^x_{s,I_s}(x_{s,I_s} - x_{s+1,I_s}) = \eta R_s(1 - t^x_{I_s}),$$
$$\eta \widehat{G}^y_{s,J_s}(y_{s,J_s} - y_{s+1,J_s}) = -\eta R_s(1 - t^y_{J_s}).$$

A scalar log-barrier inequality controls each sampled coordinate.

Proposition C.11 (Global scalar log-barrier inequality)**.** *Suppose $\eta B < 1$. Let*

$$\phi(t) := t - 1 - \log t, \qquad t > 0,$$

and

$$C_{\eta B} := \max\left\{1, \frac{1}{2(1 - \eta B)^2}\right\}.$$

For every $a \geq 0$, every $c \in [-\eta B, \eta B]$, and every $t > 0$,

$$c(1 - t) - \phi(t) - a(1 - t)^2 \leq C_{\eta B} \frac{c^2}{1 + 4a}.$$

Thus, Proposition C.11 applies with

$$c_x = \eta R_s, \quad a_x = \frac{s\eta\mu x^2_{s,I_s}}{4}, \qquad c_y = -\eta R_s, \quad a_y = \frac{s\eta\mu y^2_{s,J_s}}{4}.$$

Both signs are covered because $|c_x|, |c_y| \leq \eta B$. Every unsampled coordinate has $c = 0$ and contributes a nonpositive amount. Summing the coordinate bounds, with $1 + 4a_x = 1 + s\eta\mu x^2_{s,I_s}$ and likewise for y, yields

$$\eta \left\langle \widehat{G}_s, z_s - z_{s+1} \right\rangle - D_h(z_{s+1}, z_s) - \frac{s\eta}{2} D_q(z_{s+1}, z_s) \leq C_{\eta B}\eta^2 R_s^2 \left[\frac{1}{1 + s\eta\mu x^2_{s,I_s}} + \frac{1}{1 + s\eta\mu y^2_{s,J_s}}\right]. \tag{40}$$

For the second bracket in (39), completing the square gives

$$\eta \langle \mu(z_s - r), z_s - z_{s+1}\rangle - \frac{s\eta}{4} \|z_s - z_{s+1}\|^2_\mu = -\frac{s\eta}{4} \left\|(z_s - z_{s+1}) - \frac{2}{s}(z_s - r)\right\|^2_\mu + \frac{\eta}{s} \|z_s - r\|^2_\mu \leq \frac{\eta D_\mu^2}{s}.$$

Combining this with (40) proves the pathwise bound in Proposition C.9.

Conditionally on $\mathcal{H}_{s-1}$, the current coordinates are fixed. Using $|R_s| \leq B$ and the sampling law of Assumption 6.1, we obtain

$$\mathbb{E}_s\left[\frac{R_s^2}{1 + s\eta\mu x^2_{s,I_s}}\right] \leq B^2 \sum_{i=1}^m \frac{x_{s,i}}{1 + s\eta\mu x^2_{s,i}},$$
$$\mathbb{E}_s\left[\frac{R_s^2}{1 + s\eta\mu y^2_{s,J_s}}\right] \leq B^2 \sum_{j=1}^n \frac{y_{s,j}}{1 + s\eta\mu y^2_{s,j}}.$$

Substitution into the pathwise bound in Proposition C.9 gives its conditional bound. For $v \geq 0$ and $a > 0$, $1 + av^2 \geq 2\sqrt{a}\, v$, so $v/(1 + av^2) \leq 1/(2\sqrt{a})$. Applying this to both blocks with $a = s\eta\mu$ gives the one-step expectation bound in Proposition C.9. Finally, $\sum_{s=1}^N 1/s \leq \log(eN)$ and $\sum_{s=1}^N s^{-1/2} \leq 2\sqrt{N}$ yield the summed bound in Proposition C.9 by the tower property after conditioning on $\mathcal{H}_0$. The sum from $s = 2$ to N is zero when $N = 1$. □

C.9 Proof of Proposition C.11

Proof. If $c = 0$, the left-hand side is nonpositive. Suppose first that $c > 0$. For $t \geq 1$, all three terms on the left-hand side are nonpositive. For $0 < t \leq 1$, put $v := 1 - t \in [0, 1)$. The power series for $-\log(1 - v)$ gives

$$\phi(t) = -v - \log(1 - v) = \sum_{j=2}^{\infty} \frac{v^j}{j} \geq \frac{v^2}{2}.$$

Consequently,

$$\begin{aligned} c(1 - t) - \phi(t) - a(1 - t)^2 &\leq cv - \left(a + \frac{1}{2}\right) v^2 \\ &\leq \frac{c^2}{4a + 2} \leq \frac{c^2}{1 + 4a} \leq C_{\eta B} \frac{c^2}{1 + 4a}. \end{aligned}$$

Now suppose $c < 0$ and put $b := -c \in (0, \eta B]$. For $t \leq 1$, the left-hand side is again nonpositive. For $t = 1 + v$ with $v \geq 0$, it equals

$$\psi(v) := bv - \big(v - \log(1 + v)\big) - av^2.$$

This function attains a maximum on $[0, \infty)$: it tends to $-\infty$ quadratically when $a > 0$, and when $a = 0$ we have $\psi(v) = (b - 1)v + \log(1 + v) \to -\infty$ because $b < 1$. If the maximum is attained at zero, its value is zero. Otherwise, at a positive maximizer,

$$0 = \psi'(v) = b - \frac{v}{1 + v} - 2av, \qquad \frac{v}{1 + v} \leq b \leq \eta B, \qquad v \leq \frac{\eta B}{1 - \eta B}.$$

On this interval, $(v - \log(1 + v))'' = (1 + v)^{-2} \geq (1 - \eta B)^2$. Since the function and its first derivative vanish at zero, two integrations give

$$v - \log(1 + v) \geq \frac{(1 - \eta B)^2}{2} v^2.$$

We can therefore bound the value at any positive maximizer by completing the square:

$$\psi(v) \leq bv - \left(a + \frac{(1 - \eta B)^2}{2}\right) v^2 \leq \frac{b^2}{4a + 2(1 - \eta B)^2}.$$

The ratio $(1 + 4a)/(4a + 2(1 - \eta B)^2)$ is bounded over $a \geq 0$ by $\max\{1, 1/(2(1 - \eta B)^2)\} = C_{\eta B}$. Thus the maximum, and hence every value of ψ, is at most $C_{\eta B} b^2/(1 + 4a) = C_{\eta B} c^2/(1 + 4a)$. This proves the claim for both signs of c.

The endpoint $|R_s| = B$ is allowed because $|\eta R_s| \leq \eta B$, but the strict condition $\eta B < 1$ is essential. At $c = -1$ and $a = 0$, $c(1 - t) - \phi(t) = \log t$ has infinite supremum. In the algorithm, a is positive but proportional to the square of a current coordinate and can be arbitrarily close to zero, so no trajectory-uniform constant is obtained at $\eta B = 1$. The negative sign in the y-block estimator makes this obstruction present even when the raw payoff is nonnegative. Our proof handles signed payoffs directly and does not require a payoff shift. □

C.10 Proof of Lemma C.6

Proof. Because $p \in \operatorname{ri} \mathcal{Z}$, Proposition C.7 applies with $w = p$ for every $s \geq 2$. Recall the potential $V_s(w) = D_h(w, z_s) + (s - 1)\eta D_q(w, z_s)$ and the stability term S_s of that proposition. By (27) and (26),

$$\begin{aligned} \langle F_r(z_s), z_s - p\rangle &= \langle F_r(p), z_s - p\rangle + \langle F_r(z_s) - F_r(p), z_s - p\rangle \\ &\geq \|z_s - p\|_\mu^2 = 2D_q(p, z_s). \end{aligned} \tag{41}$$

Substituting (41) into Proposition C.7 gives

$$\mathbb{E}_s[V_{s+1}(p)] \leq V_s(p) - \eta D_q(p, z_s) + \mathbb{E}_s[S_s].$$

We refine the conditional bound in Proposition C.9 by fixing $\theta \in (0, 1/2]$. Let (v, p^v) denote either (x_s, p^x) or (y_s, p^y). For a good coordinate, $v_i \geq \theta p_i^v \geq \theta\rho$ implies

$$\frac{v_i}{1 + s\eta\mu v_i^2} \leq \frac{1}{s\eta\mu v_i} \leq \frac{1}{s\eta\mu\theta\rho}.$$

For a bad coordinate, $v_i < \theta p_i^v$ gives $p_i^v - v_i > (1-\theta)p_i^v$, so

$$\frac{v_i}{1 + s\eta\mu v_i^2} \leq v_i \leq \theta p_i^v \leq \frac{\theta(p_i^v - v_i)^2}{(1-\theta)^2 p_i^v} \leq \frac{\theta(p_i^v - v_i)^2}{(1-\theta)^2\rho}.$$

Summing the good and bad contributions in each block and using $\|\cdot\|_\mu^2 = \mu\,\|\cdot\|^2$, we obtain

$$\sum_{i=1}^{m} \frac{x_{s,i}}{1 + s\eta\mu x_{s,i}^2} \leq \frac{m}{s\eta\mu\theta\rho} + \frac{\theta}{\mu(1-\theta)^2\rho}\,\|x_s - p^x\|_\mu^2\,, \tag{42}$$

$$\sum_{j=1}^{n} \frac{y_{s,j}}{1 + s\eta\mu y_{s,j}^2} \leq \frac{n}{s\eta\mu\theta\rho} + \frac{\theta}{\mu(1-\theta)^2\rho}\,\|y_s - p^y\|_\mu^2\,. \tag{43}$$

Substituting (42) and (43) into the conditional bound in Proposition C.9 gives

$$\begin{aligned}\mathbb{E}_s[S_s] \leq \frac{\eta D_\mu^2}{s} + \frac{C_{\eta B}\eta B^2}{s}\left[\frac{m}{\mu\theta\rho} + \frac{n}{\mu\theta\rho}\right] \\ + \frac{C_{\eta B}\eta^2 B^2\theta}{\mu(1-\theta)^2\rho}\,\|z_s - p\|_\mu^2\,.\end{aligned} \tag{44}$$

The last term is $2C_{\eta B}\eta^2 B^2\theta D_q(p, z_s)/(\mu(1-\theta)^2\rho)$, which is at most $\eta D_q(p, z_s)/2$ whenever $4C_{\eta B}\eta B^2\theta \leq \mu(1-\theta)^2\rho$. The choice $\theta_\rho = \min\{1/2, \mu\rho/(16C_{\eta B}\eta B^2)\}$ in the statement satisfies this condition, since

$$(1-\theta_\rho)^2 \geq \frac{1}{4}, \qquad 4C_{\eta B}\eta B^2\theta_\rho \leq \frac{\mu\rho}{4} \leq \mu(1-\theta_\rho)^2\rho.$$

Using (44) with $\theta = \theta_\rho$ in the one-step bound above leaves

$$\mathbb{E}_s[V_{s+1}(p)] \leq V_s(p) - \frac{\eta}{2}D_q(p, z_s) + \frac{B_\rho}{s},$$

with the same B_ρ as in the statement.

The exact cancellation used in (32) and the margin $p_\ell \geq \rho$ give

$$\begin{aligned}D_h(p, u) &= -\sum_{i=1}^{m}\log(mp_i^x) - \sum_{j=1}^{n}\log(np_j^y) \\ &\leq m\log\frac{1}{m\rho} + n\log\frac{1}{n\rho} = H_\rho.\end{aligned}$$

Discarding the nonpositive drift, summing from $s = 2$ to N, and using the tower property and Lemma C.8, we obtain

$$\mathbb{E}[V_{N+1}(p) \mid \mathcal{H}_0] \leq H_\rho + C_{\text{seed}} + B_\rho\log(eN). \tag{45}$$

This also holds for $N = 1$, when the sum is empty. Since $V_{N+1}(p) \geq N\eta D_q(p, z_{N+1}) = N\eta\,\|z_{N+1} - p\|_\mu^2/2$, (45) implies

$$\begin{aligned}\mathbb{E}[\|z_{N+1} - p\|_\mu^2 \mid \mathcal{H}_0] &\leq \frac{2H_\rho + 2C_{\text{seed}} + 2B_\rho\log(eN)}{\eta N} \\ &\leq C_\ell(\rho)\frac{\log(eN)}{N},\end{aligned}$$

where the last inequality uses $\log(eN) \geq 1$ and the definition of $C_\ell(\rho)$. □

C.11 Proof of Lemma C.2

Proof. Lemma C.1 gives

$$\mathbb{E}[\|r^{k+1}-z^*\|^2] \le \left(\bar{q}^2 + \frac{4C_Y C_g\sqrt{\ell_k}}{\mu\rho_*^2}\right)\mathbb{E}[\|r^k-z^*\|^2] + \frac{C_Y C_\ell(\rho_*/2)}{\mu}\ell_k.$$

The ceiling bounds $e^{\sqrt{k+1}} \le N_k \le e^{\sqrt{k+1}}+1 \le 2e^{\sqrt{k+1}}$ imply

$$\frac{1+\sqrt{k+1}}{2e^{\sqrt{k+1}}} \le \ell_k \le \frac{1+\sqrt{k+1}+\log 2}{e^{\sqrt{k+1}}}. \tag{46}$$

Since N_k is nondecreasing and $x \mapsto (1+\log x)/x$ is nonincreasing for $x \ge 1$, ℓ_k is nonincreasing. The upper bound in (46) shows that $\ell_k \to 0$, so k_0 is finite. For $k \ge k_0$, the recursion therefore becomes $\mathbb{E}[\|r^{k+1}-z^*\|^2] \le a\mathbb{E}[\|r^k-z^*\|^2] + \frac{C_Y C_\ell(\rho_*/2)}{\mu}\ell_k$. Iterating this inequality yields

$$\mathbb{E}[\|r^{k+1}-z^*\|^2] \le a^{k-k_0+1}\mathbb{E}[\|r^{k_0}-z^*\|^2] + \frac{C_Y C_\ell(\rho_*/2)}{\mu}\sum_{j=k_0}^{k} a^{k-j}\ell_j. \tag{47}$$

For $k \ge j \ge 1$, (46) and $\sqrt{k+1}-\sqrt{j+1} \le \sqrt{k-j}$ give

$$\begin{aligned}\frac{\ell_j}{\ell_k} &\le 2\frac{1+\sqrt{j+1}+\log 2}{1+\sqrt{k+1}}e^{\sqrt{k+1}-\sqrt{j+1}} \\ &\le 2(1+\log 2)e^{\sqrt{k-j}}.\end{aligned} \tag{48}$$

Applying (24) with $q=a$, $\gamma=1$, and $-\ln q=\lambda$, we obtain

$$\begin{aligned}\sum_{j=k_0}^{k} a^{k-j}\ell_j &\le 2(1+\log 2)\ell_k \sum_{h=0}^{k-k_0} a^h e^{\sqrt{h}} \\ &\le \frac{2(1+\log 2)e^{1/(2\lambda)}}{1-\sqrt{a}}\,\ell_k.\end{aligned} \tag{49}$$

For the initial term, completing the square gives, for $h \ge 0$,

$$-\lambda h + \sqrt{h} = \frac{1}{4\lambda} - \lambda\left(\sqrt{h}-\frac{1}{2\lambda}\right)^2 \le \frac{1}{4\lambda}.$$

Together with (48) at $j=k_0$ and $\mathbb{E}[\|r^{k_0}-z^*\|^2] \le D_\mu^2/\mu$, this implies

$$\begin{aligned}a^{k-k_0+1}\mathbb{E}[\|r^{k_0}-z^*\|^2] &\le \frac{2(1+\log 2)aD_\mu^2}{\mu\ell_{k_0}}a^{k-k_0}e^{\sqrt{k-k_0}}\ell_k \\ &\le \frac{2(1+\log 2)aD_\mu^2 e^{1/(4\lambda)}}{\mu\ell_{k_0}}\,\ell_k.\end{aligned} \tag{50}$$

Combining (47), (49), and (50) proves the claim for $k \ge k_0$. For $1 \le k < k_0$, the diameter bound and monotonicity of ℓ_k give

$$\mathbb{E}[\|r^{k+1}-z^*\|^2] \le \frac{D_\mu^2}{\mu} \le \frac{D_\mu^2}{\mu\ell_{k_0}}\ell_k.$$

The definition of C_{rec} now proves the claim for every $k \ge 1$. □

C.12 Proof of Lemma C.3

Proof. Write $s := \sqrt{k+1}$. By (25), $N_k \ge e^s$, and $s \le \log(eT_k)$, we have

$$T_k \le 4se^s \le 4sN_k \le (2e+1)\log(eT_k)N_k,$$

where the last inequality uses $4 \le 2e+1$. This proves the first bound in Lemma C.3 with its stated constant. Since $N_k \le T_k$, it follows that

$$\ell_k = \frac{\log(eN_k)}{N_k} \le \log(eT_k)\frac{(2e+1)\log(eT_k)}{T_k} = (2e+1)\frac{\log^2(eT_k)}{T_k}.$$

□

Table 2: Finite-time last-iterate guarantees under noisy gradient feedback.

Work	Problem class	Guarantee	Rate
Abe et al. (2024)	Monotone games or variational inequalities	Expected gap	$\widetilde{\mathcal{O}}(T^{-1/10})$
Chen & Mazumdar (2024)	Monotone games or variational inequalities	Expected gap	$\widetilde{\mathcal{O}}(T^{-1/6})$
Abe et al. (2025)	Monotone games or variational inequalities	Expected gap	$\widetilde{\mathcal{O}}(T^{-1/7})$
Ito et al. (2026)	Monotone games or variational inequalities	L^2 gap	$\widetilde{\mathcal{O}}(T^{-1/4})$
This work	Two-player zero-sum matrix games	L^2 gap	$\boldsymbol{\widetilde{\mathcal{O}}(T^{-1/2})}$

D Additional Related-Work Comparison

For deterministic bilinear saddle-point problems, optimistic gradient descent ascent and extragradient methods achieve last-iterate convergence (Rakhlin & Sridharan, 2013a;b; Daskalakis et al., 2018; Daskalakis & Panageas, 2019; Gidel et al., 2019; Liang & Stokes, 2019; de Montbrun & Renault, 2022; Golowich et al., 2020b;a; Lee & Kim, 2021; Gorbunov et al., 2022), with tight rates for constrained monotone problems (Mertikopoulos et al., 2019; Wei et al., 2021; Lei et al., 2021; Cai et al., 2022a;b; Yoon & Ryu, 2021; Cevher et al., 2023; Cai & Zheng, 2023). These methods can be viewed as approximations to a proximal-point update (Mokhtari et al., 2020).

Stochastic gradients disrupt the deterministic per-iteration contraction, and existing methods instead provide asymptotic exact convergence or slower polynomial rates (Abe et al., 2023; 2024; Chen & Mazumdar, 2024). The optimistic family also extends to the noisy feedback setting (Hsieh et al., 2019; 2020; Azizian et al., 2021). However, its last-iterate rates degrade for merely monotone problems such as bilinear games. Regularized stochastic approximation schemes instead decay the strength of a strongly convex regularizer over the iterations, and they converge only asymptotically as the regularizer vanishes (Koshal et al., 2010; 2013; Yousefian et al., 2017). Recent payoff-perturbation and regularized-gradient methods improve these finite-time guarantees for smooth monotone games and variational inequalities (Abe et al., 2025; Ito et al., 2026). In contrast, AR-FTRL obtains contraction by tracking an anchor-regularized equilibrium over each epoch, and it recovers the stochastic $\widetilde{\mathcal{O}}(T^{-1/2})$ rate under unbiased feedback with bounded second moment. Table 2 compares this guarantee with prior finite-time last-iterate guarantees under noisy gradient feedback.

Under bandit feedback, Fiegel et al. (2025) establish an $\Omega(t^{-1/4})$ lower bound on the L^2 norm of the gap of the strategy played at each round t, for uncoupled algorithms. Theorem 6.3 attains the faster $\widetilde{\mathcal{O}}(T_k^{-1/2})$ rate for the same L^2 norm of the gap, but only for the strategy z^{T_k+1} played at each epoch boundary, and not for the strategies within an epoch. The lower bound requires the rate at every round, whereas our guarantee holds only at the epoch boundaries. The faster rate comes solely from this weaker requirement.

E Experimental Details

E.1 Payoff matrices of the structured games

The two structured games use payoff matrices normalized so that every entry lies in $[0, 1]$, under the min-max convention of (1). The biased Rock–Paper–Scissors game has the payoff matrix

$$A = \frac{1}{6}\begin{bmatrix} 3 & 4 & 0 \\ 2 & 3 & 4 \\ 6 & 2 & 3 \end{bmatrix},$$

whose unique Nash equilibrium is the interior profile $((1/5, 3/5, 1/5), (1/5, 3/5, 1/5))$. The 5×5 multiple-equilibria game of Wei et al. (2021) has the payoff matrix

$$A = \frac{1}{3}\begin{bmatrix} 2 & 1 & 3 & 3 & 3 \\ 3 & 2 & 1 & 1 & 1 \\ 1 & 3 & 2 & 2 & 2 \\ 2 & 2 & 2 & 0 & 3 \\ 2 & 2 & 2 & 3 & 0 \end{bmatrix}.$$

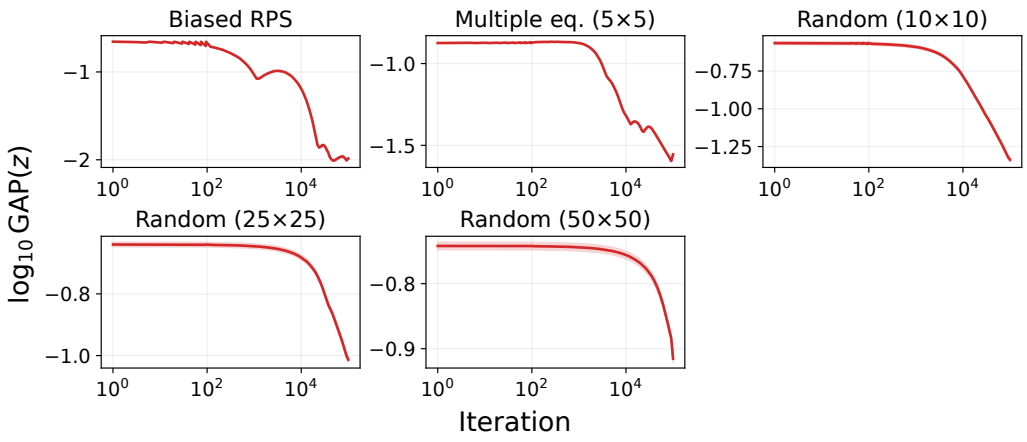


Figure 2: Gap of AR-FTRL under noisy gradient feedback at horizon $T = 10^5$, on a base-10 logarithmic scale against the iteration count. Each curve is the mean over 100 evaluation seeds, and the shaded area represents the standard error.

E.2 Tuning and evaluation protocol

Each tunable baseline selects its learning rate and, when it has one, its regularization strength from the grid $\{1/1000, 1/100, 1/10, 1\}$. We tune over 20 seeds and evaluate over 100 seeds, drawing a fresh payoff matrix for every evaluation seed of the random games. The untuned baselines use the fixed settings prescribed by their analyses.

F Additional Experimental Results

This section reports the experiments under noisy gradient feedback, in the setting of Section 7. We corrupt each coordinate of the gradient operator by independent zero-mean Gaussian noise with standard deviation $1/10$. This noise satisfies Assumption 5.1.

Figure 2 shows the AR-FTRL gap on all five games. The gap decreases on every game, as predicted by Theorems 5.3 and 5.4.