PARAMETER-FREE LAST-ITERATE CONVERGENCE OF OPTIMISTIC AND DISCOUNTED COUNTERFACTUAL REGRET MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual regret minimization (CFR) with regret matching$^+$ local minimizers is the standard way to solve two-player zero-sum games in extensive form (EFGs), but only its average strategy is guaranteed to converge. The reward-transformation framework restores them by solving a sequence of auxiliary games in which every action keeps probability at least γ and a quadratic penalty μ pulls toward a reference strategy; RTCFR$^+$ showed that plain CFR$^+$ converges in the last iterate on every such fixed game for any step size and any initialization. The variants that are actually used, namely predictive CFR$^+$ and the discounted variants DCFR$^+$ and PDCFR$^+$, had no such guarantee. We study one family of local minimizers, predictive discounted regret matching$^+$, that contains all of them. We show that the strategy sequence of any member depends on the step size and the initialization only through their ratio, and use this invariance to prove that in every perturbed regularized EFG the last iterate converges to the unique equilibrium for any step size and any initialization, provided the total discount $\sum_t(1-d_t)$ is finite. The DCFR schedule satisfies the condition exactly when its exponent exceeds one, and without prediction a constant discount cannot converge on nondegenerate matrix games unless it hits the equilibrium in finite time. Every instance has constant regret in the perturbed regularized game, hence an $O(1/T)$ average-iterate rate there, and the last iterate is an $O(\mu+\gamma)$-equilibrium of the original game. Experiments confirm the summability condition and the constant-discount obstruction, show the regret plateau the constant-regret bound predicts, and show that optimism turns the slow last-iterate convergence of regret matching$^+$ into fast convergence: with the published RTCFR$^+$ hyperparameters, RT-PCFR$^+$ ends below 10^{-15} on most benchmark EFGs and converges on a matrix game where both plain PCFR$^+$ and RTCFR$^+$ fail. Parameter-freeness is the property the reward-transformation framework needs, and it survives both optimism and discounting: within each perturbed regularized game the practical variants have the same guarantee as CFR$^+$, with no retuning of step size or initialization.

1 INTRODUCTION

Counterfactual regret minimization (CFR) (Zinkevich et al., 2007), together with its regret-matching$^+$ (RM$^+$) variants (Tammelin, 2014; Brown & Sandholm, 2019a; Farina et al., 2021; Xu et al., 2024b), powers the superhuman poker agents (Bowling et al., 2015; Moravčík et al., 2017; Brown & Sandholm, 2018; 2019b). Their guarantee is a statement about the *average* strategy; the *last* iterate of RM$^+$-based CFR may cycle even in matrix games (Farina et al., 2023; Cai et al., 2025a). Averaging costs memory and, under function approximation, a second network trained on a moving target, so last-iterate guarantees are valuable in practice. The reward-transformation (RT) framework (Perolat et al., 2021; Liu et al., 2023; Abe et al., 2024) obtains them by solving a sequence of *perturbed regularized* games whose equilibria approach the Nash set of the original game.

Meng et al. (2025a) recently proved that CFR$^+$ itself has last-iterate convergence in every such game, for any step size and any initialization of the accumulated regrets, and that this parameter-freeness is what makes the resulting RTCFR$^+$ work: with a reset of the accumulated regrets between stages the method never converges empirically. The variants that dominate the benchmarks are not CFR$^+$.

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.

Predictive CFR$^+$ (PCFR$^+$; Farina et al., 2021) and the discounted variants (DCFR$^+$, PDCFR$^+$; Brown & Sandholm, 2019a; Xu et al., 2024b) are faster in essentially every benchmark, and Meng et al. (2025a) report that an RT variant of PCFR$^+$ is competitive with RTCFR$^+$ but prove nothing about it. The proof for CFR$^+$ does not carry over: prediction adds an error term that the potential of Meng et al. (2025a) does not absorb, and discounting shrinks the accumulated regrets that the potential measures. No last-iterate guarantee in perturbed regularized games was known for either variant; Ren et al. (2025b) give best-iterate rates for discounted regret matching under a step-size condition.

The RT outer loop moves the reference between stages, so the inner solve restarts repeatedly against a changed game with the accumulated regrets it already has. A guarantee that held only for one step size or for a zero initialization would be void from the second stage on. We therefore study one family of local regret minimizers, predictive discounted RM$^+$ (PDRM$^+$), that contains RM$^+$, PRM$^+$ and the two discounted variants as instances (Table 3), and isolate two mechanisms that make the family converge in the last iterate. The first is scale invariance: the update is positively homogeneous in (θ, η), so the strategy sequence depends on the step size η and on the initialization θ^1 only through θ^1/η (Lemma 1), which removes both from every statement. The second is a comparator that shrinks along with the iterate: discounting the accumulated regrets is harmless as long as the accumulated shrinkage $\lambda_\infty = \prod_{t\geq 1} d_t$ stays positive, that is, $\sum_t (1 - d_t) < \infty$ (Assumption 1), and prediction is absorbed by the optimistic three-point inequality on the cone.

Together the two mechanisms give last-iterate convergence in every perturbed regularized game under Assumption 1 (Theorem 1). The DCFR schedule $t^\alpha/(t^\alpha + 1)$ satisfies Assumption 1 exactly when $\alpha > 1$, which includes $\alpha = 1.5$ and the $\alpha = 2.3$ of PDCFR$^+$; for $p = 0$ and a constant discount, Proposition 1 shows on nondegenerate matrix games that no trajectory converges without reaching the equilibrium exactly. Every instance has constant regret in the perturbed regularized game, hence an $O(1/T)$ average-iterate rate there and an $O(\mu + \gamma)$ last iterate in the original game (Theorem 2 and Corollary 1).

Contributions.

- We show that the strategy sequence of any PDRM$^+$ instance depends on the step size η and the initialization θ^1 only through θ^1/η (Lemma 1). For RM$^+$ this is the scaling step of Meng et al. (2025a); we show it survives prediction and discounting, which is what lets every statement below be proved at one convenient step size.
- We prove that in every perturbed regularized EFG the last iterate of any PDRM$^+$ instance converges to the unique equilibrium for any step size and any initialization, provided $\sum_t (1 - d_t) < \infty$ (Theorem 1). The DCFR schedule $t^\alpha/(t^\alpha + 1)$ satisfies Assumption 1 exactly when $\alpha > 1$, which includes $\alpha = 1.5$ and the $\alpha = 2.3$ of PDCFR$^+$; for $p = 0$ and a constant discount, Proposition 1 shows on nondegenerate matrix games that no trajectory converges without reaching the equilibrium exactly.
- We further establish that every instance has constant regret in the perturbed regularized game, that is, the accumulated regrets stay bounded, hence an $O(1/T)$ average-iterate rate there and an $O(\mu + \gamma)$-equilibrium in the original game (Theorem 2 and Corollary 1).
- We test the theory on nine benchmark EFGs and on random normal-form games (Section 6): the discount ablation matches Assumption 1 and Proposition 1, the regret bound plateaus as Theorem 2 predicts, and, with the published RTCFR$^+$ protocol used unchanged for all four instances, RT-PCFR$^+$ drives the last-iterate exploitability below 10^{-15} on six of the nine games and converges on a matrix game where both plain PCFR$^+$ and RTCFR$^+$ fail. The experiments test the inner solver under the protocol it is used with.

2 Related work

Last-iterate convergence in games. Optimistic methods (Rakhlin & Sridharan, 2013; Syrgkanis et al., 2015) converge in the last iterate in unconstrained or well-conditioned settings (Daskalakis & Panageas, 2019; Wei et al., 2021; Gorbunov et al., 2022), and Lee et al. (2021) extended this to EFGs through dilated regularizers; Cai et al. (2024) prove that forgetful dynamics are necessary for fast last-iterate rates. Regularization is the other route: entropy- or distance-regularized dynamics

converge to a regularized equilibrium, and letting the regularization vanish, or moving its reference point, recovers a Nash equilibrium (Perolat et al., 2021; Pérolat et al., 2022; Liu et al., 2023; Abe et al., 2022; 2023; 2024; Sokota et al., 2023; Liu et al., 2024b; Bernasconi et al., 2024). Our analysis lives in the second family and is agnostic to how the outer loop is scheduled. Cai et al. (2025b) convert average-iterate guarantees into last-iterate ones by a generic reduction that changes the algorithm; we analyze the unchanged practical algorithms. Lower bounds for last-iterate rates under uncoupled bandit feedback (Fiegel et al., 2025) do not apply to the full-feedback self-play setting studied here.

Regret matching and its last iterate. Farina et al. (2021) identified RM^+ with mirror descent on the cone and introduced predictive RM^+; Farina et al. (2023) exhibited the instability of RM^+ and proposed smoothed and extragradient variants; Cai et al. (2025a) proved that these smoothed variants converge in the last iterate in matrix games while RM^+ and PRM^+ may not, and Meng et al. (2025b) extended the smooth variants to games satisfying a weak Minty condition. Zhang et al. (2025) obtain scale-invariant regret matching with constant regret through a different mechanism (a state-dependent normalization). Closest to us, Meng et al. (2025a) proved that CFR^+ converges in the last iterate in perturbed regularized EFGs for any step size and initialization and built RTCFR^+ on it; Ren et al. (2025b;a) add adaptive schedules for the regularization and the perturbation and report best-iterate rates for discounted regret matching under a step-size condition. We supply the parameter-free last-iterate theory for the predictive and discounted instances that these works use empirically, and identify the summability condition on the discount, with an obstruction for constant discounts.

Discounting. Discounted CFR (Brown & Sandholm, 2019a) and its predictive form (Xu et al., 2024b) weigh recent regrets more heavily; learned or scheduled variants (Xu et al., 2022; 2024a; Zhang et al., 2026; Xu et al., 2026) tune the schedule. All of them use the form $t^\alpha/(t^\alpha+1)$ with a fixed or scheduled exponent; Assumption 1 asks for a fixed $\alpha > 1$, or for a scheduled exponent that eventually stays above $1+\epsilon$ for some $\epsilon > 0$ (pointwise $\alpha_t > 1$ is not enough: $\alpha_t = 1 + 1/\log t$ gives $t^{\alpha_t} = et$), and Proposition 1 rules out constant discounts.

3 Preliminaries

Extensive-form games. We consider two-player zero-sum EFGs, all with perfect recall. Player $i \in \{0,1\}$ acts at infosets $I \in \mathcal{I}_i$ with action set $A(I)$; $A_{\max} = \max_I |A(I)|$, H is the maximal number of decision points of one player on a path, $C_{\max}$ the maximal number of child infosets of a sequence, and $n = \sum_I |A(I)|$ the number of sequences. A behavioral strategy σ_i assigns $\sigma_i(I) \in \Delta^{|A(I)|}$ to every $I \in \mathcal{I}_i$; its sequence form $x_i \in \mathcal{X}_i$ has entries $x_i(I,a) = \pi_i^\sigma(I)\,\sigma_i(I,a)$, where $\pi_i^\sigma(I)$ multiplies player i's own probabilities along the path to I. Payoffs are bilinear, $x_0^\top A x_1$, player 0 minimizing. For a profile x the loss vectors are $\ell_0^x = Ax_1$ and $\ell_1^x = -A^\top x_0$; we write $\|\ell^x - \ell^{x'}\|_1 \le L\|x - x'\|_1$, $\|\ell^x\|_1 \le P$, $D = \max_{x,x' \in \mathcal{X}} \|x - x'\|_2$ and M the largest absolute terminal payoff. The exploitability of x is $\mathrm{expl}(x) = \max_{x_1'} x_0^\top A x_1' - \min_{x_0'} x_0'^\top A x_1$.

Perturbed regularized EFGs. Following Perolat et al. (2021); Liu et al. (2023); Abe et al. (2024); Meng et al. (2025a), the reward-transformation (RT) framework replaces the original game by a sequence of games that are both perturbed and regularized. For $\gamma > 0$ let $\Delta_\gamma^k = \{\sigma \in \Delta^k : \sigma_a \ge \gamma\}$ and let $\mathcal{X}_i^\gamma$ be the sequence-form strategies whose behavioral strategy lies in Δ_γ at every infoset. For $\mu > 0$, a reference $r = (r_0, r_1) \in \mathcal{X}$ and $\psi = \frac{1}{2}\|\cdot\|_2^2$, the perturbed regularized game is

$$\min_{\hat{x}_0 \in \mathcal{X}_0^\gamma} \max_{\hat{x}_1 \in \mathcal{X}_1^\gamma} \; \hat{x}_0^\top A \hat{x}_1 + \mu D_\psi(\hat{x}_0, r_0) - \mu D_\psi(\hat{x}_1, r_1). \tag{1}$$

Its Nash equilibrium (NE) $\hat{x}^* = \hat{x}^{*,\gamma,\mu,r}$ is unique. The losses of (1) are $\hat{\ell}_0^{\hat{x}} = A\hat{x}_1 + \mu(\hat{x}_0 - r_0)$ and $\hat{\ell}_1^{\hat{x}} = -A^\top \hat{x}_0 + \mu(\hat{x}_1 - r_1)$, and the counterfactual values are $\hat{v}_i^{\hat{x}}(I,a) = -\hat{\ell}_i^{\hat{x}}(I,a) + \sum_{I' \in C_i(I,a)} \langle \hat{v}_i^{\hat{x}}(I'), \hat{\sigma}_i(I') \rangle$. The set Δ_γ is parametrized by the full simplex through

$$\hat{\sigma}_i(I) = (1 - \alpha_I)\,\sigma_i(I) + \gamma \mathbf{1}, \qquad \alpha_I = |A(I)|\gamma \le \tfrac{1}{2}, \qquad \sigma_i(I) \in \Delta^{|A(I)|}, \tag{2}$$

a bijection with $\hat{\sigma} - \hat{\sigma}' = (1 - \alpha_I)(\sigma - \sigma')$ (we assume $\gamma \le 1/(2A_{\max})$ throughout, so $1 - \alpha_I \ge \frac{1}{2}$; the RT loop only ever shrinks γ). We write $\sigma_i^*(I)$ for the preimage of $\hat{\sigma}_i^*(I)$. Regularization makes the game μ-strongly monotone; perturbation makes behavioral strategies Lipschitz in the sequence

form on $\mathcal{X}^\gamma$. The RT outer loop shrinks γ and moves r to the current iterate; the works above establish that the sequence of NEs $\hat{x}^{*,\gamma,\mu,r}$ approaches the NE set of the original game.

CFR and regret matching$^+$. CFR (Zinkevich et al., 2007) decomposes the sequence-form regret into local regrets at infosets and minimizes each of them separately. The instantaneous counterfactual regret at I is $m_i^t(I) = v_i^t(I) - \langle v_i^t(I), \sigma_i^t(I)\rangle \mathbf{1}$. Regret matching$^+$ (RM$^+$) (Tammelin, 2014) keeps $\theta_I^{t+1} = [\theta_I^t + m_i^t(I)]_+$ and plays $\sigma_i^{t+1}(I) = \theta_I^{t+1}/\langle \theta_I^{t+1}, \mathbf{1}\rangle$; Farina et al. (2021) showed that this is online mirror descent on the cone $\mathbb{R}^{|A(I)|}_{\geq 0}$ with any step size $\eta > 0$, that predicting the next regret by the current one gives PRM$^+$ (and PCFR$^+$), and that, when $\theta^1 = 0$, the strategy sequence does not depend on η. Discounting the accumulated regrets by $t^\alpha/(t^\alpha + 1)$ gives DCFR$^+$ and PDCFR$^+$ (Brown & Sandholm, 2019a; Xu et al., 2024b).

4 Predictive discounted regret matching$^+$

We analyze one family of local regret minimizers that contains the four algorithms practitioners actually run. At every infoset $I \in \mathcal{I}_i$ the learner keeps a vector $\theta_I^t \in \mathbb{R}^{|A(I)|}_{\geq 0}$ (accumulated regret in the cone), and is specified by a step size $\eta > 0$, an initialization $\theta_I^1 \geq 0$, a discount sequence $(d_t)_{t\geq 1} \subset (0, 1]$, a prediction flag $p \in \{0, 1\}$ and a bounded initial prediction $\hat{m}^0$ (default 0). Write $[\cdot]_+ = \max(\cdot, 0)$ componentwise and $\mathrm{nrm}(\theta) = \theta/\langle \theta, \mathbf{1}\rangle$ with $\mathrm{nrm}(0)$ the uniform distribution. For $t \geq 1$ and all infosets simultaneously,

$$\tilde{\theta}_I^t = \left[d_t \theta_I^t + \eta\, p\, \hat{m}_i^{t-1}(I)\right]_+, \qquad \sigma_i^t(I) = \mathrm{nrm}(\tilde{\theta}_I^t), \qquad \hat{\sigma}_i^t(I) = (1 - \alpha_I)\sigma_i^t(I) + \gamma \mathbf{1}, \quad (3)$$

$$\hat{m}_i^t(I) = \hat{v}_i^t(I) - \langle \hat{v}_i^t(I), \sigma_i^t(I)\rangle \mathbf{1}, \qquad \theta_I^{t+1} = \left[d_t \theta_I^t + \eta\, \hat{m}_i^t(I)\right]_+, \quad (4)$$

where $\hat{v}_i^t(I) = \hat{v}_i^{\hat{x}^t}(I)$ are the counterfactual values of the perturbed regularized game (1) at the perturbed profile $\hat{x}^t$. We call (3)–(4) PDRM$^+$ (predictive discounted RM$^+$), and the CFR algorithm that runs it at every infoset PDCFR$^+$. The strategy that is played, and that enters the counterfactual values, is the perturbed $\hat{\sigma}^t$; the instantaneous regret uses the unperturbed σ^t, exactly as in RTCFR$^+$ (Meng et al., 2025a). Table 3 in Appendix A.1 lists the instances.

Two identities drive everything that follows. Because $\sigma^t = \mathrm{nrm}(\tilde{\theta}^t)$,

$$\langle \hat{m}_i^t(I), \tilde{\theta}_I^t\rangle = 0, \qquad \langle \hat{m}_i^t(I), u\rangle = \langle u, \mathbf{1}\rangle \langle \hat{v}_i^t(I), \mathrm{nrm}(u) - \sigma_i^t(I)\rangle \quad (u \geq 0,\ u \neq 0), \quad (5)$$

so in particular $\langle \hat{m}_i^t(I), \sigma_i^*(I)\rangle = \langle -\hat{v}_i^t(I), \sigma_i^t(I) - \sigma_i^*(I)\rangle$: the regret against the equilibrium point of the cone is the local progress toward σ^*.

The update is positively homogeneous in the pair (θ, η), which Lemma 1 turns into the parameter-freeness that the rest of the paper relies on.

Scale invariance. The PDRM$^+$ update (3)–(4) is positively homogeneous in the pair (θ, η), and the normalization $\mathrm{nrm}(\cdot)$ removes the scale, so the step size and the initialization of the accumulated regrets enter only through one ratio.

Lemma 1 (scale invariance)**.** *Fix the game, (d_t), p and $\hat{m}^0$. The* PDRM$^+$ *trajectories started from (θ^1, η) and from $(c\theta^1, c\eta)$, $c > 0$, satisfy $\theta'^t = c\theta^t$, $\tilde{\theta}'^t = c\tilde{\theta}^t$ and $\sigma'^t = \sigma^t$, $\hat{m}'^t = \hat{m}^t$ for all t. Hence the strategy sequence depends on (θ^1, η) only through θ^1/η.*

The proof is a two-line induction on t and uses only that $[\cdot]_+$ and $\mathrm{nrm}(\cdot)$ commute with positive scaling (Appendix A.2). For RM$^+$ this is the "linearity of the projection" step of Meng et al. (2025a), and the step-size independence of Farina et al. (2021) is the case $\theta^1 = 0$. What the lemma adds is that the identity survives prediction and discounting: the forecast $\eta p\, \hat{m}^{t-1}$ scales with η because $\hat{m}$ is a function of the play alone, and the discount multiplies θ, so both commute with the scaling.

Its use is methodological. Any property of the *strategy* sequence that is proved for one step size η_0, for every initialization $\theta^1 \geq 0$ and for the given $(d_t), p$ holds for every η, because the run (θ^1, η) is the run $(\eta_0\theta^1/\eta, \eta_0)$ and the two produce the same play. Every statement we prove below therefore holds for every step size and every initialization, and may be proved for whichever step size the argument requires; the constants depend on the effective initialization θ^1/η. Appendix C confirms the lemma numerically: runs with equal θ^1/η produce exploitability curves that agree to floating-point roundoff on Kuhn, while runs with different ratios do not.

5 ANALYSIS

Last-iterate convergence. Let $\lambda_t = \prod_{s<t} d_s$ ($\lambda_1 = 1$) and $\lambda_\infty = \prod_{s\geq 1} d_s$. The hypothesis the proof places on the discount is that the accumulated shrinkage stays bounded below.

Assumption 1 (finite total discount)**.** $\sum_{t\geq 1}(1-d_t) < \infty$*; equivalently* $\lambda_\infty > 0$.

The schedules that satisfy it are characterized below, along with the obstruction for constant discounts.

Theorem 1 (invariant last-iterate convergence)**.** *Let $\mu > 0$, $0 < \gamma \leq 1/(2A_{\max})$ and $r \in \mathcal{X}$. Suppose every infoset of both players runs* PDRM$^+$ *with a common discount sequence satisfying Assumption 1, a common $p \in \{0, 1\}$, an arbitrary step size $\eta > 0$, arbitrary initializations $\theta_I^1 \in \mathbb{R}_{\geq 0}^{|A(I)|}$ and bounded initial predictions. Then $\sum_{t\geq 1} \|\hat{x}^t - \hat{x}^*\|_2^2 < \infty$; in particular the last iterate $\hat{x}^t$ converges to the unique NE of* (1)*, and $\hat{\sigma}_i^t(I) \to \hat{\sigma}_i^*(I)$ at every infoset.*

Theorem 1 covers RTCFR$^+$ (Meng et al., 2025a, $d \equiv 1$, $p = 0$) and is new for $p = 1$ (RT-PCFR$^+$) and for every discounted instance. The infrastructure is theirs: the CFR decomposition, strong monotonicity, the local best-response property of the NE, the smoothness of the instantaneous regrets and, for $p = 0$, the potential Ψ below. New are the comparator $\lambda_{t+1}\sigma^*$ that shrinks with the iterate, which lets discounting enter only through $d_t^2 \leq 1$ and $\lambda_\infty > 0$; the optimistic three-point inequality on the cone (Lemma 2), whose error term is a prediction error that the theorem itself makes square-summable; and the boundedness of the cone iterates (Theorem 2), which turns convergence into constant regret for the whole family. The proof (Appendix A.3) has two steps, and both are short because the right object is analyzed: the cone iterate against a comparator that is *scaled along with it*.

Step 1: a small step size. Since $[\cdot]_+$ is the Euclidean projection onto the cone $K = \mathbb{R}_{\geq 0}^k$, (4) is a projected (optimistic) gradient step from the base point $d_t\theta^t$. The standard optimistic three-point inequality on K reads, for every $u \in K$,

$$\eta\langle \hat{m}^t, u - \tilde{\theta}^t\rangle \leq D_\psi(u, d_t\theta^t) - D_\psi(u, \theta^{t+1}) - D_\psi(\theta^{t+1}, \tilde{\theta}^t) + \eta\langle \hat{m}^t - p\,\hat{m}^{t-1}, \theta^{t+1} - \tilde{\theta}^t\rangle. \quad (6)$$

We take $u = u_{t+1} := \lambda_{t+1}\sigma_i^*(I)$, i.e. the equilibrium point of the cone shrunk by the same factor as the iterate. Then $D_\psi(u_{t+1}, d_t\theta^t) = d_t^2 D_\psi(u_t, \theta^t) \leq D_\psi(u_t, \theta^t)$, so the potential $\Phi_t(I) = D_\psi(u_t, \theta_I^t)$ telescopes; by (5) the left side is $\eta\lambda_{t+1}\langle -\hat{v}_i^t(I), \sigma_i^t(I) - \sigma_i^*(I)\rangle$, and Young's inequality turns the last term into $\frac{\eta^2}{2}\|\hat{m}^t - p\hat{m}^{t-1}\|^2$. For $p = 1$ this error is a prediction error, bounded through the smoothness of the perturbed game by $C(\|\hat{x}^t - \hat{x}^*\|^2 + \|\hat{x}^{t-1} - \hat{x}^*\|^2)$; for $p = 0$ it does not vanish and we use instead the potential $D_\psi(u_t, \theta^t) + \eta\langle -\hat{m}_i^*(I), \theta_I^t\rangle$ of Meng et al. (2025a), for which discounting only adds a nonpositive term (so that (7) holds with Ψ in place of Φ and without the $t-1$ term). Summing over infosets with the weights $\zeta_I = (1-\alpha_I)\pi_i^{\hat{\sigma}^*}(I)$ converts the local progress into the sequence-form quantity $\langle \hat{\ell}_i^t, \hat{x}_i^t - \hat{x}_i^*\rangle$ (the CFR decomposition), and strong monotonicity of (1) gives

$$\eta\lambda_\infty\mu\,\|\hat{x}^t - \hat{x}^*\|_2^2 \;\leq\; \Phi_t - \Phi_{t+1} + \eta^2 C\big(\|\hat{x}^t - \hat{x}^*\|_2^2 + \|\hat{x}^{t-1} - \hat{x}^*\|_2^2\big), \quad (7)$$

which sums to $\sum_t \|\hat{x}^t - \hat{x}^*\|^2 \leq 2(\Phi_1 + \eta^2 E_0)/(\eta\lambda_\infty\mu)$ as soon as $\eta \leq \eta_0 := \lambda_\infty\mu/(4C)$.

Step 2: every step size. Given any (θ^1, η), Lemma 1 identifies the strategy sequence with that of $(\eta_0\theta^1/\eta, \eta_0)$, to which Step 1 applies. Nothing else is needed; in particular the result does not rely on the zero initialization used by Farina et al. (2021).

Remark 1 (where the discount enters)**.** Discounting appears in the proof only through $D_\psi(d_t u_t, d_t\theta^t) = d_t^2 D_\psi(u_t, \theta^t)$ and the factor $\lambda_{t+1} \geq \lambda_\infty$ in front of the progress term. If $\lambda_t \to 0$ the comparator collapses to the origin and the argument yields nothing. Proposition 1 shows that the obstruction is intrinsic to the dynamics.

Constant regret and approximation of the original game. Theorem 1 is a statement about the play. Two properties follow, one about the regret of every instance inside the perturbed regularized game, and one about what the last iterate is worth in the original game.

Theorem 2 (bounded cone iterates and constant regret)**.** *Under the hypotheses of Theorem 1, with $S := \sum_{t\geq 1}(1-d_t) < \infty$:*

(i) every accumulated regret is bounded: $B_I := \sup_t \|\theta_I^t\|_\infty < \infty$ *for all* I*;*

(ii) $\sum_{t \le T} \hat{m}_i^t(I, a) \le B_I(1 + S)/\eta$ *for all* I*,* a *and* T*;*

(iii) the sequence-form regret of each player against the losses of (1)*,* $R_i^T = \max_{x_i \in \mathcal{X}_i^\gamma} \sum_{t \le T} \langle \hat{\ell}_i^t, \hat{x}_i^t - x_i \rangle$*, satisfies* $\sup_T R_i^T \le (1 + S) \sum_{I \in \mathcal{I}_i} B_I/\eta < \infty$*, and the average iterate* $\bar{x}^T = \frac{1}{T} \sum_{t \le T} \hat{x}^t$ *is an* $O(1/T)$*-NE of* (1)*.*

The constants are explicit (Appendix A.4) and depend on (θ^1, η) *only through* θ^1/η*.*

The mechanism is the one that makes RM$^+$ work in the first place: the regret against any fixed action is at most the accumulated regret $\theta^{T+1}(I, a)/\eta$ plus what discounting removed, so bounded cone iterates mean constant regret. Boundedness is a by-product of the potential argument behind Theorem 1, and it holds for the optimistic and the plain instance alike. Optimism therefore does not decide *whether* the regret is constant; it decides how fast the last iterate settles and hence how early the plateau is reached (Figure 2, bottom). For $p = 1$ the optimistic bound $D_\psi(u, \theta^1)/\eta + \frac{\eta}{2} \sum_t \|\hat{m}^t - \hat{m}^{t-1}\|^2$ of Farina et al. (2021) gives the same conclusion by a different route, because Theorem 1 makes the prediction errors square-summable.

Corollary 1 (approximation of the original game)**.** $\mathrm{expl}(\hat{x}^{*,\gamma,\mu,r}) \le \mu D^2 + 4MHA_{\max}\gamma$*, where* M *is the largest absolute terminal payoff.*

Hence the last iterate of any PDRM$^+$-based CFR is, after the transient of Theorem 1, an $O(\mu+\gamma)$-NE of the original EFG, and the RT outer loop drives this error to zero by shrinking γ and moving r; we use exactly the outer loop of RTCFR$^+$ in Section 6.

The discount threshold. Assumption 1 is the only condition the convergence proof places on the discount. The undiscounted case $d_t \equiv 1$ qualifies trivially. For the DCFR schedule $d_t = t^\alpha/(t^\alpha + 1)$ one has $1 - d_t = 1/(t^\alpha + 1)$, so Assumption 1 holds if and only if $\alpha > 1$: the $\alpha = 1.5$ of DCFR$^+$ and the $\alpha = 2.3$ of PDCFR$^+$ (Xu et al., 2024b) qualify, while $\alpha = 1$ gives $\lambda_{t+1} = 1/(t + 1) \to 0$ and a constant discount $d < 1$ gives $\lambda_t = d^{t-1} \to 0$. This section shows that the exclusion of constant discounts is forced by the dynamics; whether every schedule with $\lambda_t \to 0$ fails is open.

Proposition 1 (constant discount, $p = 0$)**.** *Consider a two-player zero-sum game in normal form perturbed and regularized as in* (1)*, with any* $0 \le \gamma \le 1/(2A_{\max})$*, whose NE* σ^* *is fully mixed, and let* $F(\sigma) = (\hat{m}_0(\sigma), \hat{m}_1(\sigma))$ *with Jacobian* J *at* σ^**. Assume the tangent-space map* ΠJ *is invertible and that no nonzero tangent direction of one player leaves the other player's regret unchanged (Appendix B). If* $d_t \equiv d \in (0, 1)$ *and* $p = 0$*, then* $\sigma^t \to \sigma^*$ *implies* $\sigma^t = \sigma^*$ *for all large* t*: no trajectory converges to the equilibrium without reaching it exactly in finite time.*

The mechanism is elementary. Convergence forces $\hat{m}^t \to 0$, hence $\theta^t \to 0$, so the normalized update behaves like a gradient step with step size $\eta/\|\theta^t\|_1 \to \infty$: linearizing at σ^*, each step multiplies the deviation by at least $\eta s/(2\|\theta^{t+1}\|_1) \to \infty$, which is incompatible with convergence unless the deviation is exactly zero. Regularized matching pennies satisfies both nondegeneracy conditions for every $\mu \ge 0$ and $\gamma < \frac{1}{2}$ (Appendix B). For $p = 1$ the linearized recurrence has one contracting direction; empirically every constant discount we tried fails for both p (Section 6).

6 Experiments

All algorithms are implemented in one code path on top of LiteEFG (Liu et al., 2024a) and OpenSpiel (Lanctot et al., 2019); our RM$^+$ instance reproduces the released RTCFR$^+$ implementation of Meng et al. (2025a) to 10^{-15}. Games: Kuhn and Leduc poker, imperfect-information Goofspiel(4,5), Liar's Dice played with 4-, 5- or 6-sided dice, and Battleship on 2×2 and 2×3 boards (Appendix C). Every algorithm is deterministic, so there are no error bars; every number below is produced by the scripts in the supplement from the logged runs. All four PDRM$^+$ instances share one computation graph, so an iteration costs the same for each; the baselines use LiteEFG's own graphs of the same size.

Optimism decides the speed. We fix $\gamma = \mu = 10^{-2}$ and r uniform, compute $\hat{x}^*$ by a long RM$^+$ run, and report the sequence-form distance $\|\hat{x}^t - \hat{x}^*\|_2$ of RM$^+$, PRM$^+$, DRM$^+$ and PDRM$^+$ ($\alpha = 1.5$) under simultaneous updates (the regime of the theory) and under the alternating updates

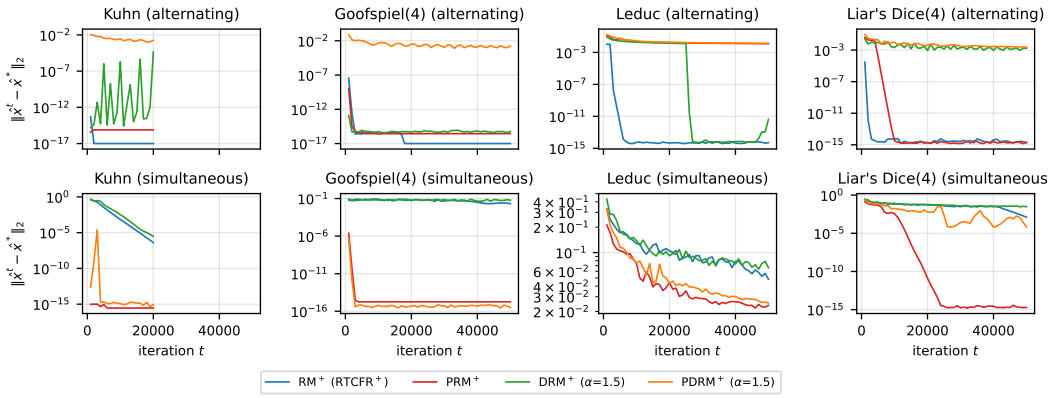

Figure 1: Distance to the equilibrium of a fixed perturbed regularized game ($\gamma = \mu = 10^{-2}$) for the four PDRM$^+$ instances, simultaneous (bottom) and alternating (top) updates. All converge (Theorem 1); optimism determines the speed under simultaneous updates.

used in practice (Figure 1). All four converge under simultaneous updates, as Theorem 1 predicts, and also under alternating updates, which the theorem does not cover; the speeds differ widely: under simultaneous updates prediction is decisive (on Goofspiel(4) PRM$^+$ reaches 10^{-14} within 5,000 iterations while RM$^+$ is still at 10^{-1} after 50,000; on Liar's Dice(4) PRM$^+$ reaches 10^{-14} while the other three stay above 10^{-4}), whereas under alternating updates RM$^+$ often locks onto the equilibrium exactly after a few thousand iterations and the predictive variants converge polynomially (log-log slopes between -0.3 and -0.8 on Leduc). On Kuhn, runs with different (θ^1, η) but equal θ^1/η coincide to floating-point roundoff ($\leq 2 \cdot 10^{-15}$), as Lemma 1 requires (Appendix C).

The discount ablation matches the condition. Figure 2 (top) varies the discount on Kuhn (50,000 iterations, both update orders, $\theta^1 \in \{0, 5\mathbf{1}\}$, with and without prediction). Every constant discount $d \in \{0.99, 0.999, 0.9999\}$ and the schedule with $\alpha = 0.5$ fail to converge (the behavioral distance in the second half of the run stays between 10^{-2} and 2.5); $\alpha = 1$, for which $\lambda_t \to 0$ but slowly, is borderline; $\alpha \geq 1.5$ and $d \equiv 1$ converge. This is what Assumption 1 and Proposition 1 predict, and it rules out the natural conjecture that any non-decreasing discount sequence would do; the run cannot separate $\alpha = 1$ from slow convergence. On regularized matching pennies, where Proposition 1 applies, the same picture holds for $p = 0$ and $p = 1$ and the effective step size $\eta/\|\theta^t\|_1$ indeed grows (Appendix C).

The regret plateaus once the iterate settles. Figure 2 (bottom) tracks the sequence-form regret upper bound of CFR, $\sum_I [\max_a \sum_{s \leq t} \hat{m}^s(I, a)]_+$, in the fixed game of Figure 1. For PRM$^+$ it is flat from the first few hundred iterations on (Kuhn: 1.476 from iteration 500 to 20,000; Goofspiel(4): 7.283), while for RM$^+$ it keeps growing (Goofspiel(4): 10.9 $\to$ 13.4; Leduc: 36.5 $\to$ 51.8), as Theorem 2 predicts: the bound is at most $\sum_I \|\theta_I^{T+1}\|_\infty/\eta$, so it plateaus exactly when the accumulated regrets stop moving, which optimism brings about much earlier. On Leduc, where PRM$^+$ has not yet converged after 20,000 iterations, its bound still grows slowly (35.9 $\to$ 40.5).

Published RTCFR$^+$ settings transfer without retuning. The RT protocol changes the game every T_u iterations and alternates, so this experiment tests the inner solvers in the setting they are used in. We run the outer loop of RTCFR$^+$ unchanged ($\gamma_0 = 10^{-10}$, $\mu = 10^{-3}$, every $T_u = 100$ iterations $\gamma \leftarrow \gamma/2$ and $r \leftarrow \hat{x}^t$, accumulated regrets carried over, alternating updates) for all four PDRM$^+$ instances, with no per-algorithm tuning (a sweep over (γ_0, μ, T_u) confirmed that these published values are the best ones for RTCFR$^+$ itself; Appendix C). Baselines with last-iterate guarantees are Reg-CFR (Liu et al., 2023), dilated optimistic mirror descent (DOMD; Lee et al., 2021) and magnetic mirror descent (MMD; Sokota et al., 2023) with the LiteEFG default hyperparameters; we also report the last iterates of CFR$^+$ and PCFR$^+$ without any regularization. Figure 3 and Table 1 show how the exploitability of the last iterate evolves in the original game over 20,000 iterations.

At their LiteEFG default settings, Reg-CFR, DOMD and MMD end one to sixteen orders of magnitude above RT-PCFR$^+$ on every game; each carries a fixed regularization weight (or a fixed reference that never moves) and therefore converges to a regularized equilibrium whose exploitability floor is

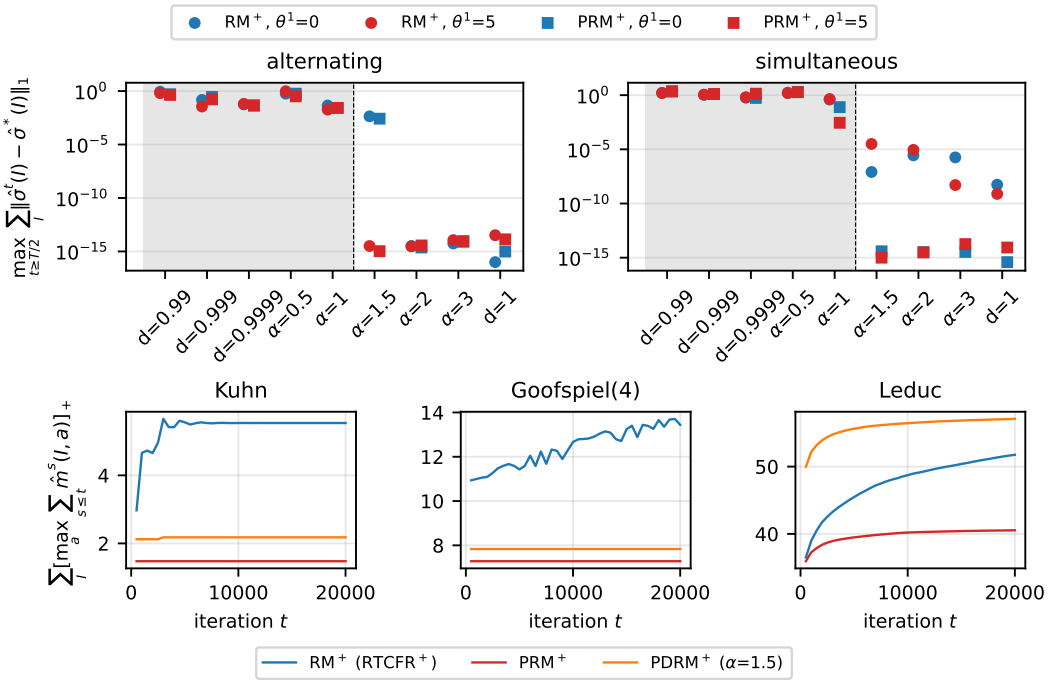

Figure 2: Two checks of the mechanism in the fixed perturbed regularized game. Top: worst behavioral distance $\sum_I \|\hat{\sigma}^t(I) - \hat{\sigma}^*(I)\|_1$ over the second half of 50,000 iterations on Kuhn as a function of the discount sequence (shaded: $\sum_t (1 - d_t) = \infty$); only schedules with finite total discount converge. Bottom: CFR regret bound; both bounds are eventually constant (Theorem 2); the optimistic instance reaches its plateau within a few hundred iterations, RM$^+$ has not within 20,000.

Table 1: Last-iterate exploitability after 20,000 iterations (bold: best RT variant).

game	RT-CFR$^+$	RT-PCFR$^+$	RT-DCFR$^+$	RT-PDCFR$^+$	Reg-CFR	DOMD	MMD	PCFR$^+$ (last)
Kuhn	$3.3{\cdot}10^{-15}$	$<\mathbf{10^{-15}}$	$2.4{\cdot}10^{-15}$	$2.4{\cdot}10^{-4}$	$2.1{\cdot}10^{-1}$	$9.8{\cdot}10^{-2}$	$1.4{\cdot}10^{-2}$	$<10^{-15}$
Leduc	$\mathbf{2.1{\cdot}10^{-8}}$	$1.2{\cdot}10^{-6}$	$4.2{\cdot}10^{-5}$	$2.3{\cdot}10^{-4}$	$4.1{\cdot}10^{0}$	$1.1{\cdot}10^{0}$	$3.8{\cdot}10^{-2}$	$7.6{\cdot}10^{-6}$
Goofspiel(4)	$3.1{\cdot}10^{-9}$	$<\mathbf{10^{-15}}$	$1.5{\cdot}10^{-15}$	$<\mathbf{10^{-15}}$	$4.6{\cdot}10^{-1}$	$3.2{\cdot}10^{-5}$	$5.3{\cdot}10^{-2}$	$<10^{-15}$
Goofspiel(5)	$3.5{\cdot}10^{-5}$	$<\mathbf{10^{-15}}$	$1.2{\cdot}10^{-6}$	$2.0{\cdot}10^{-3}$	$1.6{\cdot}10^{0}$	$1.7{\cdot}10^{-1}$	$1.1{\cdot}10^{-1}$	$5.5{\cdot}10^{-9}$
Liar's Dice(4)	$<\mathbf{10^{-15}}$	$<\mathbf{10^{-15}}$	$<\mathbf{10^{-15}}$	$<\mathbf{10^{-15}}$	$1.3{\cdot}10^{0}$	$3.6{\cdot}10^{-2}$	$5.3{\cdot}10^{-2}$	$<10^{-15}$
Liar's Dice(5)	$7.5{\cdot}10^{-7}$	$2.9{\cdot}10^{-5}$	$\mathbf{4.2{\cdot}10^{-7}}$	$3.1{\cdot}10^{-6}$	$1.5{\cdot}10^{0}$	$2.7{\cdot}10^{-1}$	$4.9{\cdot}10^{-2}$	$<10^{-15}$
Liar's Dice(6)	$2.8{\cdot}10^{-4}$	$6.2{\cdot}10^{-3}$	$\mathbf{1.6{\cdot}10^{-4}}$	$4.8{\cdot}10^{-4}$	$1.7{\cdot}10^{0}$	$3.2{\cdot}10^{-1}$	$6.2{\cdot}10^{-2}$	$5.2{\cdot}10^{-4}$
Battleship(2×2)	$2.4{\cdot}10^{-3}$	$<\mathbf{10^{-15}}$	$9.5{\cdot}10^{-4}$	$1.6{\cdot}10^{-3}$	$4.6{\cdot}10^{-1}$	$2.9{\cdot}10^{-8}$	$1.6{\cdot}10^{-6}$	$<10^{-15}$
Battleship(2×3)	$4.2{\cdot}10^{-3}$	$<\mathbf{10^{-15}}$	$2.1{\cdot}10^{-3}$	$1.4{\cdot}10^{-3}$	$5.0{\cdot}10^{-1}$	$9.3{\cdot}10^{-2}$	$2.5{\cdot}10^{-2}$	$5.1{\cdot}10^{-15}$

set by that weight. RT-PCFR$^+$ ends below 10^{-15} on six of the nine games; RTCFR$^+$ is ahead on Leduc and Liar's Dice(5,6) by factors of 22 to 57, and across the nine RTCFR$^+$ ends between 10^{-16} and 10^{-2}. Discounting does not help under the RT loop: the reference moves every 100 iterations, and forgetting the regrets accumulated against the previous reference slows the inner solve. The last iterate of plain PCFR$^+$, with no regularization and no guarantee, is excellent on these benchmarks, a fact also visible in Meng et al. (2025a); the guarantee shows on instances where the unregularized dynamics fail, which Table 2 examines.

Regularization matters when the unregularized dynamics cycle. Cai et al. (2025a) give a 3×3 game on which the last iterate of PRM$^+$ does not converge. Table 2 compares, under the simultaneous updates of the theory and with $\mu = 10^{-2}$ (the EFG protocol uses 10^{-3}; at $\mu = 10^{-3}$ the regularized inner problems are too weakly monotone for either RT method to make progress in 20,000 iterations; Appendix C), the plain and RT versions of CFR$^+$, PCFR$^+$ and PDCFR$^+$ on that game and on random 3×3 and 4×4 zero-sum games; Figure 4 in Appendix C plots the same runs. On the Cai game plain PCFR$^+$ cycles (Figure 4, left) and ends at exploitability 0.77, RTCFR$^+$ ends at 0.30, and RT-PCFR$^+$ converges to 10^{-16}. Under alternating updates the picture changes and plain PCFR$^+$

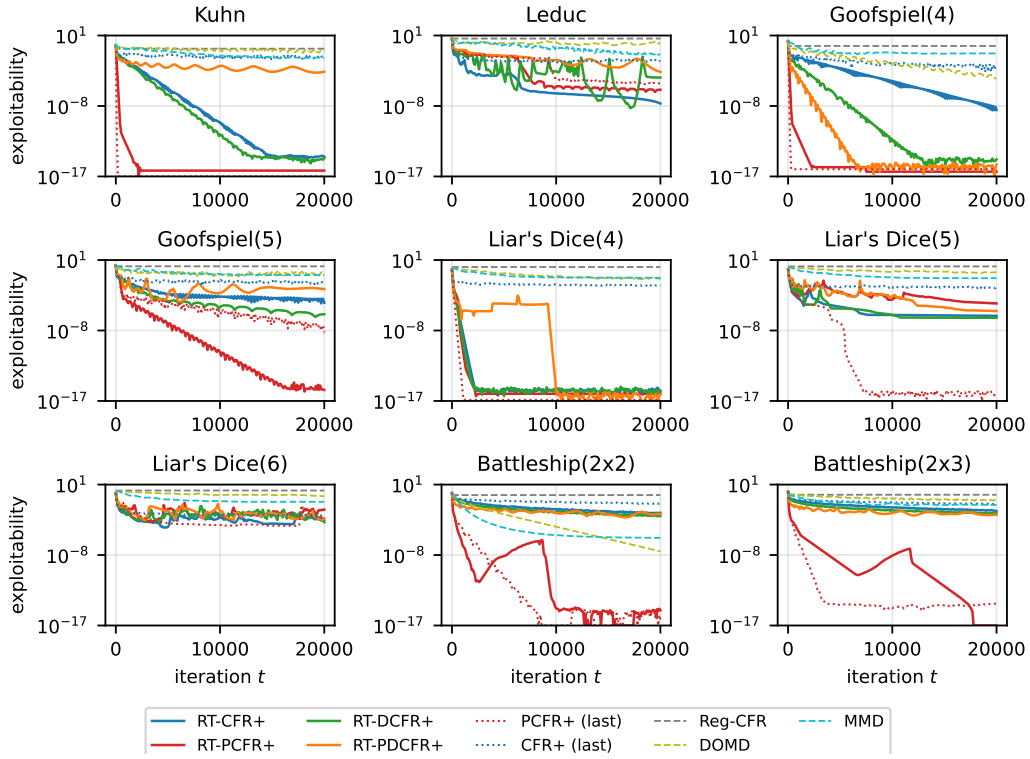

Figure 3: Last-iterate exploitability in the original game. Solid: PDRM$^+$ instances with the RTCFR$^+$ outer loop and hyperparameters; dotted: CFR$^+$ and PCFR$^+$ without regularization; dashed: other methods with last-iterate guarantees.

Table 2: Last-iterate exploitability after 20,000 iterations on the game of Cai et al. (2025a), and the fraction of random games on which it drops below 10^{-6}.

instance	updates	CFR$^+$	PCFR$^+$	PDCFR$^+$	RT-CFR$^+$	RT-PCFR$^+$	RT-PDCFR$^+$
Cai et al. game	alt.	$1.2{\cdot}10^{-1}$	$<10^{-15}$	$1.0{\cdot}10^{-3}$	$5.9{\cdot}10^{-12}$	$<10^{-15}$	$1.1{\cdot}10^{-3}$
Cai et al. game	sim.	$2.4{\cdot}10^{-1}$	$7.7{\cdot}10^{-1}$	$3.7{\cdot}10^{-1}$	$3.0{\cdot}10^{-1}$	$<10^{-15}$	$2.6{\cdot}10^{-2}$
random 3×3 (n=100), frac. $\leq 10^{-6}$	alt.	0.32	1.00	0.44	1.00	1.00	0.43
random 3×3 (n=100), frac. $\leq 10^{-6}$	sim.	0.32	0.64	0.65	0.34	0.99	0.89
random 4×4 (n=100), frac. $\leq 10^{-6}$	alt.	0.13	1.00	0.40	1.00	1.00	0.36
random 4×4 (n=100), frac. $\leq 10^{-6}$	sim.	0.13	0.69	0.60	0.13	0.95	0.88

converges on every instance we tried, which is why the benchmark results above understate the role of the guarantee.

7 Conclusion

The last iterate of predictive and discounted regret matching$^+$ converges in every perturbed regularized EFG, for any step size and any initialization, whenever the total discount is finite; the DCFR schedule satisfies the condition exactly when $\alpha > 1$, which includes $\alpha = 1.5$ and the $\alpha = 2.3$ of PDCFR$^+$, and for $p = 0$ a constant discount cannot converge on nondegenerate matrix games. The proof isolates two mechanisms: positive homogeneity, which makes step size and initialization interchangeable, and a comparator that shrinks with the iterate, which makes discounting harmless as long as the accumulated shrinkage stays bounded. Every instance has constant regret in the perturbed regularized game; empirically, optimism is what makes the last iterate fast. Our analysis is asymptotic and concerns one fixed perturbed regularized game under simultaneous updates. A last-iterate rate and a guarantee for the RT outer loop with a finite inner budget are the natural next targets.

Reproducibility statement

The anonymized code package released with this submission contains the driver for every experiment, the stored result files behind every reported number, and the numerical checks of Appendix B; its README maps each driver to the artifact it produces. The games, the protocol and the hyperparameters are documented in Appendix C. The core module implementing the PDRM$^+$ family ships as a compiled extension, so the reported results reproduce while the implementation stays unreadable, and the package documents the environment it was verified on.

AI use statement

This submission was prepared with the help of generative AI tools: they assisted with drafting, polishing and proofreading the text, and with implementing, running and analysing the experiments. All AI-assisted output was verified by the authors, who alone are responsible for the paper.

References

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

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

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. Adaptively perturbed mirror descent for learning in games. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 31–80, 2024.

Martino Bernasconi, Alberto Marchesi, and Francesco Trovò. Learning extensive-form perfect equilibria in two-player zero-sum sequential games. In *International Conference on Artificial Intelligence and Statistics*, pp. 2152–2160. PMLR, 2024.

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

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

Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1829–1836, 2019a.

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

Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng. Fast last-iterate convergence of learning in games requires forgetful algorithms. *Advances in Neural Information Processing Systems*, 37:23406–23434, 2024.

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

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. From average-iterate to last-iterate convergence in games: A reduction and its applications. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=mG9xItYI0D.

C Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *10th Innovations in Theoretical Computer Science (ITCS) conference, ITCS 2019*, 2019.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5363–5371, 2021.

Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, and Haipeng Luo. Regret Matching+: (In)Stability and Fast Convergence in Games. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=nYgs0qZJ97.

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 *Forty-second International Conference on Machine Learning*, 2025.

Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. *Advances in neural information processing systems*, 35:21858–21870, 2022.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

Chung-Wei Lee, Christian Kroer, and Haipeng Luo. Last-iterate convergence in extensive-form games. *Advances in Neural Information Processing Systems*, 34:14293–14305, 2021.

Mingyang Liu, Asuman E. Ozdaglar, Tiancheng Yu, and Kaiqing Zhang. The power of regularization in solving extensive-form games. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=bPiHuNUNv_R.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. Liteefg: An efficient python library for solving extensive-form games. *arXiv preprint arXiv:2407.20351*, 2024a.

Zongkai Liu, Chaohao Hu, Chao Yu, and Peng Sun. Regularization is enough for last-iterate convergence in zero-sum games, 2024b. URL https://openreview.net/forum?id=qjFnENGhDE.

Linjian Meng, Tianpei Yang, Youzhi Zhang, Zhenxing Ge, Shangdong Yang, Tianyu Ding, Wenbin Li, Bo An, and Yang Gao. Efficient last-iterate convergence in solving extensive-form games. In *Advances in Neural Information Processing Systems*, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/918c67cd7910578f1637beb56717f455-Abstract-Conference.html.

Linjian Meng, Youzhi Zhang, Zhenxing Ge, Tianyu Ding, Shangdong Yang, Zheng Xu, Wenbin Li, and Yang Gao. Last-iterate convergence of smooth regret matching$^+$ variants in learning nash equilibria. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=v771lscnlS.

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

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8525–8535. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/perolat21a.html.

Julien Pérolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling mean field games by online mirror descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 1028–1037, 2022.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26, 2013.

Hang Ren, Xiaozhen Sun, Tianzi Ma, Jiajia Zhang, and Xuan Wang. Last-iterate convergence in adaptive regret minimization for approximate extensive-form perfect equilibrium. In *28th European Conference on Artificial Intelligence, ECAI 2025, including 14th Conference on Prestigious Applications of Intelligent Systems, PAIS 2025*, pp. 4686–4693. IOS Press BV, 2025a.

Hang Ren, Yulin Wu, Shuhan Qi, Jiajia Zhang, Xiaozhen Sun, Tianzi Ma, and Xuan Wang. Efficient last-iterate convergence in regret minimization via adaptive reward transformation. *arXiv preprint arXiv:2509.13653*, 2025b.

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 *The Eleventh International Conference on Learning Representations*, 2023.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems*, 28, 2015.

Oskari Tammelin. Solving large imperfect information games using cfr+. *arXiv preprint arXiv:1407.5042*, 2014.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=dx11_7vm5_r.

Hang Xu, Kai Li, Haobo Fu, Qiang Fu, and Junliang Xing. Autocfr: Learning to design counterfactual regret minimization algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5244–5251, 2022.

Hang Xu, Kai Li, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. Dynamic discounted counterfactual regret minimization. In *The Twelfth International Conference on Learning Representations*, 2024a.

Hang Xu, Kai Li, Bingyun Liu, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. Minimizing weighted counterfactual regret with optimistic online mirror descent. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5272–5280, 2024b.

Hang Xu, Kai Li, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. Deep (predictive) discounted counterfactual regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17284–17292, 2026.

Brian Hu Zhang, Ioannis Anagnostides, and Tuomas Sandholm. Scale-invariant regret matching and online learning with optimal convergence: Bridging theory and practice in zero-sum games. *arXiv preprint arXiv:2510.04407*, 2025.

Naifeng Zhang, Stephen Marcus McAleer, and Tuomas Sandholm. Faster game solving via hyperparameter schedules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17319–17326, 2026.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20, 2007.

A PROOFS

Throughout, $D_\psi(u,v) = \frac{1}{2}\|u-v\|_2^2$. We use the following facts from Meng et al. (2025a), valid for the perturbed regularized game (1) with $\gamma, \mu > 0$.

(L-A) (decomposition; their Lemma 4.3) For $x, x' \in \mathcal{X}$ with behavioral σ, σ' and any player i, $\sum_{I\in\mathcal{I}_i} \pi_i^{\sigma'}(I)\langle -v_i^x(I), \sigma_i(I) - \sigma_i'(I)\rangle = \langle \ell_i^x, x_i - x_i'\rangle$; the identity holds verbatim for the losses and counterfactual values of (1).

(L-B) (strong monotonicity; their Lemma D.1) $\sum_i \langle \hat{\ell}_i^{\hat{x}}, \hat{x}_i - \hat{x}_i^*\rangle \geq \mu\|\hat{x} - \hat{x}^*\|_2^2$ for all $\hat{x} \in \mathcal{X}^\gamma$.

(L-C) (the NE is a local best response; their Lemma 4.4) $\hat{m}_i^*(I) := \hat{v}_i^{\hat{x}^*}(I) - \langle \hat{v}_i^{\hat{x}^*}(I), \sigma_i^*(I)\rangle \mathbf{1} \leq 0$ componentwise; equivalently $\langle -\hat{m}_i^*(I), \theta\rangle \geq 0$ for all $\theta \geq 0$.

(L-D) (smoothness of instantaneous regrets; their Eq. (14) with the perturbation map made explicit) For all $\hat{x} \in \mathcal{X}^\gamma$ with behavioral σ and $\hat{m}_i(I) = \hat{v}_i^{\hat{x}}(I) - \langle \hat{v}_i^{\hat{x}}(I), \sigma_i(I)\rangle \mathbf{1}$: $\|\hat{m}_i(I) - \hat{m}_i^*(I)\|_2^2 \leq C_1\|\hat{x} - \hat{x}^*\|_1^2$ with $C_1 = A_{\max}^2\big(12(L+\mu)^2 + 28(P+2\mu D)^2(A_{\max}C_{\max} + 1)^2\gamma^{-2H}\big)$. The constant follows from their Eq. (14), $\|\hat{m} - \hat{m}^*\|_2^2 \leq 6A_{\max}^2\|\hat{v} - \hat{v}^*\|_2^2 + 4A_{\max}^2(P+2\mu D)^2\|\sigma_i(I) - \sigma_i^*(I)\|_2^2$, their Lemma D.3 for the first term, and $\|\sigma_i(I) - \sigma_i^*(I)\|_2 \leq \|\sigma_i(I) - \sigma_i^*(I)\|_1 = \|\hat{\sigma}_i(I) - \hat{\sigma}_i^*(I)\|_1/(1-\alpha_I) \leq 2\|\hat{\sigma}_i(I) - \hat{\sigma}_i^*(I)\|_1$ for the second (the factor $(1-\alpha_I)^{-1}$ is where $\gamma \leq 1/(2A_{\max})$ enters; it is absent from their Eq. (15)), followed by their Lemma D.4. Summing over infosets and using $\|\cdot\|_1^2 \leq n\|\cdot\|_2^2$: $\sum_{i,I}\|\hat{m}_i(I) - \hat{m}_i^*(I)\|_2^2 \leq C\|\hat{x} - \hat{x}^*\|_2^2$, $C := |\mathcal{I}|\, n\, C_1$.

(L-E) (their Lemma D.4) $\|\hat{\sigma}_i - \hat{\sigma}_i'\|_1 \leq \frac{A_{\max}C_{\max}+1}{\gamma^H}\|\hat{x}_i - \hat{x}_i'\|_1$ on $\mathcal{X}^\gamma$.

A.1 INSTANCES

DCFR$^+$ of Xu et al. (2024b) updates $R^t = \left[R^{t-1}\frac{(t-1)^\alpha}{(t-1)^\alpha+1} + r^t\right]_+$ and PDCFR$^+$ plays $\mathrm{nrm}\left(\left[R^t\frac{t^\alpha}{t^\alpha+1} + r^t\right]_+\right)$, i.e. the prediction is formed from the same discounted base point as the next update. This is (3)–(4) with $d_t = (t-1)^\alpha/((t-1)^\alpha+1)$, whose first element is 0 (the initialization is discarded at $t=1$); Assumption 1 and Theorem 1 apply verbatim from $t=2$ on. We use $d_t = t^\alpha/(t^\alpha+1)$ in the experiments so that the initialization is never discarded and the invariance statements are nonvacuous.

Table 3: Instances of PDRM$^+$. The discount $d_t = t^\alpha/(t^\alpha+1)$ is the DCFR schedule of Brown & Sandholm (2019a), used with $\alpha = 1.5$ by DCFR$^+$ and with the grid-searched $(\alpha, \gamma) = (2.3, 5)$ by PDCFR$^+$ (Xu et al., 2024b) (up to the index shift above); we use $\alpha = 1.5$ for every instance, varying it in Figure 2, and Assumption 1 holds for every $\alpha > 1$. Prediction uses the current regret as the forecast of the next one.

local minimizer	d_t	p	CFR algorithm (with the RT outer loop)
RM$^+$ (Tammelin, 2014)	1	0	CFR$^+$ / RTCFR$^+$ (Meng et al., 2025a)
PRM$^+$ (Farina et al., 2021)	1	1	PCFR$^+$ / RT-PCFR$^+$
discounted RM$^+$	$t^\alpha/(t^\alpha+1)$	0	DCFR$^+$ / RT-DCFR$^+$
predictive discounted RM$^+$	$t^\alpha/(t^\alpha+1)$	1	PDCFR$^+$ / RT-PDCFR$^+$

A.2 TWO ELEMENTARY LEMMAS

Lemma 2 (optimistic three-point inequality on the cone)**.** *Let $K = \mathbb{R}^k_{\geq 0}$, $b \in K$, $g, q \in \mathbb{R}^k$, $\eta > 0$, $\tilde{\theta} = [b + \eta q]_+$ and $\theta^+ = [b + \eta g]_+$. Then for every $u \in K$, $\eta\langle g, u - \tilde{\theta}\rangle \leq D_\psi(u,b) - D_\psi(u,\theta^+) - D_\psi(\theta^+,\tilde{\theta}) - D_\psi(\tilde{\theta},b) + \eta\langle g - q, \theta^+ - \tilde{\theta}\rangle$. With $q = 0$ (so $\tilde{\theta} = b$) the plain three-point inequality $\eta\langle g, u - \theta^+\rangle \leq D_\psi(u,b) - D_\psi(u,\theta^+) - D_\psi(\theta^+,b)$ holds.*

Proof. $[\cdot]_+$ is the Euclidean projection onto K, so $\theta^+ = \arg\min_{\theta\in K}\{-\eta\langle g,\theta\rangle + D_\psi(\theta,b)\}$ and the three-point inequality of the proximal step gives $\eta\langle g, u - \theta^+\rangle \leq D_\psi(u,b) - D_\psi(u,\theta^+) - D_\psi(\theta^+,b)$

for every $u \in K$. Applying the same inequality to $\tilde{\theta}$ with comparator $\theta^+ \in K$ gives $\eta\langle q, \theta^+ - \tilde{\theta}\rangle \le D_\psi(\theta^+, b) - D_\psi(\theta^+, \tilde{\theta}) - D_\psi(\tilde{\theta}, b)$. Add the two and use $\eta\langle g, u - \tilde{\theta}\rangle = \eta\langle g, u - \theta^+\rangle + \eta\langle g - q, \theta^+ - \tilde{\theta}\rangle + \eta\langle q, \theta^+ - \tilde{\theta}\rangle$. □

Proof of Lemma 1. Induction on t. If $\theta'^t = c\theta^t$ and $\hat{m}'^{t-1} = \hat{m}^{t-1}$, positive homogeneity of $[\cdot]_+$ gives $\tilde{\theta}'^t = \left[c\, d_t\theta^t + c\eta p\hat{m}^{t-1}\right]_+ = c\tilde{\theta}^t$, hence $\sigma'^t = \mathrm{nrm}(c\tilde{\theta}^t) = \sigma^t$ (also when $\tilde{\theta}^t = 0$). Identical strategy profiles give identical counterfactual values, so $\hat{m}'^t = \hat{m}^t$, and $\theta'^{t+1} = [cd_t\theta^t + c\eta\hat{m}^t]_+ = c\theta^{t+1}$. □

A.3 Proof of Theorem 1

Step 1 ($\eta \le \eta_0$). Fix $I \in \mathcal{I}_i$ and put $u_t := \lambda_t\sigma_i^*(I) \in K$, so $u_{t+1} = d_t u_t$. Apply Lemma 2 with $b = d_t\theta_I^t$, $g = \hat{m}_i^t(I)$, $q = p\,\hat{m}_i^{t-1}(I)$, $u = u_{t+1}$: by (5) the left side equals $\eta\lambda_{t+1}\langle\hat{m}_i^t(I), \sigma_i^*(I)\rangle = \eta\lambda_{t+1}\langle -\hat{v}_i^t(I), \sigma_i^t(I) - \sigma_i^*(I)\rangle$; on the right, $D_\psi(d_t u_t, d_t\theta_I^t) = d_t^2 D_\psi(u_t, \theta_I^t) \le D_\psi(u_t, \theta_I^t)$, and Young's inequality $\eta\langle a, b\rangle \le \frac{\eta^2}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$ with $a = \hat{m}^t - p\hat{m}^{t-1}$, $b = \theta^{t+1} - \tilde{\theta}^t$ absorbs $-D_\psi(\theta^{t+1}, \tilde{\theta}^t) = -\frac{1}{2}\|b\|^2$; we drop $-D_\psi(\tilde{\theta}^t, d_t\theta^t) \le 0$. With $\Phi_t(I) := D_\psi(u_t, \theta_I^t) \ge 0$,

$$\eta\lambda_{t+1}\langle -\hat{v}_i^t(I), \sigma_i^t(I) - \sigma_i^*(I)\rangle \le \Phi_t(I) - \Phi_{t+1}(I) + \tfrac{\eta^2}{2}\|\hat{m}_i^t(I) - p\,\hat{m}_i^{t-1}(I)\|_2^2. \tag{8}$$

Case $p = 1$: $\|\hat{m}^t - \hat{m}^{t-1}\|^2 \le 2\|\hat{m}^t - \hat{m}^*\|^2 + 2\|\hat{m}^{t-1} - \hat{m}^*\|^2$. *Case* $p = 0$: the error in (8) is $\frac{\eta^2}{2}\|\hat{m}^t\|^2$ and does not vanish. Instead use the plain three-point inequality, $\eta\langle\hat{m}^t, u_{t+1} - \theta^{t+1}\rangle \le D_\psi(u_t, \theta^t) - D_\psi(u_{t+1}, \theta^{t+1}) - D_\psi(\theta^{t+1}, d_t\theta^t)$, and decompose, using $\langle\hat{m}^t, d_t\theta^t\rangle = 0$ from (5), $\eta\langle\hat{m}^t, \theta^{t+1}\rangle = \eta\langle\hat{m}^t - \hat{m}^*, \theta^{t+1} - d_t\theta^t\rangle + \eta\langle\hat{m}^*, \theta^{t+1}\rangle - \eta d_t\langle\hat{m}^*, \theta^t\rangle$. Young's inequality absorbs the first term into $D_\psi(\theta^{t+1}, d_t\theta^t)$ at the cost $\frac{\eta^2}{2}\|\hat{m}^t - \hat{m}^*\|^2$. With $\Psi_t(I) := D_\psi(u_t, \theta_I^t) + \eta\langle -\hat{m}_i^*(I), \theta_I^t\rangle \ge 0$ (L-C),

$$\eta\lambda_{t+1}\langle -\hat{v}_i^t(I), \sigma_i^t(I) - \sigma_i^*(I)\rangle \le \Psi_t(I) - \Psi_{t+1}(I) - \eta(1-d_t)\langle -\hat{m}_i^*(I), \theta_I^t\rangle + \tfrac{\eta^2}{2}\|\hat{m}_i^t(I) - \hat{m}_i^*(I)\|_2^2, \tag{9}$$

and the third term on the right is nonpositive by (L-C) and is dropped.

Aggregation. Multiply (8) (resp. (9)) by $\zeta_I := (1-\alpha_I)\pi_i^{\hat{\sigma}^*}(I) \in [0,1]$ and sum over $I \in \mathcal{I}_i$ and i. Since $(1-\alpha_I)(\sigma_i^t(I) - \sigma_i^*(I)) = \hat{\sigma}_i^t(I) - \hat{\sigma}_i^*(I)$, (L-A) gives $\sum_I \zeta_I\langle -\hat{v}_i^t(I), \sigma_i^t(I) - \sigma_i^*(I)\rangle = \langle\hat{\ell}_i^t, \hat{x}_i^t - \hat{x}_i^*\rangle$, and (L-B) gives $\sum_i\langle\hat{\ell}_i^t, \hat{x}_i^t - \hat{x}_i^*\rangle \ge \mu\|\hat{x}^t - \hat{x}^*\|_2^2$. Bounding $\zeta_I \le 1$ in the error terms and using (L-D), with $\Phi_t := \sum_{i,I}\zeta_I\Phi_t(I)$ and $\Psi_t := \sum_{i,I}\zeta_I\Psi_t(I)$,

$$\eta\lambda_{t+1}\mu\|\hat{x}^t - \hat{x}^*\|_2^2 \le \Phi_t - \Phi_{t+1} + \eta^2 C\big(\|\hat{x}^t - \hat{x}^*\|_2^2 + \|\hat{x}^{t-1} - \hat{x}^*\|_2^2\big) \quad (p = 1,\ t \ge 2), \tag{10}$$

$$\eta\lambda_{t+1}\mu\|\hat{x}^t - \hat{x}^*\|_2^2 \le \Psi_t - \Psi_{t+1} + \tfrac{\eta^2}{2}C\|\hat{x}^t - \hat{x}^*\|_2^2 \quad (p = 0). \tag{11}$$

For $p = 1$ and $t = 1$, (L-D) does not apply to the initial prediction and the inequality reads $\eta\lambda_2\mu\|\hat{x}^1 - \hat{x}^*\|^2 \le \Phi_1 - \Phi_2 + \eta^2 C\|\hat{x}^1 - \hat{x}^*\|^2 + \eta^2 E_0$ with $E_0 := \sum_{i,I}\|\hat{m}_i^0(I) - \hat{m}_i^*(I)\|_2^2 < \infty$. Summing over $t \le T$, using $\lambda_{t+1} \ge \lambda_\infty > 0$ and $\Phi_{T+1}, \Psi_{T+1} \ge 0$,

$$(\eta\lambda_\infty\mu - 2\eta^2 C)\sum_{t\le T}\|\hat{x}^t - \hat{x}^*\|_2^2 \le \Phi_1 + \eta^2 E_0 \quad (p = 1),$$

$$(\eta\lambda_\infty\mu - \tfrac{\eta^2}{2}C)\sum_{t\le T}\|\hat{x}^t - \hat{x}^*\|_2^2 \le \Psi_1 \quad (p = 0).$$

For $\eta \le \eta_0 := \lambda_\infty\mu/(4C)$ (resp. $\lambda_\infty\mu/C$) the bracket is at least $\eta\lambda_\infty\mu/2$, so $\sum_{t\ge1}\|\hat{x}^t - \hat{x}^*\|_2^2 \le 2(\Phi_1 + \eta^2 E_0)/(\eta\lambda_\infty\mu)$ (resp. $2\Psi_1/(\eta\lambda_\infty\mu)$).

Step 2 (every η and θ^1). Given (θ^1, η) let $c = \eta_0/\eta$. By Lemma 1 the run $(c\theta^1, \eta_0)$ has the same strategy sequence $\{\hat{x}^t\}$, and Step 1 applies to it because it holds for every initialization in the cone. Hence $\sum_t\|\hat{x}^t - \hat{x}^*\|^2 < \infty$ for the original run, $\hat{x}^t \to \hat{x}^*$, and (L-E) transfers the convergence to the perturbed behavioral strategies. □

A.4 Proof of Theorem 2

(i) Small step size. Let $\eta \le \eta_0$ and sum (10) over $t \le T$ (with the $t = 1$ term of the main proof): $\Phi_{T+1} \le \Phi_1 + \eta^2 E_0 - (\eta\lambda_\infty\mu - 2\eta^2 C)\sum_{t\le T}\|\hat{x}^t - \hat{x}^*\|_2^2 \le \Phi_1 + \eta^2 E_0$ for $p = 1$, and likewise $\Phi_{T+1} \le \Psi_{T+1} \le \Psi_1$ for $p = 0$ from (11). Write $\bar{\Phi}$ for the right-hand side. Since every summand of $\Phi_{T+1} = \sum_{i,I}\zeta_I D_\psi(u_{T+1}, \theta_I^{T+1})$ is nonnegative and $\zeta_I \ge (1-\alpha_I)\gamma^{H-1} > 0$ (each own probability on the path to I is at least γ under $\hat{\sigma}^*$),

$$\|\theta_I^{T+1}\|_\infty \le \|\theta_I^{T+1} - u_{T+1}\|_2 + \|u_{T+1}\|_2 \le \sqrt{2\bar{\Phi}/\zeta_I} + 1 =: B_I,$$

using $\|u_{T+1}\|_2 = \lambda_{T+1}\|\sigma_i^*(I)\|_2 \le 1$.

(i) General step size. Given (θ^1, η), Lemma 1 gives $\theta^t = (\eta/\eta_0)\theta'^t$ where θ' is the run $(\eta_0\theta^1/\eta, \eta_0)$, to which the previous paragraph applies with its own $\bar{\Phi}'$ computed from $\eta_0\theta^1/\eta$ and $\hat{m}^0$. Hence $B_I = (\eta/\eta_0)B_I'$ and $B_I/\eta = B_I'/\eta_0$ depends on θ^1/η only.

(ii) Componentwise, $\theta_a^{t+1} = [d_t\theta_a^t + \eta\hat{m}_a^t]_+ \ge d_t\theta_a^t + \eta\hat{m}_a^t$, i.e. $\eta\hat{m}_a^t \le \theta_a^{t+1} - \theta_a^t + (1-d_t)\theta_a^t$. Summing over $t \le T$ and dropping $-\theta_a^1 \le 0$, $\eta\sum_{t\le T}\hat{m}_a^t \le \theta_a^{T+1} + \sum_{t\le T}(1-d_t)\theta_a^t \le B_I(1+S)$.

(iii) Let $x_i' \in \mathcal{X}_i^\gamma$ have behavioral strategy $\hat{\sigma}_i'(I) = (1-\alpha_I)u'(I) + \gamma\mathbf{1}$, $u'(I) \in \Delta$. By (L-A) applied at every t to $(\hat{x}^t, x')$, $\hat{\sigma}^t - \hat{\sigma}' = (1-\alpha_I)(\sigma^t - u')$ and $\langle -\hat{v}^t, \sigma^t - u'\rangle = \langle \hat{m}^t, u'\rangle$ for $u' \in \Delta$,

$$\sum_{t\le T}\langle \hat{\ell}_i^t, \hat{x}_i^t - x_i'\rangle = \sum_I \pi_i^{\hat{\sigma}'}(I)(1-\alpha_I)\sum_{t\le T}\langle \hat{m}_i^t(I), u'(I)\rangle \le \sum_{I\in\mathcal{I}_i}\Big[\max_a \sum_{t\le T}\hat{m}_i^t(I,a)\Big]_+ \le (1+S)\sum_{I\in\mathcal{I}_i}\frac{B_I}{\eta},$$

using $0 \le \pi_i^{\hat{\sigma}'}(I)(1-\alpha_I) \le 1$ and (ii). The objective of (1) is convex in $\hat{x}_0$ and concave in $\hat{x}_1$ with gradients $\hat{\ell}_0, -\hat{\ell}_1$, so the saddle-point gap of $\bar{x}^T$ is at most $(R_0^T + R_1^T)/T$ (Meng et al., 2025a, Theorem F.1). □

Explicit constant. For $\eta \le \eta_0$ and $p = 1$, $\bar{\Phi} = \sum_{i,I}\zeta_I\frac{1}{2}\|\sigma_i^*(I) - \theta_I^1\|_2^2 + \eta^2 E_0$; for $p = 0$, $\bar{\Phi} = \sum_{i,I}\zeta_I\big[\frac{1}{2}\|\sigma_i^*(I) - \theta_I^1\|_2^2 + \eta\langle -\hat{m}_i^*(I), \theta_I^1\rangle\big]$. For general η replace θ^1 by $\eta_0\theta^1/\eta$ and η by $\eta_0 = \lambda_\infty\mu/(4C)$. The bound grows with $\|\theta^1\|/\eta$, with E_0, with $1/\lambda_\infty$ (through η_0 and S), with $1/\mu$ and, through C and $\zeta_I \ge \gamma^{H-1}/2$, with γ^{-H}: *any* effective initialization converges and has constant regret, but the constant is not uniform over θ^1/η.

A.5 Proof of Corollary 1

For $x_1 \in \mathcal{X}_1$ with behavioral σ_1 let $\tilde{x}_1 \in \mathcal{X}_1^\gamma$ be the sequence form of $(1-\alpha_I)\sigma_1(I) + \gamma\mathbf{1}$. Against the fixed $\hat{x}_0^*$ (and chance) the leaf distributions induced by σ_1 and $\tilde{\sigma}_1$ have total variation at most the sum, over the at most H decision points of player 1 on a path, of $\frac{1}{2}\|\sigma_1(I) - \tilde{\sigma}_1(I)\|_1 = \frac{\alpha_I}{2}\|\sigma_1(I) - \mathbf{1}/|A(I)|\|_1 \le \alpha_I \le A_{\max}\gamma$; hence $|\hat{x}_0^{*\top}A\hat{x}_1 - \hat{x}_0^{*\top}A\tilde{x}_1| \le 2MHA_{\max}\gamma$. By optimality of $\hat{x}_1^*$ in (1) against $\hat{x}_0^*$, $\hat{x}_0^{*\top}A\tilde{x}_1 - \mu D_\psi(\tilde{x}_1, r_1) \le \hat{x}_0^{*\top}A\hat{x}_1^* - \mu D_\psi(\hat{x}_1^*, r_1) \le \hat{x}_0^{*\top}A\hat{x}_1^*$, so $\max_{x_1\in\mathcal{X}_1}\hat{x}_0^{*\top}Ax_1 \le \hat{x}_0^{*\top}A\hat{x}_1^* + \mu\max_x D_\psi(x, r_1) + 2MHA_{\max}\gamma$. The symmetric bound for the minimizing player contributes the same two terms; adding the two sides and using $D_\psi(x,r) \le \frac{1}{2}D^2$ gives $\mu D^2 + 4MHA_{\max}\gamma$. □

B Constant discounting: proof of Proposition 1

Setting. Two-player zero-sum game in normal form, one decision point per player with k_i actions, perturbed and regularized as in (1) with $\mu > 0$ and $0 \le \gamma \le 1/(2A_{\max})$: the played strategy is $\hat{\sigma}_i = (1-\alpha_i)\sigma_i + \gamma\mathbf{1}$, losses $\hat{\ell}_0 = A\hat{\sigma}_1 + \mu(\hat{\sigma}_0 - r_0)$, $\hat{\ell}_1 = -A^\top\hat{\sigma}_0 + \mu(\hat{\sigma}_1 - r_1)$, values $\hat{v}_i = -\hat{\ell}_i$, instantaneous regrets $\hat{m}_i(\sigma) = \hat{v}_i - \langle\hat{v}_i, \sigma_i\rangle\mathbf{1}$ (unperturbed σ_i, as in (4)), $F(\sigma) = (\hat{m}_0(\sigma), \hat{m}_1(\sigma))$ (a polynomial map of σ). Let σ^* be the preimage of the unique NE $\hat{\sigma}^*$; it is fully mixed by assumption; then (L-C) gives $\hat{m}_i^* \le 0$ and $\langle\hat{m}_i^*, \sigma_i^*\rangle = 0$ with $\sigma_i^* > 0$, hence $F(\sigma^*) = 0$. Let $T = \{(\delta_0, \delta_1) : \langle\delta_i, \mathbf{1}\rangle = 0\}$, $\Pi(z_0, z_1) = (z_0 - \langle z_0, \mathbf{1}\rangle\sigma_0^*,\ z_1 - \langle z_1, \mathbf{1}\rangle\sigma_1^*) \in T$, and $J = \nabla F(\sigma^*)$.

Nondegeneracy (N). (N1) $\Pi J : T \to T$ is invertible, with least singular value $s > 0$. (N2) $A\delta_1 \notin \mathrm{span}(\mathbf{1})$ for $0 \ne \delta_1 \in T_1$ and $A^\top\delta_0 \notin \mathrm{span}(\mathbf{1})$ for $0 \ne \delta_0 \in T_0$.

Proof of Proposition 1. (a) *Exact hits are absorbing.* If $\sigma^t = \sigma^*$ then $\hat{m}^t = 0$, so $\theta_i^{t+1} = [d\theta_i^t]_+ = d\theta_i^t$ for both players (also if $\theta_i^t = 0$) and $\sigma_i^{t+1} = \mathrm{nrm}(d\theta_i^t) = \mathrm{nrm}(\theta_i^t) = \sigma_i^t = \sigma_i^*$. Hence it suffices to derive a contradiction from

$$(\star) \qquad \sigma^t \to \sigma^* \quad \text{and} \quad \delta^t := \sigma^t - \sigma^* \neq 0 \text{ for all } t.$$

(b) $\theta^t \to 0$: $\|\theta^{t+1}\| \le d\|\theta^t\| + \eta\|\hat{m}^t\|$ and $\hat{m}^t = F(\sigma^t) \to F(\sigma^*) = 0$. Hence $\bar{\rho}^t := \max_i \langle \theta_i^t, \mathbf{1}\rangle \to 0$.

(c) *For each player, $\theta_i^t \neq 0$ for all large t.* Take t large enough that $\sigma_{i,a}^s > 0$ for all a and $s \ge t$ (possible under $(\star)$ since $\sigma_i^* > 0$), and suppose $\theta_i^t = 0$. Then σ_i^t is uniform and $\theta_i^{t+1} = [\eta\hat{m}_i^t]_+$. Because $\langle \hat{m}_i^t, \sigma_i^t\rangle = 0$ with σ_i^t uniform, $\sum_a \hat{m}_{i,a}^t = 0$; if $\hat{m}_i^t \neq 0$ some component is negative and $\sigma_i^{t+1} = \mathrm{nrm}([\eta\hat{m}_i^t]_+)$ has a zero component, a contradiction. So $\hat{m}_i^t = 0$, $\theta_i^{t+1} = 0$, and by induction $\theta_i^s = 0$, $\sigma_i^s =$ uniform $= \sigma_i^*$ and $\hat{m}_i^s = 0$ for all $s \ge t$. Now $\hat{m}_i(\sigma_i^*, \sigma_j)$ is affine in σ_j with value 0 at σ_j^* and derivative $\mp(1-\alpha_j)(I - \mathbf{1}\sigma_i^{*\top})A^{(i)}\delta_j$ in direction δ_j ($A^{(0)} = A$, $A^{(1)} = A^\top$), which vanishes iff $A^{(i)}\delta_j \in \mathrm{span}(\mathbf{1})$; by (N2), $\hat{m}_i^s = 0$ for $s \ge t$ forces $\delta_j^s = 0$, i.e. $\sigma^s = \sigma^*$, contradicting $(\star)$. Consequently $\rho_i^t := \langle \theta_i^t, \mathbf{1}\rangle > 0$ and $\sigma_i^t = \theta_i^t/\rho_i^t \to \sigma_i^* > 0$ give $\theta_i^t > 0$ componentwise for $t \ge t_2$. Then $\theta^{t+1} = [d\theta^t + \eta\hat{m}^t]_+$ with $\theta^{t+1} > 0$ componentwise forces $\theta^{t+1} = d\theta^t + \eta\hat{m}^t$: the updates are linear for $t \ge t_2$.

(d) Taking $\langle \cdot, \mathbf{1}\rangle$, $\rho_i^{t+1} = d\rho_i^t + \eta\langle \hat{m}_i^t, \mathbf{1}\rangle$; subtracting σ_i^* times this from $\rho_i^{t+1}\sigma_i^{t+1} = d\rho_i^t\sigma_i^t + \eta\hat{m}_i^t$ yields

$$X^{t+1} := \left(\rho_i^{t+1}\delta_i^{t+1}\right)_i = d\,X^t + \eta\,\Pi\hat{m}^t.$$

By differentiability $\hat{m}^t = J\delta^t + \mathrm{rem}_t$ with $\|\mathrm{rem}_t\| \le c_t\|\delta^t\|$, $c_t \to 0$, and $\delta^t \in T$. Since $\|X^{t+1}\| \le \bar{\rho}^{t+1}\|\delta^{t+1}\|$ and $\|\Pi J\delta^t\| \ge s\|\delta^t\|$ by (N1),

$$\bar{\rho}^{t+1}\|\delta^{t+1}\| \;\ge\; \|X^{t+1}\| \;\ge\; \eta s\|\delta^t\| - d\bar{\rho}^t\|\delta^t\| - \eta\|\Pi\|c_t\|\delta^t\| \;\ge\; \tfrac{\eta s}{2}\|\delta^t\|$$

for $t \ge t_3$, because $\bar{\rho}^t \to 0$ and $c_t \to 0$. As $\bar{\rho}^{t+1} \to 0$ there is t_4 with $\eta s/(2\bar{\rho}^{t+1}) \ge 2$ for $t \ge t_4$, so $\|\delta^{t+1}\| \ge 2\|\delta^t\|$ for all large t and $\|\delta^t\| \to \infty$ (recall $\delta^t \neq 0$), contradicting $\delta^t \to 0$. □

Example. Regularized matching pennies, $A = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, r uniform, $0 \le \gamma < \frac{1}{2}$: $\hat{\sigma}^*$ and σ^* are uniform, and in the tangent coordinates $e = (1,-1)/\sqrt{2}$ of each player $\Pi J = (1-2\gamma)\left(\begin{smallmatrix} -\mu & -2 \\ 2 & -\mu \end{smallmatrix}\right)$ (up to the sign convention of the off-diagonal block), with $|\det| = (1-2\gamma)^2(\mu^2+4) > 0$, so (N1) holds for every $\mu \ge 0$ and $\gamma < \frac{1}{2}$; (N2) holds because $Ae = 2e \notin \mathrm{span}(\mathbf{1})$. We verified ΠJ by finite differences for $\mu \in \{0, 0.1, 1\}$ and $\gamma \in \{0, 0.1, 0.3, 0.45\}$.

The predictive case. For $p = 1$ an exact hit is *not* absorbing: with $d = \frac{1}{2}$, $\eta = 1$, $\theta^1 = (1,2)$ and $\hat{m}^0 = (\frac{1}{2}, 0)$ against a uniform opponent, player 0 plays the uniform equilibrium at $t = 1$, so $\hat{m}^1 = 0$, but then plays $\mathrm{nrm}(d^2\theta^1) = (\frac{1}{3}, \frac{2}{3})$ at $t = 2$; the prediction carries memory that the strategy does not show. The linearization of the played cone point $\tilde{\theta}^t$ gives $X^{t+1} = dX^t + \eta\Pi((1+d)J\delta^t - dJ\delta^{t-1}) + o(\cdot)$ with X^t built from $\tilde{\theta}^t$; the frozen-coefficient companion operator has one family of eigenvalues of size $\Theta(\eta/\bar{\rho}) \to \infty$ and one of size $\to d/(1+d) < 1$, so a purely spectral argument does not exclude convergence along the contracting direction. We state non-convergence of PDRM$^+$ with constant discount for $p = 1$ as an empirical finding only.

C Experimental details and additional results

Implementation. All PDRM$^+$ instances share one LiteEFG computation graph: the counterfactual values of (1) are obtained from the original utilities by adding the sequence-form regularization gradient $-\mu(\hat{x}_i - r_i)$ at every sequence, exactly as in the released RTCFR$^+$ code, whose exploitability trajectory on Kuhn our RM$^+$ instance reproduces with maximal deviation $1.4 \cdot 10^{-15}$ over 2,000 iterations. Discounting uses $d_t = t^\alpha/(t^\alpha + 1)$ applied to θ^t before the update and, in the predictive case, d_{t+1} applied to θ^{t+1} inside the prediction, which is the PDCFR$^+$ form of Xu et al. (2024b) (Appendix A.1). Alternating updates recompute player 2's counterfactual values after player 1 has updated; simultaneous updates compute both from the old profile. Exploitability is LiteEFG's exact best-response computation in the original game; the sequence-form distance in the convergence experiment is computed from the perturbed behavioral strategies through the infoset tree exported from OpenSpiel.

Games. Kuhn poker (12 infosets), Leduc poker (936), Goofspiel with 4 and 5 cards, imperfect information, descending point order (162 and 2,124 infosets), Liar's Dice with one 4/5/6-sided die per player (1,024, 5,120 and 24,576 infosets), and Battleship on a 2×2 and a 2×3 board with a single ship of length 2, three shots and no repeated shots (3,286 and 81,027 infosets). Simultaneous-move games are converted to turn-based form by OpenSpiel.

Certifying the reference point. For any $\hat{x} \in \mathcal{X}^\gamma$ let $\varepsilon_i(\hat{x}) := \sum_{I \in \mathcal{I}_i} \max_a [\hat{m}_i(I,a)]_+$ with $\hat{m}_i(I)$ the instantaneous regrets at $\hat{x}$. By (L-A) with comparator $\hat{x}^*$ and $\langle -\hat{v}_i(I), \sigma_i(I) - \sigma_i^*(I) \rangle = \langle \hat{m}_i(I), \sigma_i^*(I) \rangle \le \max_a [\hat{m}_i(I,a)]_+$, $\langle \hat{\ell}_i^{\hat{x}}, \hat{x}_i - \hat{x}_i^* \rangle \le \varepsilon_i(\hat{x})$, and (L-B) gives

$$\|\hat{x} - \hat{x}^*\|_2 \le \sqrt{(\varepsilon_0(\hat{x}) + \varepsilon_1(\hat{x}))/\mu}.$$

The right-hand side is computable from the reference run alone. For every reference point we report it next to the distances.

Reference points and invariance. Reference points $\hat{x}^*$ are RM$^+$ (alternating) runs of three times the length of the compared runs (150,000 iterations); the final points of all converging runs agree with each other to the precision that their own convergence permits (the plateau values 1.476 and 7.283 quoted in Section 6, and the pairwise sequence-form distances behind them, are stored with the released runs). Invariance check: on Kuhn with 20,000 iterations, the exploitability trajectories of $(\theta^1, \eta) = (0, 1)$ and $(0, 100)$, and of $(\mathbf{1}, 1)$ and $(100\mathbf{1}, 100)$, differ by at most $2 \cdot 10^{-16}$ and $2 \cdot 10^{-15}$ (PRM$^+$ and PDRM$^+$), while $(\mathbf{1}, 1)$ versus $(\mathbf{1}, 100)$ differ by 10^{-2} and $(\mathbf{1}, 1)$ versus $(\mathbf{1}, 0.01)$ by 0.66. On Leduc the same comparison gives $3 \cdot 10^{-15}$ (PRM$^+$) and $1.1 \cdot 10^{-3}$ (PDRM$^+$): two runs at the same ratio separate by roundoff that the longer trajectory amplifies. Lemma 1 is an identity in exact arithmetic; at this precision its numerical realisation is exact on Kuhn and approximate on Leduc.

Regularized matching pennies. For regularized matching pennies ($\mu = 0.1$, r uniform, $\eta = 1$, random $\theta^1 \in [0.5, 1.5]^2$, 20,000 iterations) the final distance to the uniform equilibrium is $3 \cdot 10^{-11}$ (RM$^+$) and $2 \cdot 10^{-16}$ (PRM$^+$) for $d = 1$, and between 0.44 and 1.0 for every $d \in \{0.999, 0.99, 0.9, 0.5\}$ and both p (worst distance over the second half of the run); the effective step size $\eta / \|\theta^t\|_1$ at the end of the run grows from 0.03 ($d = 1$, RM$^+$) to 0.9 ($d = 0.5$). The finite-difference Jacobian in tangent coordinates is $\left(\begin{smallmatrix} -\mu & -2 \\ 2 & -\mu \end{smallmatrix}\right)$ for $\mu \in \{0, 0.1, 1\}$ (determinants 4, 4.01, 5), confirming the nondegeneracy conditions of Proposition 1.

Protocol and baseline hyperparameters. Protocol sweep: for RT-CFR$^+$ and RT-PCFR$^+$ on Kuhn, Leduc and Goofspiel(4), and on Liar's Dice(4) we ran $\gamma_0 \in \{10^{-10}, 10^{-3}, 10^{-2}\}$, $\mu \in \{10^{-3}, 10^{-2}\}$, $T_u \in \{100, 500\}$ for 20,000 iterations. Larger μ leaves a visible bias floor between reference updates on Leduc (10^{-4} to 10^{-2}), and larger T_u slows the outer loop on Leduc as well; we keep the published RTCFR$^+$ values $(10^{-10}, 10^{-3}, 100)$ for every RT variant, so that the comparison to Meng et al. (2025a) is direct. Baseline hyperparameters are the LiteEFG defaults (Reg-CFR: entropy regularizer, $\kappa = 1$, $\tau = 0.1$; DOMD: $\eta = 0.1$, entropy; MMD: $\eta = 0.005$, $\tau = 0.05$). Average iterates of CFR$^+$ (linear weights) and of PCFR$^+$, DCFR$^+$, PDCFR$^+$ (quadratic weights) are logged as well and reach between 10^{-12} and 10^{-4} after 20,000 iterations on the six smaller games.

Random normal-form games. Random games have i.i.d. uniform payoffs in $[-1, 1]$; 100 games per size, seed 2026. The RT protocol is $\mu = 10^{-2}$, $T_u = 100$, $r \leftarrow \sigma^t$ (γ plays no role in normal-form games). On the game of Cai et al. (2025a) with simultaneous updates, RT-PCFR$^+$ reaches $2 \cdot 10^{-16}$ for $\mu \in \{10^{-2}, 10^{-1}\}$ and 0.12 for $\mu = 10^{-3}$, while RT-CFR$^+$ ends at 0.30, 0.75 and 0.73 respectively.

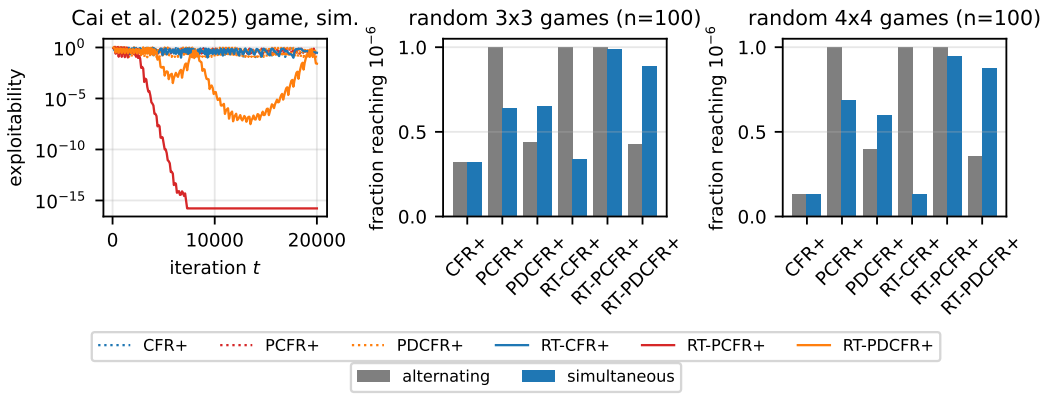

Figure 4: Left, the game of Cai et al. (2025a) under simultaneous updates; middle and right, the fraction of random 3×3 and 4×4 games solved to 10^{-6} by the last iterate. Same runs as Table 2.