IMPROVING LAST-ITERATE CONVERGENCE IN MONTE CARLO COUNTERFACTUAL REGRET MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual regret minimization (CFR) and its Monte Carlo sampled variants (MCCFR) are the standard solvers for two-player zero-sum extensive-form games and the engine behind superhuman poker programs. Their guarantees cover the *average* strategy, which is cheap to track offline. A deployment that acts with the current strategy, for instance online adaptation to a changing opponent, has only the *last* iterate to act on. Under MCCFR sampling the last iterate of the regret-matching family settles at a noise floor and improves far more slowly than the average. We introduce reward-transformed MCCFR (RT-MCCFR$^+$), a regret matching plus (RM$^+$) algorithm whose last iterate lowers this floor and keeps falling from 5M to 40M node visits in our sweeps. A perturbation keeps every reach probability bounded below, an anchor regularization makes each local subproblem strongly concave, and a windowed average of the recent iterates is the output that the regularization pulls toward, at constant memory per information set. On five games under external sampling (ES) and outcome sampling (OS) with 10 seeds, the windowed anchor is below the best baseline under the same window rule in six of ten cells and above it in four, one cell resolved in each direction, and below every baseline last iterate; the raw last iterate is below every RM-family last iterate. Multiplying the budget by eight from an exploration transient lowers the anchor by a factor of 1.6 to 23 where RM$^+$ moves by at most 30%, and past that budget no anchor change is resolved. On the ES noise axis the floor rises with exponent 0.47 and bootstrap half-width 0.10 in the total sampling variance, consistent with the square-root heuristic of the self-normalized RM$^+$ step size, and it falls roughly inversely in the anchor strength over the working range. We prove an inverse-square-root last-iterate convergence bound, in squared distance to the regularized fixed point, in a stylized push regime.

1 INTRODUCTION

Counterfactual regret minimization (CFR) (Zinkevich et al., 2007; Li & Huang, 2026) is the standard solver for two-player zero-sum extensive-form games (EFGs). It runs a regret minimizer at every information set on counterfactual values, and the *average* of the played strategies converges to a Nash equilibrium at the rate of the accumulated regret. Its sampled variants, Monte Carlo CFR (MCCFR) (Lanctot et al., 2009), replace the full tree traversal by a sampled one, under external sampling (ES) of the opponent and chance or outcome sampling (OS) of a single trajectory. Sampling is what makes CFR applicable to games too large to traverse. DeepStack (Moravčík et al., 2017) and Libratus (Brown & Sandholm, 2018) solved heads-up no-limit Texas hold'em, Pluribus (Brown & Sandholm, 2019b) reached superhuman play in multiplayer poker, and Deep CFR (Brown et al., 2019) and ReBeL (Brown et al., 2020) scaled regret minimization and search to large imperfect-information games with function approximation. The fastest tabular solvers today, predictive regret matching (PRM$^+$) (Farina et al., 2021) and the scale-invariant IREG-PRM$^+$ (Zhang et al., 2025), accelerate the same regret dynamics. All of these run offline against the full game or a full training budget and report an average, distilled, or re-searched policy. The average is cheap to track alongside the current iterate, and the regret bounds of CFR-style solvers are stated for it.

A different setting asks the solver to keep playing: online adaptation to a changing opponent, or path following along a game whose payoffs drift. The average can still be accumulated online, but it spans the whole history, including iterates from a regime that no longer applies. The acting policy then has

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.

to come from the recent trajectory. Under MCCFR sampling the last iterates of the regret-matching (RM) family, regret matching plus (RM^+) and its predictive variants, settle at a noise floor. The sampled estimates keep pushing the iterate around, so the NashConv of the iterate, its distance from equilibrium, stops decreasing or decreases only slowly; on Kuhn and Leduc we measure within-run slopes near zero for the RM family (Table 3). The theory of last-iterate play covers normal-form games under full feedback (Meng et al., 2025b; Zhang et al., 2025) and, in extensive-form games, the full-feedback $RTCFR^+$ of Meng et al. (2025a). Under MCCFR sampling, where the RM family is what practitioners run, the last iterate has no comparable guarantee, and the floor is what limits it.

We introduce reward-transformed MCCFR (RT-$MCCFR^+$), an RM^+ algorithm whose last iterate lowers the noise floor under MCCFR sampling and falls as the budget grows in our sweeps. At each information set we mix the RM^+ output with a uniform distribution, with weight γ, which keeps every reach probability bounded below. We then regularize the observed counterfactual value toward a slowly moving *anchor* with strength μ, which makes the local subproblem strongly concave, the property behind our last-iterate bound. The anchor is a *windowed average* of the recent played strategies, a single-policy output whose memory is constant in the number of iterations and whose variance is far below the raw iterate's.

The theory is a stylized statement for a single information set with two actions, and it explains the noise floor. Sampling makes the accumulated regret of the base RM^+ minimizer grow like the square root of the iteration count, so its self-normalized step size shrinks at the same rate and the update behaves as a stochastic approximation whose noise variance falls with it. The anchor contraction balances that noise against the drift. Weighting each player's potential by its own regret norm gives an inverse-square-root last-iterate bound, in squared distance to the regularized fixed point, with no condition on the ratio of the two players' step sizes (Theorem 4.1); the coefficient in front depends on the per-step noise variance and the anchor strength and stays a heuristic.

The experiments follow the same two quantities. On five games under ES and OS with 10 seeds, with the same window rule given to every baseline, the anchor is lower in six of ten cells and higher in four, one difference resolved in each direction (Table 1), and below every baseline last iterate; the raw last iterate is below every RM-family last iterate. A touch here is one visited tree node. From 5M to 40M touches the RM^+ last iterate moves by at most 30% and the anchor falls by a factor of 1.6 to 23; past 40M no anchor change is resolved at 10 seeds (Table 2, §B.9). Injecting noise into the counterfactual values on Kuhn moves the measured floor with exponent 0.47 and half-width 0.10 in the total variance, and lowering the anchor strength from 1 to 0.3 raises it with exponent -0.87 and half-width 0.11, consistent with the square-root heuristic and roughly inverse in the anchor strength (Figure 2).

Our contributions are:

1. We introduce RT-$MCCFR^+$, an RM-based algorithm whose windowed anchor is built for the last-iterate regime under MCCFR sampling. The anchor is compared against every baseline last iterate and against a matched-window baseline on five games under ES and OS sampling with 10 seeds (Tables 1 and 2).
2. We measure the noise floor: it rises with the total sampling variance along the square-root exponent of the self-normalized RM^+ step size and falls with the anchor strength down to the freeze threshold of the anchor (Figure 2).
3. We prove an inverse-square-root last-iterate convergence guarantee in a stylized push regime, stated for a single information set (Theorem 4.1). The proof weights each player's potential by its own regret norm, which removes any condition on the ratio of the two players' step sizes.

2 Related Work

Regret matching and its predictive variants. CFR (Zinkevich et al., 2007; Li & Huang, 2026) and RM^+ (Tammelin, 2014) decompose overall regret into per-information-set regret via counterfactual values and converge by outputting the average strategy. Discounting the accumulated regrets and strategy weights accelerates convergence in practice (Brown & Sandholm, 2019a), and smoothing-based first-order methods give faster rates for the average iterate in sequential games (Hoda et al., 2010; Kroer et al., 2015). Predictive variants (Farina et al., 2021; Zhang et al., 2025) accelerate regret matching with Blackwell-predictive updates, regret circuits (Farina et al., 2019b) show how

such minimizers compose, and behaviorally constrained minimizers (Farina et al., 2017) restrict the played strategy. All of these are evaluated on the average policy; under MCCFR sampling the last iterates of the RM family settle at the noise floor we measure in §5. RT-MCCFR$^+$ can wrap the same base minimizers, so our contribution is orthogonal to these speed-ups.

Last-iterate convergence. The theory of last-iterate play is richest in normal-form games under full feedback: smoothed or scale-invariant variants of RM$^+$ (Cai et al., 2025a; Meng et al., 2025b; Zhang et al., 2025), optimistic mirror descent and optimistic gradient methods (Anagnostides et al., 2022; Daskalakis & Panageas, 2019; Gorbunov et al., 2022), and cautious optimism (Soleymani et al., 2025) all converge. In extensive-form games, dilated distance-generating functions (Farina et al., 2019c) and their stable-predictive variants (Farina et al., 2019a) bring optimistic and predictive updates to sequential play, and regularized variants attain last-iterate convergence under full feedback (Liu et al., 2023). All of them assume exact losses. Under MCCFR sampling the per-round counterfactual estimates are noisy, and the RM-family baselines in §5 stall or improve only slowly; guarantees for the average or best iterate do not transfer to the last one (Cai et al., 2026). Average-to-last reductions (Cai et al., 2025b) wrap an average-iterate algorithm into a last-iterate one. They require exact gradient feedback to telescope the running average, and under CFR self-play the wrapped algorithm still returns the average. RT-MCCFR$^+$ is the base minimizer itself. Closest to us is RTCFR$^+$ (Meng et al., 2025a), the first RM-based last-iterate algorithm for EFGs, which proves last-iterate convergence of CFR$^+$ on a sequence of perturbed regularized EFGs under full feedback. RT-MCCFR$^+$ differs by studying the RM-based last iterate under MCCFR *sampling*, where the noise floor appears, by making the windowed anchor, a constant-memory average of the recent iterates, both the regularizer's reference and the output, and by measuring the noise floor as an explicit function of sampling variance and anchor strength. PFTRL (Masaka et al., 2025) is also perturbed, but FTRL-based and needs a learning rate; its reference strategy is refreshed on a visit clock and is not the output, whereas our anchor is both the regularizer's reference and the reported policy. Its reversed-KL perturbation has zero conditional variance, which makes it exact under external sampling; the implementation's numerical guard does not fire in our runs.

Sampling and variance control. The importance weights of MCCFR (Lanctot et al., 2009) grow as reach probabilities shrink, which is what destabilizes last-iterate play under sampling. A long line of work reduces the variance of the sampled counterfactual estimates during training: efficient sampling schemes (Johanson et al., 2012; Burch et al., 2012; Li et al., 2026b) and control-variate baselines such as VR-MCCFR (Schmid et al., 2019). These schemes speed up the average iterate by improving the estimator of its gradient. AIVAT (Burch et al., 2018; Li et al., 2026a) attacks a different estimator: it is a control variate for *evaluating* a fixed profile, not for training one. The last iterate is a third object: its noise floor is set by the per-step sampling variance σ^2 and the contraction strength of the update. Perturbing the played strategy bounds the reach probabilities from below, as in eq. (1); this is the mechanism behind the uniform fallback of Masaka et al. (2025), and our theory makes the role of the bound explicit. The noise-floor measurement in §5 connects the views: lowering σ^2 by a variance-reduction scheme should lower the floor at about the square-root rate measured on the ES axis of §5.

3 PRELIMINARIES

Extensive-form games. A two-player zero-sum extensive-form game (EFG) with imperfect information is a tuple $(\mathcal{N}, H, Z, A, \sigma_c, u, \mathcal{I}, P)$: players $\mathcal{N} = \{1, 2\}$, histories H with terminal set Z, actions $A(h)$, chance probabilities σ_c, a payoff $u(z) \in [-U, U]$ at each terminal, information sets $\mathcal{I}$ partitioning the histories that the acting player cannot distinguish, and a player function P; both players have perfect recall. We write $u_i(z)$ for the payoff to player i, $P(I)$ for the player acting at I, and, because perfect recall gives every history of I the same action set, A_I for that set and $\Delta(A_I)$ for the simplex over it. A *strategy profile* $\sigma = (\sigma_1, \sigma_2)$ assigns $\sigma_I \in \Delta(A_I)$ to each information set. We write $\pi^\sigma(h)$ for the reach probability of h, $\pi^\sigma_{-i}(h)$ for the contribution of the opponent and chance, $\pi^\sigma_i(h)$ for the player's own contribution, and $\pi^\sigma(h, z)$ for the probability of reaching z from h. The *counterfactual value* of action a at I is $v_i(\sigma, I, a) = \sum_{h \in I} \sum_z \pi^\sigma_{-i}(h)\, \pi^\sigma(ha, z)\, u_i(z)$. The *best-response value* of a profile is $\mathrm{BR}_i(\sigma) = \max_{\sigma'_i} \sum_z u_i(z) \pi^{\sigma'_i, \sigma_{-i}}(z)$, and the NashConv $\mathrm{expl}(\sigma) = \mathrm{BR}_1(\sigma) + \mathrm{BR}_2(\sigma)$ measures the distance to Nash equilibrium. A function is μ-strongly concave on a convex set if adding $\frac{\mu}{2}\|y - x\|^2$ makes it concave there.

Regret matching. CFR (Zinkevich et al., 2007) runs a local regret minimizer at each information set. The minimizer maintains a cumulative regret vector $R_I = (R_I(a))_{a \in A_I}$ with $R_I(a) = \sum_t (u_t(a) - \sum_a \sigma_I^t(a) u_t(a))$, where $u_t(a)$ is the counterfactual value of action a at iteration t and σ_I^t is the strategy played at I then. It plays the regret-matching rule (Hart & Mas-Colell, 2000) $\sigma_I(a) = R_I^+(a) / \sum_{a'} R_I^+(a')$, where $R^+ = \max(R, 0)$ and a zero numerator leaves σ_I uniform. RM$^+$ (Tammelin, 2014) instead stores the vector clipped at zero after every update, $\theta_{I,t} = (R_I(a)^+)_{a \in A_I}$, plays $\sigma_I(a) = \theta_{I,t}(a) / S_{I,t}$ with $S_{I,t} = \|\theta_{I,t}\|_1$, and is parameter-free. The standard output is the *average strategy* $\bar{\sigma}_I = \sum_t w_t \sigma_I^t / \sum_t w_t$ with $w_t = \pi_i^{\sigma^t}(I)$, the player's own reach of I at iteration t. This average converges to a Nash equilibrium. *Last-iterate* convergence is the property that the current strategy σ_I^t itself converges. Under sampling this fails for regret matching because the sampling noise keeps perturbing the iterate; we measure this floor in §5.

Sampled regret matching. MCCFR (Lanctot et al., 2009) samples a subset of the tree and rescales the observed counterfactual values so that the expected gradient equals the full-feedback one. *External sampling* (ES) samples the opponent's and chance's actions and branches over every action of the acting player at each information set reached under the sampled actions; *outcome sampling* (OS) samples a single history and reweights it by importance weights. The rescaled estimate is the *sampled counterfactual value* $\hat{u}_I = (\hat{u}_I(a))_{a \in A_I}$. We call the minimizer that consumes it the *base* minimizer, and the strategy it outputs at the moment an information set is visited the *executed* strategy. In both samplers variance control reduces to a bound on the reach probabilities, and the reach probabilities are what our perturbation bounds in eq. (1).

Budget units. One *touch* is one tree node entered by a traversal, terminal and chance nodes included; the touch budget B of a run is the total over both players. The outer window T_u is a number of touches, so a run of budget B refreshes the anchors about B/T_u times. The theorem counts update rounds T at one information set, not touches; §B.11 states how the two relate.

Predictive baselines. PRM$^+$ (Farina et al., 2021) replaces the observed counterfactual value by a prediction of the next one, through the Blackwell-approachability view of regret matching, and is a state-of-the-art full-feedback solver. IREG-PRM$^+$ (Zhang et al., 2025) is its scale-invariant, parameter-free extension with $T^{-1/2}$ best-iterate convergence in normal-form games; we use it with the exponential-moving-average (EMA) prediction of the source, which smooths the observed values with a decaying weight. Both output the average strategy; under MCCFR sampling their last iterates settle at a noise floor in our experiments.

4 METHOD

RT-MCCFR$^+$ wraps an RM$^+$ regret minimizer at every information set with a *reward-transformed* update that mixes in a small uniform distribution and regularizes toward a slowly moving anchor. The anchor is a windowed average of the recent played strategies, which gives a constant-memory single-policy output with low variance. Algorithm 1 states the full loop.

Reward transformation. Following the reward-transformed regret matching of Meng et al. (2025a), RT-MCCFR$^+$ maintains a base RM$^+$ minimizer per information set. Writing $\hat{x}_I$ for the strategy executed at I and q_I for the output of the base minimizer, it executes the mixture

$$\hat{x}_I = (1 - \gamma)\, q_I + \frac{\gamma}{|A_I|}, \tag{1}$$

so that every action is played with probability at least $\gamma / |A_I|$. For $\gamma > 0$ this floor keeps every reach probability, hence each information set's visit probability, bounded below (§B.7). The base minimizer observes the transformed counterfactual value

$$\tilde{u}_I \;=\; \hat{u}_I \;-\; \mu\,(\hat{x}_I - r_I), \tag{2}$$

where $\hat{u}_I$ is the sampled counterfactual value vector at I and r_I is the anchor. The regularizer enters once per visit, at strength μ relative to the visit-conditional value estimate. Over a full-feedback sequence, (2) is the gradient of a game whose local subproblem at I is μ-strongly concave in $\hat{x}_I$ around r_I. This local concavity is the property Theorem 4.1 exploits.

Algorithm 1: RT-MCCFR$^+$ (windowed-anchor RT-RM$^+$)

Input: touch budget B, window T_u, perturbation γ_0, anchor strength μ, decay $d \in (0,1)$, floor $\gamma_{\min}$
Global: $\gamma \leftarrow \gamma_0$, touch counter $n \leftarrow 0$, next refresh $\tau \leftarrow T_u$; per I: $\theta_I \leftarrow 0$ (RM$^+$ regret vector), $\hat{x}_I \leftarrow r_I \leftarrow \mathrm{Unif}(A_I)$, $\mathrm{acc}_I \leftarrow 0$, $c_I \leftarrow 0$; fallback threshold EPS (§A.1).;
while $n < B$ **do**
  Sample one block (one ES or OS traversal per player); n counts every visited tree node. At each visited I of the traverser: ; `/* perturb: execute a mixture */`
  $S_I \leftarrow \|\theta_I\|_1$; $q_I \leftarrow \theta_I / S_I$ if $S_I >$ EPS, else $\mathrm{Unif}(A_I)$;
  $\hat{x}_I \leftarrow (1-\gamma)\, q_I + \gamma/|A_I|$;
  $\tilde{u}_I \leftarrow \hat{u}_I - \mu(\hat{x}_I - r_I)$; `/* reward transform: pull toward anchor */`
  $\theta_I \leftarrow [\theta_I + \tilde{u}_I - \langle q_I, \tilde{u}_I \rangle \mathbf{1}]^+$;;
  $\mathrm{acc}_I \mathrel{+}= \hat{x}_I$;;
  $c_I \mathrel{+}= 1$;
  if $n \geq \tau$; `/* global refresh, about` B/T_u `per run */`
  then
    foreach I **do**
      $r_I \leftarrow \mathrm{acc}_I / c_I$ if $c_I > 0$, else $r_I \leftarrow$ current policy of I; then $\mathrm{acc}_I \leftarrow 0$, $c_I \leftarrow 0$
    $\gamma \leftarrow \max(\gamma \cdot d,\ \gamma_{\min})$;;
    $\tau \leftarrow \tau + T_u$;
Output: anchor strategy r_I at each I

Windowed anchor. We need a single-policy output drawn from the recent trajectory. The current strategy $\hat{x}_I$ carries the noise floor we study in §5, so we maintain a *windowed anchor* instead. Over each outer window of T_u touches we accumulate $\mathrm{acc}_I \mathrel{+}= \hat{x}_I$ and $c_I \mathrel{+}= 1$, and at the window boundary we set $r_I \leftarrow \mathrm{acc}_I / c_I$. This running average, one vector and one counter per information set, keeps the anchor on the recent part of the trajectory, where the perturbation has decayed and the strategy is closest to the equilibrium. RT-MCCFR$^+$ reports the anchor r_I.

Outer schedule. Every T_u tree touches on the global counter, so a run of budget B refreshes the anchors about B/T_u times (§B.11): all anchors move to their window averages and the single perturbation is halved, $\gamma \leftarrow \max(\gamma d, \gamma_{\min})$, following the schedule of Meng et al. (2025a). The anchor strength μ is held fixed. The anchor is the mean of the $\hat{x}$ iterates over one window, so a short window tracks the decaying perturbation but inherits the noise of the raw iterates, and a long window averages over early, high-γ iterates still far from equilibrium. All sampled experiments use $T_u = \mathrm{budget}/12$, $d = 0.5$, $\gamma_{\min} = 10^{-5}$ and $\mu = 1.0$; the ablation in §5 halves and doubles T_u, and the noise-floor scan in §5 covers μ down to 0.01. The anchor must stay effective: below $\mu \approx 0.3$ on Kuhn the regularization can freeze an action and stall the iterate (§A.5), setting the working range's lower end.

Convergence of the last iterate. The guarantee holds under assumption (M), a single information set with two actions, which makes it a 2×2 normal-form subproblem, and it bounds the distance of the perturbed last iterate to the *regularized* fixed point $\hat{x}^\star$ of the perturbed game. The hypotheses are game and parameter conditions stated in §A.1, and no assumption is made on the trajectory. §5 probes the tree-level regime.

Theorem 4.1 (Push-regime last-iterate convergence)**.** *Let the game and parameter conditions (M), (A-var), (A-int), (A-push) and (A-ε'') in §A hold. Then for the RT-MCCFR$^+$ last iterate $\hat{x}^T$ and its fixed point $\hat{x}^\star$,*

$$\mathbb{E}\,\|\hat{x}^T - \hat{x}^\star\|^2 \;\leq\; K\, T^{-1/2} \qquad \text{for all } T \geq T_0, \tag{3}$$

where K, T_0 are explicit functions of the game, of the fixed window parameters (γ, μ) and of the regularity constants of (A-var) and (A-int), assembled in §A.

The proof in §A is a fresh-start, fixed-anchor statement, and §A.2 collects its notation. At fixed perturbation γ, anchor strength μ and any fixed anchor r in the executed box, in particular any window average of Algorithm 1, the single-set 2×2 dynamics started from zero regrets contract to their regularized fixed point $\hat{x}^\star(\gamma, \mu, r)$, which lies within $O(U/\mu)$ of r for a payoff bounded by U. Assumption (A-push), $|\kappa_A| \leq \mu\delta_0/2$, asks that the anchor strength dominates the transformed best-response curvature, where κ_A is the payoff coupling of the subgame and δ_0 is the interior margin

Table 1: Single-policy output NashConv at the final touch count, mean over 10 seeds. Baselines report their last iterate; RM$^+$ window is the plain window average of RM$^+$ iterates; RT-MCCFR$^+$ reports the raw last iterate and the windowed anchor of the same run; best windowed is the lowest matched-window baseline output, the RM$^+$ window column included (§B.14; M, P, I, K, R: windowed RM$^+$, PRM$^+$, IREG-PRM$^+$, PFTRL-KL, PFTRL-RKL). Bold: lowest single-policy output in the row, windowed columns included. *: the anchor is lower than every baseline last iterate. Paired intervals: §B.1.

Game / sampler	RM$^+$	PRM$^+$	IREG-PRM$^+$	PFTRL-KL (loc.)	PFTRL-RKL	RM$^+$ window	RT-MCCFR$^+$ last	RT-MCCFR$^+$ anchor	best windowed
Kuhn (ES)	0.0995	0.1086	0.1169	0.0191	0.0264	0.0056	0.0153	0.0041$*$	**0.0038**^{P}
Kuhn (OS)	0.1107	0.1124	0.1338	0.0534	0.0590	0.0128	0.0325	0.0085$*$	**0.0076**^{R}
Leduc (ES)	0.4841	0.4337	0.3980	0.3990	0.3266	0.0908	0.2728	**0.0708**$*$	0.0832^{I}
Leduc (OS)	0.4580	0.5102	0.4475	1.4770	0.8867	0.2291	0.3762	**0.1857**$*$	0.2069^{R}
Liar's Dice 1d6 (ES)	0.2253	0.2259	0.2220	1.1589	1.1849	0.1399	0.1519	**0.1002**$*$	0.1335^{I}
Liar's Dice 1d6 (OS)	0.3206	0.3286	0.3114	0.3329	0.3463	0.2698	0.2761	**0.2542**$*$	0.2698^{M}
Goofspiel(4) (ES)	0.7322	0.8623	0.6220	0.2729	0.2615	0.1124	0.4820	**0.1113**$*$	0.1124^{M}
Goofspiel(4) (OS)	0.8390	0.9044	0.9003	0.7096	0.7229	0.1985	0.8330	**0.1226**$*$	0.1307^{R}
Liar's Dice 1d4 (ES)	0.0870	0.0940	0.0842	0.2372	0.3119	0.0391	0.0555	0.0376$*$	**0.0360**^{I}
Liar's Dice 1d4 (OS)	0.1052	0.1012	0.1013	0.0499	0.0640	0.0623	0.0725	0.0451$*$	**0.0301**^{K}

of the fixed point; this dominance makes the drift coercive. (A-var) and (A-ε'') enter only through the regret-norm window, and (A-int) keeps the fixed point inside the box where the potential is compared.

Two features of RM$^+$ under sampling set the rate. The regret norms grow like $\sqrt{t}$, so the implicit step size $1/S_{I,t}$ is a $t^{-1/2}$ schedule, where $S_{I,t} = \|\theta_{I,t}\|_1$ is the ℓ_1 norm of the RM$^+$ regret vector at I. This is the supply lemma (Lemma A.3), which uses the variance floor. Weighting each player's potential by its own regret norm $\|\theta_{I,t}\|_2$ makes weight times step size a bounded quantity, so the anchor term dominates the zero-sum coupling without any condition on the ratio of the two players' step sizes. On matching pennies with $\gamma = 0.05$ and the uniform anchor the hypotheses hold for every $\mu \geq 20$, and the measured squared distance on that instance decays as $T^{-0.5}$ at $\mu \in \{20, 50\}$ and at $\mu = 1$ alike (§A.5). The result covers one information set at a fixed anchor: it bounds the distance of that subproblem to its regularized fixed point and does not by itself give a contraction of the coupled tree dynamics, and the outer schedule that halves γ between windows is set by the empirical protocol of §5. The sufficient conditions are conservative relative to the experimental $\mu = 1.0$.

5 Experiments

We compare RT-MCCFR$^+$ with three regret-matching baselines, RM$^+$, PRM$^+$ and IREG-PRM$^+$ with EMA prediction, and two follow-the-regularized-leader (FTRL) baselines, PFTRL with a reversed Kullback–Leibler (RKL) regularizer and a localized PFTRL with the KL regularizer (§B.8), on Kuhn poker, Leduc hold'em, Liar's Dice with one d6 (1d6), Goofspiel(4) and Liar's Dice with one four-sided die per player (LD(4), 1,024 information sets). All arms share the same implementation, machines and touch budget per game, 2×10^7 touches except 4×10^7 on Leduc, under both external sampling (ES) and outcome sampling (OS), with 10 seeds. Tree touches are the cost axis; one RT-MCCFR$^+$ run of 2×10^7 touches on LD(4) takes about 312s on one core. We report NashConv of the *last* strategy $\hat{x}^T$ and, for RT-MCCFR$^+$, of the *windowed anchor* r_I, the method's output. A control arm, ***RM$^+$ window***, applies the same window average to plain RM$^+$ iterates with no perturbation and no anchor regularization. Fitted exponents are written as mean $\pm$ half-width of a 95% bootstrap interval over seeds, and paired significance tests for every comparison are in §B.1. RT-MCCFR$^+$ uses $\gamma_0 = 0.1$, $\mu = 1.0$ and $T_u = \mathrm{budget}/12$ on every game, and RM$^+$, PRM$^+$ and IREG-PRM$^+$ use their published settings. PFTRL was tuned on Kuhn, and the selected step size, the smallest value of the grid, is used on every other game without further tuning (§B.12).

Single-policy output. Table 1 compares the single-policy outputs, with every baseline given the same window-average output rule (last column, §B.14). The windowed anchor of RT-MCCFR$^+$ is below the best windowed baseline in six of ten settings, significantly on 1d6-ES only, and above it in four, significantly on LD(4)-OS (§B.14). Against baseline last iterates the anchor is below every one. Its raw last iterate is below every RM-family last iterate and below the better FTRL variant in seven.

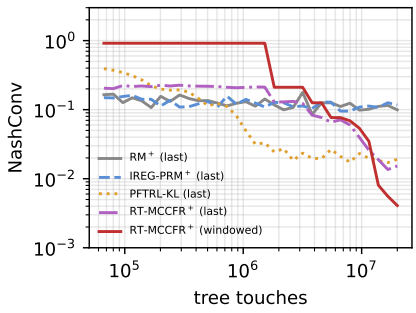

Figure 1: NashConv on Kuhn-ES, mean over 10 seeds. The raw last iterate and the windowed anchor of RT-MCCFR$^+$ decrease over the 2×10^7 touches shown, while the RM$^+$ last iterate plateaus.

Table 2: Budget sweep with a budget-independent window $T_u = 2 \times 10^7/12$, mean NashConv over 10 shared seeds; budgets in touches (M $= 10^6$). Bold: RT-MCCFR$^+$'s anchor at the two operating budgets.

Cell	Budget	RM$^+$ last	RT-MCCFR$^+$ last	RT-MCCFR$^+$ anchor	RM$^+$ window
Kuhn-ES	5M	0.1255	0.0812	0.0767	0.0033
Kuhn-ES	20M	0.0995	0.0153	**0.0041**	0.0056
Kuhn-ES	40M	0.0882	0.0111	**0.0034**	0.0060
Kuhn-OS	5M	0.1251	0.0812	0.0817	0.0130
Kuhn-OS	20M	0.1107	0.0325	**0.0085**	0.0128
Kuhn-OS	40M	0.1006	0.0192	**0.0066**	0.0100
Leduc-ES	5M	0.5269	0.3560	0.1900	0.1248
Leduc-ES	20M	0.4456	0.3073	**0.0945**	0.1252
Leduc-ES	40M	0.4841	0.2698	**0.0899**	0.1408
Leduc-OS	5M	0.6384	0.5618	0.4052	0.3675
Leduc-OS	20M	0.4844	0.4023	**0.2359**	0.2970
Leduc-OS	40M	0.4580	0.3903	**0.2461**	0.3098

Most of the single-policy gain over the last iterates comes from the window itself. The reward transformation lowers the anchor further than the RM$^+$ window in all ten settings and lowers the raw last iterate (Table 4). The three exceptions, Goofspiel(4) under both samplers and LD(4)-OS, are the cells where the PFTRL last iterate is still decreasing at the budget (Table 3); even there the anchor stays below PFTRL. Paired tests over the shared seeds confirm the anchor's ordering against every baseline last iterate in every cell except LD(4)-OS against PFTRL-KL (§B.1). PFTRL reaches NashConv above one on two cells because its Kuhn-tuned step size is unsuited to those games.

The control column separates the two ingredients of the anchor. Window averaging alone lowers the RM$^+$ output by a factor between 1.2 and 17.8, and RT-MCCFR$^+$'s anchor is lower than the control in all ten cells, by a factor above 1.1 in seven and by at most 1.07 on Goofspiel(4)-ES, LD(4)-ES and 1d6-OS. The window average is a better output when the trajectory it averages is itself pulled toward a fixed point by the perturbation and the anchor. The full-trajectory average moves the other way: the perturbation inflates it (§B.2), and the comparison target of RT-MCCFR$^+$ is the last iterate.

Does the last iterate keep decreasing? A single budget cannot distinguish a stationary floor from slow convergence, so we look at the trajectory in two ways. Table 3 fits a power law $\mathrm{expl}(t) \propto t^\beta$ to every arm's last iterate over eight of the 32 log-spaced checkpoints in $[B/4, B]$; it is a finite-budget diagnostic, and a floor gives $\beta \approx 0$. The RM-family baselines are flat on Kuhn and Leduc, with $|\beta| \leq 0.18$ in eleven of twelve cells, and decrease slowly on the three larger games, always from a higher level. RT-MCCFR$^+$'s last iterate has the steepest slope of its row on Kuhn and LD(4)-ES, a slope comparable to PFTRL on LD(4)-OS, and a moderate one on 1d6. PFTRL decreases on Goofspiel(4), LD(4) and 1d6-OS, which is why it overtakes RT-MCCFR$^+$'s raw last iterate on Goofspiel(4) and LD(4)-OS, and is flat on Kuhn and Leduc. Two cases are different in kind: on Leduc the RT-MCCFR$^+$ slope is weak although its level is the lowest, and on Goofspiel(4)-OS the last iterate keeps a floor while the anchor still beats every baseline. Table 2 repeats the Kuhn and Leduc runs at three budgets with one window $T_u = 2 \times 10^7/12$ held fixed across budgets, in place of the per-budget $T_u = \mathrm{budget}/12$ of Table 1, so its Leduc rows differ from Table 1 at the same budget; §B.9 continues them to 8×10^7 touches. Past the transient no anchor change is resolved at 10 seeds: from 2 to 4×10^7 touches its mean falls in three cells and rises on Leduc-OS, and from 4 to 8×10^7 it rises in three of four. The raw last iterate falls over both ranges (§B.1). The eightfold ratios from 5×10^6 to 4×10^7 are therefore gains out of the exploration transient, not a rate: after three windows $\gamma \approx 0.0125$ has not reached its floor and the anchor sits above the plain window average. Over that range the RM$^+$ last iterate improves by at most 30%, while the RT-MCCFR$^+$ anchor and last iterate fall by factors of 1.6–23 and 1.3–7.3. Figure 1 shows the Kuhn-ES curves.

Which part of the anchor matters. Table 4 replaces the windowed anchor by alternatives on Kuhn and Leduc. Anchoring to the last iterate raises the anchor output to the level of the last iterate while moving the last output itself by 13% or less on three of the four cells. The window average

Table 3: Last-iterate power-law slope β over eight of the 32 log-spaced checkpoints in $[B/4, B]$, mean over 10 seeds, for each arm's raw last iterate; a finite-budget diagnostic.

Game / sampler	RM$^+$	PRM$^+$	IREG-PRM$^+$	PFTRL-KL (loc.)	PFTRL-RKL	RT-MCCFR$^+$ last
Kuhn (ES)	-0.18	-0.06	-0.00	-0.24	-0.07	-1.50
Kuhn (OS)	-0.13	+0.02	-0.27	-0.00	+0.07	-1.02
Leduc (ES)	-0.04	-0.15	-0.17	-0.32	-0.07	-0.11
Leduc (OS)	-0.14	-0.13	-0.16	+0.16	+0.01	-0.24
Liars Dice 1d6 (ES)	-0.26	-0.25	-0.25	-0.16	-0.14	-0.44
Liars Dice 1d6 (OS)	-0.36	-0.36	-0.35	-0.54	-0.58	-0.39
Goofspiel(4) (ES)	-0.19	-0.10	-0.29	-0.61	-0.68	-0.26
Goofspiel(4) (OS)	-0.20	-0.18	-0.02	-0.35	-0.40	+0.06
Liar's Dice 1d4 (ES)	-0.36	-0.36	-0.36	-0.80	-0.62	-0.91
Liar's Dice 1d4 (OS)	-0.46	-0.56	-0.40	-1.05	-0.98	-0.82

Table 4: Windowed-anchor ablation: last/anchor output NashConv at the final touch count, mean over 10 seeds, on Kuhn and Leduc. The two rows above the last isolate the perturbation (μ=0) and the anchor regularization (γ=0) of the windowed rule.

Anchor rule	Kuhn (ES)	Kuhn (OS)	Leduc (ES)	Leduc (OS)
windowed T_u=$B/12$	0.0153/0.0041	0.0325/0.0085	0.2728/0.0708	0.3762/0.1857
last-iterate	0.0206/0.0207	0.0353/0.0355	0.3029/0.3029	0.3578/0.3579
EMA β=0.9	0.0902/0.2020	0.0942/0.2033	0.2694/1.3452	0.3293/1.2880
EMA β=0.99	0.1681/0.7902	0.1686/0.7909	0.4423/4.3572	0.4571/4.3238
T_u=$B/24$ (frequent)	0.0145/0.0044	0.0270/0.0089	0.2698/0.0899	0.3903/0.2461
T_u=$B/6$ (sparse)	0.0328/0.0343	0.0321/0.0269	0.2934/0.0832	0.3263/0.1488
perturb. only (μ=0)	0.1184/0.0041	0.1233/0.0109	0.3936/0.0866	0.4491/0.2227
anchor only (γ=0)	0.0123/0.0045	0.0296/0.0088	0.2791/0.0739	0.3724/0.1887
no perturb. / no anchor (γ=μ=0)	0.0995/0.0056	0.1107/0.0128	0.4841/0.0908	0.4580/0.2291

is therefore what lowers the reported output, through the variance of the anchor; the drift of the raw iterate is set by the contraction of the anchor regularization. Exponential smoothing at the 12 refreshes with $\beta = 0.9$ or 0.99 keeps 28% or 89% of the initial anchor, leaving its error one to two orders of magnitude above the windowed anchor. Changing T_u also changes the number of γ halvings, so the window rows are a joint schedule sensitivity. With the halving clock held fixed at $B/12$ and the anchor window at $B/24$ or $B/6$, so that every run ends at $\gamma = 2.4 \times 10^{-5}$, the pattern of Table 4 survives on the same cells and 10 seeds (Table 13): the more frequent refresh raises the Leduc anchor by up to 45%, while the sparse window costs a factor of 7.4 (ES) and 2.8 (OS) on the Kuhn anchor and doubles the Kuhn-ES last output. The window length, not the number of halvings, drives these rows, so high-perturbation iterates in the final window do not explain the $B/6$ cost. The last row removes the perturbation and the anchor together ($\gamma = \mu = 0$), where the last iterate sits at the plain RM$^+$ level and the window average is the control column of Table 1; the two rows above it run the ingredients separately. Removing the anchor ($\mu = 0$) leaves the last iterate at or above the RM$^+$ level on Kuhn and only part of the way down on Leduc; removing the perturbation ($\gamma = 0$) is statistically indistinguishable from the full method in every cell, with a paired t test at $p \geq 0.4$ and a gap within 0.01. The anchor regularization is what lowers the last iterate; the perturbation adds a reach floor that helps only on the deeper tree, where Leduc-ES moves from 0.484 to 0.394 at $p = 0.01$, and is invisible on Kuhn.

What sets the noise floor. A step-size heuristic drawn from the proof of Theorem 4.1 predicts the level of the last iterate under sampling. The RM$^+$ regret norm grows like $\sigma_{\text{tot}}\sqrt{t}$, where σ_{tot}^2 is the per-step variance of the transformed counterfactual values, so the update is a stochastic approximation with step $\propto 1/(\sigma_{\text{tot}}\sqrt{t})$ and noise variance σ_{tot}^2. Balancing the contraction μ·step against the injected variance $\text{step}^2\sigma_{\text{tot}}^2$ gives $\mathbb{E}\|\hat{x} - \hat{x}^\star\|^2 \propto \sigma_{\text{tot}}/(\mu\sqrt{t})$. We test this on Kuhn at fixed $\gamma = 0.05$, a fixed uniform anchor and no outer updates, injecting zero-mean noise of variance σ^2 into the counterfactual values on top of the natural ES or OS variance. The response variable D_{floor} is the mean squared distance of the last iterate to $\hat{x}^\star(\gamma, \mu)$, computed by a full-feedback run

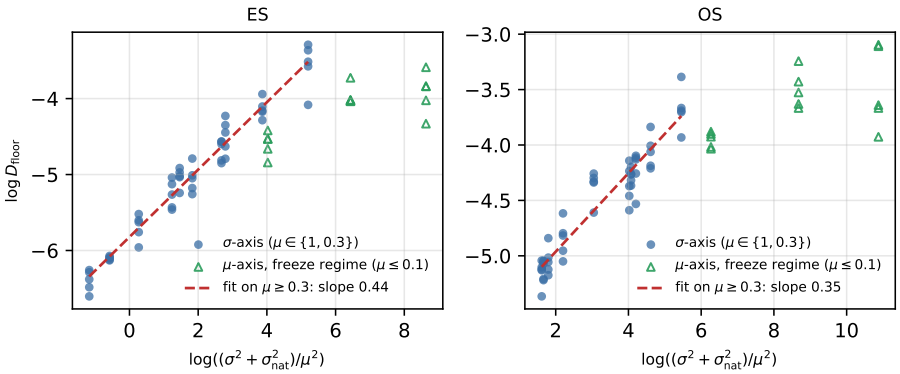

Figure 2: Noise floor on Kuhn: mean squared distance of the last iterate to the regularized fixed point against $(\sigma^2 + \hat{\sigma}^2_{\text{nat}})/\mu^2$ (log-log), under ES (left) and OS (right). Circles: σ-axis at $\mu \in \{1, 0.3\}$; hollow triangles: μ-axis at $\sigma = 0.5$ in the freeze regime $\mu \le 0.1$. Dashed: least-squares line over $\mu \ge 0.3$, slope 0.44 ± 0.03 (ES) and 0.35 ± 0.03 (OS); the prediction is $1/2$.

at the same μ (Figure 2). The grid crosses $\sigma \in \{0, 0.5, 1, 2, 4\}$ with $\mu \in \{1.0, 0.3\}$, and a second sweep varies μ down to 0.01 at $\sigma = 0.5$, under both samplers with 5 seeds. On each σ-axis we fit $\log D_{\text{floor}} = c + p \log(\sigma^2 + \sigma^2_{\text{nat}})$ with the natural variance σ^2_{nat} free, so the exponent is read off without assuming the injected noise dominates (§B.4). Along the σ-axis the ES exponent in $\sigma^2 + \hat{\sigma}^2_{\text{nat}}$ is 0.47 ± 0.10 at $\mu = 1$ and 0.45 ± 0.15 at $\mu = 0.3$, matching the predicted $1/2$ with a natural variance $\hat{\sigma}^2_{\text{nat}} \approx 0.3$. Under OS the natural variance is an order of magnitude larger, so the injected range does not identify the exponent. Along the μ-axis the ES floor falls as $\mu^{-0.87 \pm 0.11}$ between $\mu = 0.3$ and 1, roughly inverse in μ (the interval excludes an exact $1/\mu$), and flattens below $\mu \approx 0.1$, where the anchor no longer dominates the coupling and the iterate enters the freeze regime of §A.5. Against the predicted composite $(\sigma^2 + \hat{\sigma}^2_{\text{nat}})/\mu^2$ the working-range points in Figure 2 lie on a line of slope 0.44 ± 0.03 (ES) and 0.35 ± 0.03 (OS), short of the heuristic $1/2$, most under OS where the natural variance is not identified (§B.4). The sampling variance is the primary lever on the floor and the anchor strength a bounded one. Under OS the reference solver contracts to the fixed point of a rescaled game, at most 3.8×10^{-4} from the ES reference over the μ grid (§B.5).

Rate check in the proven regime. Table 6 checks the rate of Theorem 4.1 on matching pennies itself, inside the proven regime ($\mu \in \{20, 50\}$) and outside it ($\mu = 1$), under both samplers (§A.5). It reports the seed-mean log-log slope of $\|\hat{x}^t - \hat{x}^\star\|^2$ with its seed-bootstrap interval, and the constant $\sqrt{T}\,\mathbb{E}\|\hat{x}^T - \hat{x}^\star\|^2$ at the final touch count. Every slope lies within a few hundredths of the $T^{-1/2}$ of (3), and the constant decreases in μ as the $O(\sigma/\mu)$ heuristic above predicts. At $\mu = 1$ the rate persists where only the sufficient conditions fail.

Initialization dependence. On a two-stage game with an off-path information set, the unregularized RM$^+$ last iterate leaves player 2's probability of the weakly dominant action at 0.0, 0.5 or 1.0 depending on the start, while RT-MCCFR$^+$ puts it at 1.0 from all five starts, at a last-iterate NashConv of 7.6×10^{-7}, the exploration floor of that run (§B.13).

6 Conclusion

We introduced RT-MCCFR$^+$, a reward-transformed RM$^+$ algorithm whose last iterate lowers the noise floor under MCCFR sampling. A γ-perturbation keeps reach probabilities bounded below, a μ-anchor makes local subproblems strongly concave, and a windowed average is both the anchor and the output. Over five games, two samplers and 10 seeds the anchor is below the best matched-window baseline in six of ten cells, one difference resolved in each direction, and below every baseline last iterate; the raw last iterate is below every RM-family baseline. Its floor rises with exponent 0.47 ± 0.10 in the ES sampling variance and falls with the anchor strength; a stylized theory gives an $O(T^{-1/2})$ squared-distance bound to the regularized fixed point. Transferring the anchor to parameterized policies and reducing the average-policy cost are the main open directions.

AI USE STATEMENT

We used generative AI tools for manuscript preparation (drafting, polishing, proofreading) and for the experiments (implementation, execution, figures, numerical analysis). They were not used to conceive the contributions, derive or verify the theory, or formulate the claims. All AI-assisted output was reviewed and cross-checked by the authors, who take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The proof of Theorem 4.1 and all assumptions are in Appendix A; Appendix B documents the experimental protocol, per-cell means, paired tests, per-seed endpoints, hyperparameters, and tuning. Anonymous code, with per-seed CSV files, is provided; the core modules are distributed as compiled binaries and the full code will be released upon acceptance.

REFERENCES

Ioannis Anagnostides, Gabriele Farina, Ioannis Panageas, and Tuomas Sandholm. Optimistic mirror descent either converges to nash or to strong coarse correlated equilibria in bimatrix games. *Advances in Neural Information Processing Systems*, 35:16439–16454, 2022.

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.

Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *International Conference on Machine Learning*, pp. 793–802. PMLR, 2019.

Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games. *Advances in Neural Information Processing Systems*, 33:17057–17069, 2020.

Neil Burch, Marc Lanctot, Duane Szafron, and Richard Gibson. Efficient monte carlo counterfactual regret minimization in games with many player actions. *Advances in Neural Information Processing Systems*, 25, 2012.

Neil Burch, Martin Schmid, Matej Moravcik, Dustin Morrill, and Michael Bowling. Aivat: A new variance reduction technique for agent evaluation in imperfect information games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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. URL https://openreview.net/forum?id=LWeVVPuIx0.

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.

Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng. Separation of nonergodic uniform convergence rates for regularized learning in games. *Operations Research*, 2026.

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

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Regret minimization in behaviorally-constrained zero-sum games. In *International Conference on Machine Learning*, pp. 1107–1116. PMLR, 2017.

Gabriele Farina, Christian Kroer, Noam Brown, and Tuomas Sandholm. Stable-predictive optimistic counterfactual regret minimization. In *International Conference on Machine Learning*, pp. 1853–1862. PMLR, 2019a.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Regret circuits: Composability of regret minimizers. In *ICML*, Proceedings of Machine Learning Research, pp. 1863–1872. PMLR, 2019b.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Optimistic regret minimization for extensive-form games via dilated distance-generating functions. *Advances in Neural Information Processing Systems*, 32, 2019c.

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.

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.

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

Samid Hoda, Andrew Gilpin, Javier Pena, and Tuomas Sandholm. Smoothing techniques for computing nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2):494–512, 2010.

Michael Johanson, Nolan Bard, Marc Lanctot, Richard G Gibson, and Michael Bowling. Efficient nash equilibrium approximation through monte carlo counterfactual regret minimization. In *AAMAS*, pp. 837–846, 2012.

Christian Kroer, Kevin Waugh, Fatma Kilinç-Karzan, and Tuomas Sandholm. Faster first-order methods for extensive-form game solving. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pp. 817–834, 2015.

Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. Monte carlo sampling for regret minimization in extensive games. *Advances in Neural Information Processing Systems*, 22, 2009.

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

Boning Li and Longbo Huang. Real-time parallel counterfactual regret minimization. *arXiv preprint arXiv:2605.19928*, 2026.

Boning Li, Yu Chen, and Longbo Huang. Av-aivat: 74x cheaper agent evaluation with certified anytime-valid stopping in imperfect-information games. *arXiv preprint arXiv:2608.06362*, 2026a.

Boning Li, Yu Chen, and Longbo Huang. Correlated chance sampling for monte carlo counterfactual regret minimization. *arXiv preprint arXiv:2607.27035*, 2026b.

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.

Wataru Masaka, Mitsuki Sakamoto, Kenshi Abe, Kaito Ariu, Tuomas Sandholm, and Atsushi Iwasaki. On the power of perturbation under sampling in solving extensive-form games, 2025. URL https://arxiv.org/abs/2501.16600.

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.

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.

Martin Schmid, Neil Burch, Marc Lanctot, Matej Moravcik, Rudolf Kadlec, and Michael Bowling. Variance reduction in monte carlo counterfactual regret minimization (vr-mccfr) for extensive form games using baselines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2157–2164, 2019.

Ashkan Soleymani, Georgios Piliouras, and Gabriele Farina. Faster rates for no-regret learning in general games via cautious optimism. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 518–529, 2025.

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

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.

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 Proof of Theorem 4.1

A.1 Setting and assumptions

The theorem is proved for a single information set I: a two-player zero-sum 2×2 normal-form game with payoff $A \in [-U, U]^{2\times 2}$, run as Algorithm 1 restricted to I within one window, i.e. with fixed (γ, μ, r), every policy query (the traverser's own $\hat{x}_I$ and the strategy from which an opponent action is sampled) recomputed from the current regret vector, where the anchor $r = (r_1, r_2)$ is any fixed pair of strategies in the executed box, $r_{i,1} \in [\gamma/2, 1 - \gamma/2]$ (the window average of executed strategies, and the initial uniform anchor, lie there); alternating updates, bounded dither $|\xi_{i,j}| \le \tau_d$, and the perturbative parameters $\mu > 0$ and $\gamma \in [0, 1)$ (so $\tilde{\mu} = \mu(1 - \gamma) > 0$); the RM$^+$ regrets start at 0. Each player observes its whole payoff column against one sampled opponent action. In the external-sampling implementation on the two-node tree of a 2×2 game this is the second mover's feedback; the first mover's two rows are evaluated against two independent opponent draws, and its payoff difference then has conditional variance $p(1-p)\big[(A_{11} - A_{12})^2 + (A_{21} - A_{22})^2\big] \ge p(1-p)\kappa_A^2/2$, at least half of the shared-column variance (with equality on matching pennies, §A.5). (A-var) below is stated for the shared-column oracle; under the implemented sampler the first mover's floor has $\kappa_A^2/2$ in place of κ_A^2, and $\sigma^2_{\min}$ always denotes the *smaller* of the two players' floors, which is the value entering every constant of the proof. The outcome-sampling checks in §A.5 are empirical. Let $\hat{x}^\star = \hat{x}^\star(\gamma, \mu, r)$ denote the regularized perturbed equilibrium of the window and $x^\star_{\mathrm{b}}$ its base-strategy representation (§A.2). The conditions below are explicit game and parameter conditions; they are the exact hypotheses under which the statement is proven. Extending the result from this single-set setting to full trees requires an additional projection argument for $n \ge 3$ action sets (§A.5); the tree-level regime is what the experiments in §5 probe.

(M) The single-information-set setting above.

(A-var) The per-step sampling variance is bounded below, $\sigma^2_{\min} = \frac{\gamma}{2}(1-\frac{\gamma}{2})\kappa_A^2 + 2\sigma_d^2 > 0$, with κ_A the payoff coupling and σ_d^2 the dither variance defined in §A.2; the derivation of the formula is (7), which uses the fresh-sampling property stated there. Namely, each player's round-t randomization (the opponent's action, drawn from the opponent's executed strategy $\hat{x}_{-i}$, and two dither components) is fresh at its pre-update filtration $\mathcal{H}_t^{(i)}$: driven by uniform variables independent of the past, with conditional mean 0 and variance σ_d^2 for the dither and mutual conditional independence of the three components given $\mathcal{H}_t^{(i)}$, so that $\mathbb{E}[\tilde{D}_i^2 \mid \mathcal{H}_t^{(i)}] \geq \sigma^2_{\min}$ holds for *both* players at their actual update time (player 2 at $\mathcal{H}_t^{(2)} = \mathcal{F}_{t+1/2}$); under the implemented sampler the first mover's floor has $\kappa_A^2/2$ in place of κ_A^2, and $\sigma^2_{\min}$ is the smaller of the two.

(A-int) The perturbed fixed point is interior with margin $\delta_0 \in (0, \frac{1}{2})$: $z^\star \in [\delta_0, 1-\delta_0]^2$ (in particular $z^\star \in (0,1)^2$); an explicit sufficient condition is (41) of Lemma A.8: the anchor's base coordinate $\bar{z}$ lies $\sqrt{2}U/\tilde{\mu}$ inside the margin. Every use below requires only this closed margin.

(A-push) The transformed best-response curvature is dominated by the anchor strength, $|\kappa_A| \leq \mu\delta_0/2$.

(A-ε'') *(explicit form)* The norm-plus fallback threshold satisfies

$$\mathrm{EPS} \;\leq\; \min\Big(\tfrac{G_D}{4},\; \tfrac{\sigma_{\min}}{8}\Big), \tag{4}$$

with G_D defined in §A.2. (These two bounds are exactly the uses made in Lemma A.3; $G_D, \sigma_{\min}$ are global constants, so (4) is a condition on the threshold EPS alone. Both bounds are nondecreasing in μ, so the implemented $\mathrm{EPS} = 10^{-12}$ satisfies (4) for every μ as soon as $\sigma_{\min} \geq 8 \cdot 10^{-12}$. On the regret window used in the proof one has $S_{i,t} \geq 4G_D > \mathrm{EPS}$, so no fallback step is ever counted there.)

No condition on the ratio of the two players' step sizes is needed: the weighted Lyapunov function of §A.2 absorbs the step-size asymmetry pointwise (Lemma A.9).

A.2 Notation

Fix the perturbed iterate $\hat{x}^t$; player i's base strategy is $x_{\mathrm{b},i} = (z_i, 1-z_i)$ with $z_i \in [0,1]$, and by the single-set reduction the dynamics are the two-dimensional recursion $z^t = (z_1^t, z_2^t) \in [0,1]^2$ with interior fixed point $z^\star \in [\delta_0, 1-\delta_0]^2$ (A-int); the executed strategy is $\hat{x}_{i,1} = (1-\gamma)z_i + \gamma/2$ (the perturbation mixes in the uniform distribution; the anchor r enters only the reward transform), so $\hat{x} - \hat{x}^\star = (1-\gamma)(x_{\mathrm{b}} - x_{\mathrm{b}}^\star)$ componentwise. Write $\rho_i := 2r_{i,1} - 1 \in [-(1-\gamma), 1-\gamma]$ for the anchor's action bias and $\bar{z}_i := (r_{i,1} - \gamma/2)/(1-\gamma) \in [0,1]$ for the base coordinate whose executed strategy equals the anchor. Player i's RM$^+$ regret vector is $\theta_{i,t} \in \mathbb{R}^2_+$ with $\theta_{i,0} = 0$, regret norm $S_{i,t} = \|\theta_{i,t}\|_1$, and, on a non-fallback step ($S_{i,t} > \mathrm{EPS}$, so that $z_i^t = \theta_{i,t,1}/S_{i,t}$), step size $a_{i,t} = 1/S_{i,t}$ and

$$\bar{a}_t := \frac{a_{1,t} + a_{2,t}}{2}$$

(the arithmetic mean; the constant C_w of Lemma A.13 is stated at this scale, $a_{1,t} + a_{2,t} = 2\bar{a}_t$). Every use of $a_{i,t}$, of the identity $N_{i,t} = S_{i,t}\sqrt{\varphi(z_i^t)}$ below and of the virtual increments is on the event $\{\varrho > t\}$ of (8), where $S_{i,t} \geq 4G_D > \mathrm{EPS}$ for both players; off that event these quantities are not used. Alternation uses the half-step filtration $\mathcal{F}_{t+1/2} = \mathcal{F}_t \vee \sigma(\text{player 1's round-}t\text{ sample and dither})$. Player i's *pre-update filtration* at round t is $\mathcal{H}_t^{(1)} = \mathcal{F}_t$ (player 1 samples and moves at t) and $\mathcal{H}_t^{(2)} = \mathcal{F}_{t+1/2}$ (player 2 samples and moves after player 1's round-t update). All sampling and dither hypotheses below are stated relative to $\mathcal{H}_t^{(i)}$. The statistics $a_{1,t}, a_{2,t}, N_{1,t}, N_{2,t}, z^t$ (and hence $\Phi_i'(z_i^t)$) are $\mathcal{F}_t$-measurable, hence $\mathcal{H}_t^{(i)}$-measurable for both i.

Weighted state-dependent Lyapunov function. Write $e^t = z^t - z^\star$ and $\varphi(v) = v^2 + (1-v)^2$. On the simplex segment $[0,1]$, $\varphi \in [\frac{1}{2}, 1]$ and $|\varphi'| \leq 2$; on all of $\mathbb{R}$ only $\varphi \geq \frac{1}{2}$ holds (φ' is

unbounded there, with $|\varphi'| \le 4$ on the extended domain $\mathcal{Z}^+ = [-\frac{1}{2}, \frac{3}{2}]^2$ of the following paragraph). Two weights appear in the normalized dynamics and both are removed by the choice of Lyapunov function. First, the expected virtual increment of player i is $a_{i,t}\varphi(z_i^t)f_i(z^t)$ (Lemma A.6), so the plain quadratic $\|z - z^\star\|^2$ sees the coupling as $\tilde{\kappa}e_1e_2(\varphi(z_1) - \varphi(z_2))$, which no fixed norm cancels (Remark A.10); the potential

$$V(z) = \sum_{i=1}^{2} \Phi_i(z_i), \qquad \Phi_i(v) := \int_{z_i^\star}^{v} \frac{s - z_i^\star}{\varphi(s)}\, ds, \tag{5}$$

has $\Phi_i'(v) = (v - z_i^\star)/\varphi(v)$ and cancels φ pointwise. Second, the two players run with different step sizes $a_{1,t} \neq a_{2,t}$, so the drift seen by V is $\sum_i a_{i,t}e_if_i$, whose skew part $\tilde{\kappa}e_1e_2(a_{1,t} - a_{2,t})$ vanishes only when the two step sizes agree. The proof therefore runs the recursion on the *weighted* function

$$V_t := \sum_{i=1}^{2} N_{i,t}\, \Phi_i(z_i^t), \qquad N_{i,t} := \|\theta_{i,t}\|_2 = S_{i,t}\sqrt{\varphi(z_i^t)}, \tag{6}$$

where the identity for $N_{i,t}$ follows from $\theta_{i,t} = S_{i,t}(z_i^t, 1 - z_i^t)$. The weight times the step size is $N_{i,t}a_{i,t} = \sqrt{\varphi(z_i^t)} \in [1/\sqrt{2}, 1]$, a state-dependent quantity confined to a fixed interval; the skew part of the drift is then bounded by $\frac{1}{2}(1 - 1/\sqrt{2})|\tilde{\kappa}|\,\|e\|^2$ and is dominated by the anchor term $\sqrt{2}\tilde{\mu}\|e\|^2$ (Lemma A.9(V2)). The weight itself moves by at most $G_D^2/(2N_{i,t}) = O(a_{i,t})$ per step (orthogonality (17) and nonexpansive clipping), which is second order relative to the drift and is charged to the constant C_w. Since $N_{i,t} \asymp \sqrt{t}$ on the regret window and $\mathbb{E}\|e^t\|^2 = O(t^{-1/2})$ is the target upper rate, $\mathbb{E}W_t$ is an $O(1)$ quantity, and the recursion of Lemma A.13 is of the form $w_{t+1} \le (1 - c/\sqrt{t})w_t + C'/\sqrt{t}$ for $w_t = \mathbb{E}W_t$. Lemma A.9 proves (V1) $\frac{1}{2}\|e\|^2 \le V(z) \le \|e\|^2$ (so $V \le 2$ on $[0,1]^2$) and (V2) the weighted drift bound. Each Φ_i is C^2 and single-valley on $\mathbb{R}$ (Φ_i' has the sign of $v - z_i^\star$), which is the property used by the projection lemma. The virtual (unprojected) point z_v^{t+1} lives in the extended domain $\mathcal{Z}^+ = [-\frac{1}{2}, \frac{3}{2}]^2$, where the Hessian is bounded by $L_V^+ = 26$ (Lemma A.12).

Symbols and constants (definition, dependence, first use).

- $\kappa_A := A_{11} - A_{12} - A_{21} + A_{22}$: the 2×2 payoff coupling. First use: (A-var).
- σ_d^2: per-component dither variance. The randomization entering player i's round-t payoff difference, that is, the *opponent's* action, drawn from the opponent's *executed* strategy $\hat{x}_{-i} = (1-\gamma)x_{\mathrm{b},-i} + \frac{\gamma}{2}\mathbf{1}$ as it stands at player i's update time, together with the two dither components $\xi_{i,1}, \xi_{i,2}$ of the observed payoff, is *fresh* at the pre-update filtration $\mathcal{H}_t^{(i)}$: the dither has conditional mean $\mathbb{E}[\xi_{i,j} \mid \mathcal{H}_t^{(i)}] = 0$, $|\xi_{i,j}| \le \tau_d$, and conditional variance σ_d^2, and the opponent's action draw, $\xi_{i,1}$ and $\xi_{i,2}$ are mutually conditionally independent given $\mathcal{H}_t^{(i)}$ (the uniform variables driving them are independent of the past; the action itself has the $\mathcal{H}_t^{(i)}$-measurable law $\hat{x}_{-i}$, so it is not unconditionally independent of earlier randomness). This is the standard fresh-sampling property of MCCFR; stating it relative to $\mathcal{H}_t^{(i)}$, not $\mathcal{F}_t$, is what lets it survive the player-2 half-step enlargement $\mathcal{H}_t^{(2)} = \mathcal{F}_{t+1/2}$. First use: (A-var).
- $\sigma_{\min}^2$: the per-step variance lower bound of (A-var). For player i, D_i is the payoff difference to the *opponent*'s sampled action, dither included when present ("un-transformed" below means before the anchor regularization, not before dither); conditioning on $\mathcal{H}_t^{(i)}$, the opponent's mixing probability at player i's sampling time, $p := \hat{x}_{-i,1} \in [\gamma/2, 1 - \gamma/2]$, is $\mathcal{H}_t^{(i)}$-measurable: for $i = 1$ the opponent (player 2) has not moved this round, so $p = \hat{x}_{2,1}^t$ is $\mathcal{F}_t$-measurable, while for $i = 2$ the opponent (player 1) has just moved, so $p = \hat{x}_{1,1}^{t+1}$ is $\mathcal{F}_{t+1/2}$-measurable. E.g. for $i = 1$, D_1 takes $A_{11} - A_{21} + (\xi_{1,1} - \xi_{1,2})$ with probability $p = \hat{x}_{2,1}^t$ and $A_{12} - A_{22} + (\xi_{1,1} - \xi_{1,2})$ otherwise. Sampling and the two dither components are mutually conditionally independent given $\mathcal{H}_t^{(i)}$, so $\mathrm{Var}(\xi_{i,1} - \xi_{i,2} \mid \mathcal{H}_t^{(i)}) = 2\sigma_d^2$ and

$$\mathrm{Var}(D_i \mid \mathcal{H}_t^{(i)}) = p(1-p)\kappa_A^2 + 2\sigma_d^2 \;\ge\; \sigma_{\min}^2 := \tfrac{\gamma}{2}(1 - \tfrac{\gamma}{2})\kappa_A^2 + 2\sigma_d^2, \tag{7}$$

(player 2 identical up to sign, with the roles of the players swapped).

- $\tilde{\mu} := \mu(1-\gamma)$, $\tilde{\kappa} := (1-\gamma)\kappa_A$: anchor and coupling after folding the perturbation into the update.
- $G_D := 2U + 2\tau_d + 2\mu$: deterministic bound on $|\tilde{D}_i|$ (the sampling increments satisfy $|D_i| \le 2(U+\tau_d)$, $\tilde{D}_i = D_i - \tilde{\mu}(2z_i - 1) + \mu\rho_i$, $|2z_i - 1| \le 1$, $|\rho_i| \le 1-\gamma$, so the anchor terms contribute at most $2\tilde{\mu} \le 2\mu$).
- $F_\infty := 2U + 2\mu$: bound on the drift field, $|f_i| \le F_\infty$ (Lemma A.7).
- $L_V^+ = 26$: extended Hessian bound on $\mathcal{Z}^+$ (Lemma A.12); $L_V = 10$ is the bound on $[0,1]$.
- $m_w := \sqrt{2}\,\tilde{\mu} - \frac{1}{2}\big(1 - \frac{1}{\sqrt{2}}\big)|\tilde{\kappa}|$: the weighted drift coefficient of Lemma A.9(V2); under (A-push), $|\tilde{\kappa}| \le \tilde{\mu}/4$ and $m_w \ge 1.37\,\tilde{\mu} > 0$.
- $C_w := 114\,G_D^2 + 4|\tilde{\kappa}|G_D$: assembly constant of the one-step recursion (50) (weight drift, Taylor residual, quadratic term and alternating deviation; the accounting is in the proof of Lemma A.13).
- $\Delta z_i^v := z_{v,i}^{t+1} - z_i^t$: virtual (unprojected) increment; by Lemma A.6, $\Delta z_i^v = \tilde{D}_i \varphi(z_i^t)/(S_{i,t} + \tilde{D}_i(1 - 2z_i^t))$, with the expansion and amplitude bound of (37) on $\{\varrho > t\}$.
- EPS > 0: the norm-plus fallback threshold (§A.1, (A-ε'')); a step is a *fallback* step when $S_{i,t} \le$ EPS, at which the strategy resets to uniform.
- $t_0^\dagger$, $t_0(T)$, T_0: burn-in and horizon thresholds of (54)–(56) in the proof.

Truncation. To keep the recursion trackable while the regret norms are in their two-sided window, truncate:

$$\varrho = \min\Big\{t \ge t_0 :\ \exists i,\ S_{i,t} \notin [s_1\sqrt{t}, s_2\sqrt{t}]\Big\}, \qquad W_t = V_t \mathbb{1}\{\varrho > t\}, \tag{8}$$

with V_t the weighted function (6), $s_1 = \sqrt{c_Y/2}$, $s_2 = \sqrt{3}\,G_D$ and c_Y as in Lemma A.3; $\{\varrho > t\} \in \mathcal{F}_t$, so no clipping indicator enters any $\mathcal{F}_t$-conditional quantity. The truncation carries no final-clip tail: on $\{\varrho > t\}$ every counted step is a non-fallback step with $S_{i,t} \ge 4G_D$, W_T decays through (50) directly, and the restoration step (62) in Step 3 reconstructs the untruncated value.

A.3 Four lemmas

The proof rests on four ingredients: a product-form recursion inequality *(lem:chung)* and a *supply* lemma bounding the regret-norm growth (both discharged in Steps 1–2 below), a projection-domination lemma *(lem:proj)*, and the drift-decomposition engine behind the truncated one-step recursion *(lem:rec)*. We state and prove them in that order.

Lemma A.1 (Product-form recursion bound)**.** *Suppose $w_t \ge 0$, $C' \ge 0$ and*

$$w_{t+1} \le \Big(1 - \frac{c}{\sqrt{t}}\Big)w_t + \frac{C'}{\sqrt{t}}, \qquad t_0 \le t < T, \tag{9}$$

where $1 \le t_0 < T$ and $c \in (0, \sqrt{t_0}]$ (in particular $c > 0$). Then, with $P_{t_0,T} := \prod_{t=t_0}^{T-1}(1 - c/\sqrt{t}) \in [0,1]$,

$$w_T \ \le\ P_{t_0,T}\,w_{t_0} + \frac{C'}{c}\big(1 - P_{t_0,T}\big) \ \le\ w_{t_0}\,e^{-2c(\sqrt{T} - \sqrt{t_0})} + \frac{C'}{c}. \tag{10}$$

The constant in front of the forcing term depends only on (c, C'), not on t_0 or T.

Proof. Write $\beta_t := c/\sqrt{t} \in (0,1]$ for $t \ge t_0$ (as $c \le \sqrt{t_0} \le \sqrt{t}$), so that $C'/\sqrt{t} = (C'/c)\beta_t$ and (9) reads $w_{t+1} \le (1-\beta_t)w_t + (C'/c)\beta_t$. Subtracting C'/c from both sides,

$$w_{t+1} - \frac{C'}{c} \ \le\ (1-\beta_t)\Big(w_t - \frac{C'}{c}\Big), \qquad t_0 \le t < T, \tag{11}$$

an inequality that holds for any sign of $w_t - C'/c$. Since every factor $1 - \beta_t$ lies in $[0,1)$, iterating (11) from t_0 to T gives $w_T - C'/c \le P_{t_0,T}(w_{t_0} - C'/c)$, i.e. the first bound of (10). For the

second, $1 - P_{t_0,T} \le 1$ (with $C' \ge 0$), and, using $1 - x \le e^{-x}$ and, since $x \mapsto x^{-1/2}$ is decreasing, $\sum_{t=t_0}^{T-1} t^{-1/2} \ge \int_{t_0}^{T} x^{-1/2}\,dx = 2(\sqrt{T} - \sqrt{t_0})$,

$$P_{t_0,T} = \prod_{t=t_0}^{T-1}\Big(1 - \frac{c}{\sqrt{t}}\Big) \;\le\; e^{-2c(\sqrt{T}-\sqrt{t_0})}. \tag{12}$$

□

Remark A.2 (Why the product form is needed). An inductive (constant-propagated) proof of a bound $w_T \le K$ from (9) must close from the burn-in t_0 and can only propagate $K \ge w_{t_0}$; with $w_{t_0} \lesssim \sqrt{t_0}$ (the weighted function at the burn-in, Step 2 below) and $t_0 = t_0(T) = \Theta(\log T)$ this would inject a $\sqrt{\log T}$ factor into the rate. The explicit product form keeps the forcing constant C'/c free of t_0 and confines the burn-in value to the exponentially decaying homogeneous term, which is scheduled away in (56). Positivity of c is also needed: in the application $c = m_w/s_2 > 0$ because (A-push) gives $m_w \ge 1.37\tilde{\mu} > 0$ (Lemma A.9(V2)). The bound (10) is the exact fixed-point form of the recursion: C'/c is the stationary value of $w_{t+1} = (1 - \beta_t)w_t + (C'/c)\beta_t$, and $P_{t_0,T}$ is its exact homogeneous propagator.

Lemma A.3 (Supply: regret-norm growth)**.** *Assume (M), (A-var), (A-int), (A-push) and (A-ε''), with $\theta_i^0 = 0$. Let*

$$\tilde{\mu} := \mu(1-\gamma), \qquad c_\pi := \frac{\tilde{\mu}\,\delta_0}{4}, \tag{13}$$

$$c_Y := \min\Big(\frac{\sigma_{\min}^2\,\delta_0^2}{4},\; c_\pi^2,\; \frac{15\,\sigma_{\min}^2}{64}\Big) = \min\Big(\frac{\sigma_{\min}^2\,\delta_0^2}{4},\; c_\pi^2\Big), \tag{14}$$

$$s_1 := \sqrt{\frac{c_Y}{2}}, \qquad s_2 := \sqrt{3}\,G_D, \qquad c_6 := \frac{c_Y^2}{32\,G_D^4}, \tag{15}$$

where (A-ε'') is used through (4)*. Then for every player $i = 1, 2$ and every $t \ge 1$:*

(i) (deterministic upper bound.) *$Y_{i,t} \le \frac{3}{2}\,G_D^2\,t$, hence $S_{i,t} \le \sqrt{2Y_{i,t}} \le s_2\sqrt{t}$;*

(ii) (probabilistic lower bound, two-player union.)

$$\Pr\Big(\exists\, t \ge t_0,\; i \in \{1,2\} :\; S_{i,t} < s_1\sqrt{t}\Big) \;\le\; \frac{2\,e^{-c_6 t_0}}{1 - e^{-c_6}}, \tag{16}$$

which is independent of any horizon T (monotone limit of the infinite-window event). The factor 2 *in* (16) *is the two-player union bound.*

Proof. Fix player i and drop the subscript where unambiguous. Let $\mathcal{H}_t^{(1)} = \mathcal{F}_t$ and $\mathcal{H}_t^{(2)} = \mathcal{F}_{t+1/2}$ (§A.2): player 1 samples before moving, player 2 after player 1's update. In the layer bounds below the opponent coordinate enters only through the coupling bound $|\tilde{\kappa} e_{-i}| \le |\tilde{\kappa}|$, so the half-step timing shift for player 2 is immaterial. Write $\mathcal{H}_t := \mathcal{H}_t^{(i)}$, $S := S_{i,t}$, $Y := \|\theta\|_2^2$, and let $z := z_i^t$ denote the *played* base coordinate: on a non-fallback step ($S >$ EPS) $z = \theta_1/S$, so $\theta = (Sz,\, S(1-z))$; on a fallback step ($S \le$ EPS, including $S = 0$) the algorithm plays $z = \frac{1}{2}$ irrespective of θ, and only Case 5.1 below concerns such steps. The RM$^+$ increment on a step is $g = \tilde{D}\,(1 - z,\, -z)$ with $\tilde{D} := D - \tilde{\mu}(2z - 1) + \mu\rho_i$ and $|D| \le 2(U + \tau_d)$, hence $|\tilde{D}| \le G_D := 2U + 2\tau_d + 2\mu$; $\omega := z^2 + (1-z)^2 \in [\frac{1}{2}, 1]$ satisfies $\|g\|_2^2 = \tilde{D}^2\omega$.

Step 1 (orthogonality and inner-product identities). The RM$^+$ update is $g = \tilde{u} - \langle \tilde{u}, x_b\rangle \mathbf{1}$ with $x_b = \theta/S$ on a non-fallback step. Since $\langle \mathbf{1}, \theta\rangle = S$, the *orthogonality identity* holds

$$\langle g, \theta\rangle = \langle \tilde{u}, \theta\rangle - \langle \tilde{u}, x_b\rangle\, S = S\big(\langle \tilde{u}, x_b\rangle - \langle \tilde{u}, x_b\rangle\big) = 0. \tag{17}$$

More generally, for every $y \in \Delta$ we have $\langle \mathbf{1}, \theta - Sy\rangle = S - S = 0$, so the *inner-product identity*

$$\langle g, \theta - Sy\rangle = S\,\langle \tilde{u},\, x_b - y\rangle \tag{18}$$

links the θ-space update to the x-space field at scale S. Identity (17) gives $\langle g, \theta\rangle = 0$ on every non-fallback step; on a fallback step $x_b = \mathbf{1}/2 \ne \theta/S$ and Cauchy–Schwarz replaces it by $|\langle g, \theta\rangle| \le \|g\|_2\,\|\theta\|_2 \le G_D\,\mathrm{EPS}$ ($\|\theta\|_2 \le \|\theta\|_1 = S \le$ EPS).

Step 2 (energy inequality). Projection is non-expansive from the origin, $\|[\theta+g]_+\|_2^2 \le \|\theta+g\|_2^2$, so $Y_{t+1} \le Y_t + 2\langle g,\theta\rangle + \|g\|_2^2$. On a non-fallback step (17) and $\|g\|_2^2 \le G_D^2$ give $Y_{t+1} \le Y_t + G_D^2$; on a fallback step the Step 1 inner-product bound gives $Y_{t+1} \le Y_t + G_D^2 + 2G_D\,\mathrm{EPS}$. With $\mathrm{EPS} \le G_D/4$ ((4)), every step obeys

$$Y_{t+1} \;\le\; Y_t + \tfrac{3}{2}\,G_D^2, \tag{19}$$

and $\theta^0 = 0$ sums (19) to $Y_t \le \frac{3}{2}\,G_D^2\,t$. Since $\|\theta\|_1 \le \sqrt{2}\,\|\theta\|_2$,

$$S_{i,t} \;\le\; \sqrt{2Y_{i,t}} \;\le\; \sqrt{3}\,G_D\,\sqrt{t} \;=\; s_2\sqrt{t}, \tag{20}$$

deterministically for every path. The cross term in $\|\theta+g\|_2^2$ is killed by (17), so the energy accumulates only the squared increments $\|g\|_2^2$; this is what yields the $\sqrt{t}$ scaling (a naive per-step bound on S alone would only give $S_t = O(tG_D)$, and the constant-increment example $A = \left(\begin{smallmatrix} M & M+\kappa \\ 0 & 0 \end{smallmatrix}\right)$ below shows the cancellation is essential).

Step 3 (at most one clipped component; lossless clipping). The components of $g = \tilde{D}(1-z,-z)$ have opposite signs or one is zero, and a component with $g_j \ge 0$ cannot clip since $\theta_j \ge 0$; hence at most one component satisfies $\theta_j + g_j < 0$. On a non-fallback step, (17) gives the exact expansion $\Delta := \|\theta^{t+1}\|_2^2 - Y_t = \|g\|_2^2 - (\theta_j+g_j)^2$ if component j clips (single term by the preceding line), and $\Delta = \|g\|_2^2 = \tilde{D}^2\omega$ if neither clips. Expanding the clip term,

$$\Delta = \|g\|_2^2 - \big(g_j^2 + 2\theta_j g_j + \theta_j^2\big) = g_{3-j}^2 - 2\theta_j g_j - \theta_j^2 \;\ge\; g_{3-j}^2 + \theta_j^2 \;\ge\; 0, \tag{21}$$

where the inequality uses the clip condition $g_j < -\theta_j \le 0$, i.e. $-2\theta_j g_j \ge 2\theta_j^2$. Clipping is therefore pointwise lossless: Y_t never decreases on a non-fallback step, and a clipped step's gain is reduced from $\|g\|_2^2$ to at least $g_{3-j}^2 + \theta_j^2$.

Step 4 (layer drift). Let $f(z) := \mathbb{E}[\tilde{D} \mid \mathcal{H}_t^{(i)}]$, the conditional mean of the sampled increment; with $e := z - z^\star$ one has the field $-2\tilde{\mu}e_i + \sigma_i\tilde{\kappa}e_{-i}$ for player i ($\sigma_i = +1$ for player 1, $\sigma_i = -1$ for player 2), where $\tilde{\kappa} := (1-\gamma)\kappa_A$ and $|e_{-i}| \le 1$. For player 1 this is the field at (z_1^t, z_2^t); for player 2, who samples against her already-updated opponent, the coupling is at $e_1(z_1^{t+1})$ instead. Only the own-coordinate term $-2\tilde{\mu}e_i$ and the coupling bound $|\tilde{\kappa}e_{-i}| \le |\tilde{\kappa}|$ enter the layer bounds below, so the timing shift cancels. Fix player 1 in the *upper layer* $z \ge 1 - \delta_0/2$. By (A-int), $z_1^\star \le 1-\delta_0$, so $e_1 \ge (1-\delta_0/2) - (1-\delta_0) = \delta_0/2$, and by (A-push), $|\kappa_A| \le \mu\delta_0/2$, hence

$$f_1 \;\le\; -2\tilde{\mu}\cdot\frac{\delta_0}{2} + (1-\gamma)|\kappa_A| \;\le\; -\tilde{\mu}\delta_0 + \frac{\tilde{\mu}\delta_0}{2} \;=\; -\frac{\tilde{\mu}\delta_0}{2} \;=\; -2c_\pi. \tag{22}$$

Since $\theta_2 = S(1-z)$ with $z \ge 1 - \delta_0/2 \ge \frac{1}{2}$,

$$\mathbb{E}[g_{i,2} \mid \mathcal{H}_t^{(i)}] = -z\,\mathbb{E}[\tilde{D}_i \mid \mathcal{H}_t^{(i)}] \;\ge\; \big(1-\tfrac{\delta_0}{2}\big)\cdot 2c_\pi \;\ge\; c_\pi, \qquad \mathbb{E}[\tilde{D}_i \mid \mathcal{H}_t^{(i)}] = f_1 \;\le\; -2c_\pi \;<\; 0. \tag{23}$$

The *lower layer* $z \le \delta_0/2$ is symmetric: $e_1 \le \delta_0/2 - \delta_0 = -\delta_0/2$, so $f_1 \ge \tilde{\mu}\delta_0 - |\tilde{\kappa}| \ge 2c_\pi$ and $\mathbb{E}[g_{i,1} \mid \mathcal{H}] = (1-z)f_1 \ge c_\pi$; player 2 obeys the same bounds in its own coordinate. We shall use these as: in the upper (lower) layer, $\mathbb{E}[\tilde{D}_i \mid \mathcal{H}] \le -2c_\pi$ ($\ge 2c_\pi$).

Step 5 (uniform expected gain). We prove

$$\mathbb{E}[\Delta Y_i^t \mid \mathcal{H}_t^{(i)}] \;\ge\; c_Y \tag{24}$$

for *every* reachable state on *every* step, clipping or fallback included, with no condition on S. The four cases partition the state space: fallback, then the non-fallback states split by the layer of z.

Case 5.1 (fallback, $S \le \mathrm{EPS}$). On a fallback step the strategy is $x = \mathbf{1}/2$, so the regret increment is $g = \tilde{D}(\frac{1}{2}, -\frac{1}{2})$, whose two components have opposite signs. Since $\theta \ge 0$, the clip cannot shrink the favourable component, so after the update *at least one* component of θ^{t+1} is $\ge \frac{|\tilde{D}|}{2}$, hence $Y' \ge \tilde{D}^2/4$ pointwise, while $Y \le S^2 \le \mathrm{EPS}^2$. By (A-var) $\mathbb{E}[\tilde{D}^2 \mid \mathcal{H}] \ge \sigma_{\min}^2$ (the shift $\tilde{D} - D = \mu\rho_i$ is deterministic on a fallback step, so $\mathbb{E}[\tilde{D}^2 \mid \mathcal{H}] \ge \mathrm{Var}(D \mid \mathcal{H})$), and $\mathrm{EPS} \le \sigma_{\min}/8$ ((4)) gives

$$\mathbb{E}[\Delta \mid \mathcal{H}] \;\ge\; \frac{\mathbb{E}[\tilde{D}^2 \mid \mathcal{H}]}{4} - \mathrm{EPS}^2 \;\ge\; \frac{\sigma_{\min}^2}{4} - \frac{\sigma_{\min}^2}{64} \;=\; \frac{15\,\sigma_{\min}^2}{64}. \tag{25}$$

This is the third term of c_Y; since $\delta_0 < \frac{1}{2}$ it never binds ($\delta_0^2/4 < \frac{1}{16} < \frac{15}{64}$), which is why (14) reduces to a two-term minimum.

Case 5.2 (upper layer $z \geq 1 - \delta_0/2$, non-fallback). If $\tilde{D} \leq 0$, then $g_2 = -\tilde{D}z \geq 0$ and component 2 cannot clip; component 1 clips only if $|\tilde{D}|(1-z) > \theta_1 = Sz$. When $S \geq G_D$ this is impossible, since $|\tilde{D}|(1-z) \leq G_D \cdot \frac{\delta_0}{2} < \frac{G_D}{2} \leq \frac{S}{2} \leq Sz$, so no clip occurs and $\Delta = \tilde{D}^2\omega \geq \frac{\tilde{D}^2}{2}$; when $S < G_D$ and component 1 clips, (21) gives $\Delta \geq g_2^2 = \tilde{D}^2 z^2 \geq \frac{\tilde{D}^2}{4}$. If $\tilde{D} > 0$, (21) gives $\Delta \geq 0$. Hence $\Delta \geq \mathbb{1}\{\tilde{D} \leq 0\}\, \frac{\tilde{D}^2}{4}$ pointwise. Writing $\tilde{D}^- := \max(-\tilde{D}, 0) \in [0, G_D]$, (23) gives $\mathbb{E}[\tilde{D}^- \mid \mathcal{H}] \geq -\mathbb{E}[\tilde{D} \mid \mathcal{H}] \geq 2c_\pi$ ($\mathbb{E}[\tilde{D}^+ \mid \mathcal{H}] \geq 0$), so conditional Jensen yields $\mathbb{E}[(\tilde{D}^-)^2 \mid \mathcal{H}] \geq (\mathbb{E}[\tilde{D}^- \mid \mathcal{H}])^2 \geq 4c_\pi^2$, and

$$\mathbb{E}[\Delta \mid \mathcal{H}] \;\geq\; \frac{\mathbb{E}[(\tilde{D}^-)^2 \mid \mathcal{H}]}{4} \;\geq\; c_\pi^2. \tag{26}$$

Case 5.3 (lower layer $z \leq \delta_0/2$, non-fallback). Mirror: $\tilde{D} \geq 0 \Rightarrow g_1 \geq 0$ (component 1 cannot clip); component 2 clips only if $\tilde{D}z > S(1-z) \geq S/2$ with $\tilde{D}z \leq G_D\delta_0/2$, impossible for $S \geq G_D$, while for $S < G_D$ (21) gives $\Delta \geq g_1^2 = \tilde{D}^2(1-z)^2 \geq \frac{\tilde{D}^2}{4}$. So $\Delta \geq \mathbb{1}\{\tilde{D} \geq 0\}\, \frac{\tilde{D}^2}{4}$ pointwise, and $\mathbb{E}[\tilde{D} \mid \mathcal{H}] \geq 2c_\pi$ gives $\mathbb{E}[\Delta \mid \mathcal{H}] \geq c_\pi^2$ by the same Jensen argument.

Case 5.4 (interior $\delta_0/2 < z < 1 - \delta_0/2$, non-fallback). No clip: $\Delta = \tilde{D}^2\omega \geq \frac{\tilde{D}^2}{2}$; clip at 2: (21) gives $\Delta \geq g_1^2 = \tilde{D}^2(1-z)^2 \geq \frac{\tilde{D}^2\delta_0^2}{4}$; clip at 1: $\Delta \geq g_2^2 = \tilde{D}^2 z^2 \geq \frac{\tilde{D}^2\delta_0^2}{4}$. Since $\delta_0 \leq 1$, $\Delta \geq \frac{\tilde{D}^2\delta_0^2}{4}$ in all three subcases, and (A-var) $\mathbb{E}[\tilde{D}^2 \mid \mathcal{H}] \geq \sigma_{\min}^2$ gives

$$\mathbb{E}[\Delta \mid \mathcal{H}] \;\geq\; \frac{\delta_0^2}{4}\, \mathbb{E}[\tilde{D}^2 \mid \mathcal{H}] \;\geq\; \frac{\sigma_{\min}^2\, \delta_0^2}{4}. \tag{27}$$

Minimizing (25), (26) and (27) over the four cases yields (24) with c_Y as in (14).

Step 6 (Azuma closure). Set $M_t := \sum_{s<t}\big(\Delta_s - \mathbb{E}[\Delta_s \mid \mathcal{H}_s]\big)$, a martingale on the (player-wise) filtration (not to be confused with the truncated Lyapunov variable W_t of (8)). Increment size: a non-fallback step has $\Delta \in [0, G_D^2]$ by (21) and Step 2; a fallback step has $\Delta \in [-\mathrm{EPS}^2,\ G_D^2 + 2G_D\,\mathrm{EPS}]$ (Step 2), so $|\Delta_s - \mathbb{E}[\Delta_s \mid \mathcal{H}_s]| \leq 2G_D^2$ for every s ($\mathrm{EPS} \leq G_D/4$). Since $Y_t = \sum_{s<t}\Delta_s \geq c_Y\, t + M_t$ by (24), Azuma–Hoeffding gives

$$\Pr\Big(M_t \leq -\tfrac{c_Y}{2}\, t\Big) \;\leq\; \exp\Big(-\frac{(c_Y t/2)^2}{2t\,(2G_D^2)^2}\Big) \;=\; e^{-c_6 t}, \tag{28}$$

and on the complement $Y_t \geq \frac{c_Y}{2}\, t$, so $S_{i,t} \geq \|\theta^t\|_2 = \sqrt{Y_t} \geq s_1\sqrt{t}$ with $s_1 = \sqrt{c_Y/2}$. Taking the union over $t \geq t_0$ and over the two players (this is precisely the factor 2 in (16)),

$$\Pr\Big(\exists\, t \geq t_0,\ i \in \{1,2\} :\ S_{i,t} < s_1\sqrt{t}\Big) \;\leq\; 2\sum_{t \geq t_0} e^{-c_6 t} \;\leq\; \frac{2\, e^{-c_6 t_0}}{1 - e^{-c_6}}, \tag{29}$$

the infinite-window event following by the monotone limit $T \to \infty$. This proves (ii).

Remark (necessity of (A-int)). (A-var) alone cannot yield any lower bound. For the payoff $A = \left(\begin{smallmatrix} M & M+\kappa \\ 0 & 0 \end{smallmatrix}\right)$ with the uniform anchor, $\gamma > 0$, $\kappa \neq 0$ and $\min(M, M+\kappa) > \max(\mu(1-\gamma), 2\,\mathrm{EPS}) + 2\tau_d$, one has $\kappa_A = -\kappa \neq 0$, so $\sigma_{\min}^2 > 0$ and (A-var) holds. Player 1's un-transformed payoff difference is M or $M + \kappa$, both exceeding $\tilde{\mu} + 2\tau_d$. Starting from $\theta^0 = 0$ the state freezes: the fallback step reaches $\theta^1 = (\tilde{D}_1^0/2, 0)$ with $\tilde{D}_1^0 \geq \min(M, M+\kappa) - 2\tau_d > 2\,\mathrm{EPS}$, so $S_{1,1} > \mathrm{EPS}$ and the fallback is left at once; thereafter $z = 1$, $\tilde{D} = D - \tilde{\mu} \geq \min(M, M+\kappa) - 2\tau_d - \tilde{\mu} > 0$ a.s., $g = (0, -\tilde{D})$ leaves component 1 unchanged and clips component 2 every step, so $\theta^{t+1} = \theta^t$ and $S_{1,t}$ is constant in t: no $s_1 > 0$ exists. The excluded hypothesis is interiority. Suppose the root $z^\star$ of f lay in $(0,1)^2$; then the opponent's executed strategy at $z^\star$ is in the executed box, $\hat{x}_{2,2}^\star \in [\gamma/2, 1 - \gamma/2]$, and the first fixed-point equation $2\tilde{\mu}z_1^\star = M + \tilde{\mu} + \kappa\,\hat{x}_{2,2}^\star$ gives $2\tilde{\mu}z_1^\star \geq \min(M, M+\kappa) + \tilde{\mu} > 2\tilde{\mu}$, i.e. $z_1^\star > 1$, a contradiction. Hence $z^\star \notin (0,1)^2$: the example is in the class excluded by (A-int). (The unrestricted affine root may leave the box through either coordinate; only interiority is being refuted.) □

Lemma A.4 (Projection domination)**.** *Let (M) hold and* EPS $\le G_D/4$ *(the first bound of (A-ε'')). Fix an update of player i that is* not *a fallback step and satisfies $S := S_{i,t} \ge 2G_D$. Let $\theta_j = \theta_{i,j} \ge 0$ $(j = 1, 2)$ be the two RM$^+$ component regrets, $\theta_1+\theta_2 = S$, so that the base coordinate is $z = z_i^t = \theta_1/S \in [0, 1]$; let $\tilde{D}_i := D_i - \tilde{\mu}(2z - 1) + \mu\rho_i$ be the* realized *transformed payoff difference of the step, $D_i = u_{i,1}-u_{i,2}$ the sampled un-transformed difference with $|D_i| \le 2(U+\tau_d)$, so $|\tilde{D}_i| \le G_D$; write $g = (g_1, g_2) = (\tilde{D}_i(1-z), -\tilde{D}_i z)$ for the pre-clip regret increment. The* virtual *(unprojected) iterate is*

$$z_v^{t+1} \;:=\; \frac{\theta_1 + g_1}{\theta_1 + g_1 + \theta_2 + g_2} \;=\; \frac{Sz + \tilde{D}_i(1 - z)}{S + \tilde{D}_i(1 - 2z)}. \tag{30}$$

Then, pointwise on the noise realization,

$$z_i^{t+1} = \mathrm{proj}_{[0,1]}\big(z_v^{t+1}\big), \tag{31}$$

with clipping at 1 iff $z_v^{t+1} > 1$, clipping at 0 iff $z_v^{t+1} < 0$, and equality otherwise. Moreover, for any $\Phi : \mathbb{R} \to \mathbb{R}$ that is single-valley at $z_i^\star \in [0, 1]$, i.e. nonincreasing on $(-\infty, z_i^\star]$, nondecreasing on $[z_i^\star, \infty)$, one has pointwise

$$\Phi\big(z_i^{t+1}\big) \;\le\; \Phi\big(z_v^{t+1}\big), \tag{32}$$

with equality when no clipping occurs. The Lyapunov components of (5) *are of this type: for $\varphi(v) = v^2 + (1 - v)^2 \ge \frac{1}{2} > 0$ on all of $\mathbb{R}$, the potential $\Phi_i(v) = \int_{z_i^\star}^{v} \frac{s-z_i^\star}{\varphi(s)}\, ds$ satisfies $\Phi_i'(v) = (v - z_i^\star)/\varphi(v)$, hence is single-valley at $z_i^\star$; applying* (32) *coordinatewise to the separable $V(z) = \Phi_1(z_1) + \Phi_2(z_2)$ gives $V(z^{t+1}) \le V(z_v^{t+1})$ pointwise, whether or not clipping occurs.*

Proof. **(S0.1) Normalization identities.** The virtual value of $\theta + g$ is $(\theta_1 + g_1)/(S + g_1 + g_2)$; its numerator is $\theta_1 + g_1 = Sz + \tilde{D}_i(1-z)$ and its denominator is $S + g_1 + g_2 = S + \tilde{D}_i(1-z) - \tilde{D}_i z = S + \tilde{D}_i(1-2z)$, giving (30). Directly subtracting and using $\theta_2 = S(1-z)$, $\theta_2 + g_2 = S(1-z) - \tilde{D}_i z$,

$$1 - z_v^{t+1} = \frac{\theta_2 + g_2}{S + \tilde{D}_i(1 - 2z)}, \qquad z_v^{t+1} = \frac{\theta_1 + g_1}{S + \tilde{D}_i(1 - 2z)}. \tag{33}$$

(S0.2) Denominator positivity. Since $z \in [0, 1]$ and $|\tilde{D}_i| \le G_D$,

$$S + \tilde{D}_i(1 - 2z) \;\ge\; S - |\tilde{D}_i|\,|1 - 2z| \;\ge\; S - G_D \;\ge\; \tfrac{S}{2} \;>\; 0, \tag{34}$$

the last inequality using $S \ge 2G_D$. If in addition the update clips at 2, then $\theta_2 + g_2 < 0$, i.e. $S(1 - z) < \tilde{D}_i z$ (so $\tilde{D}_i > 0$); from $S(1 - z) < \tilde{D}_i z \le G_D z$ it follows that $z > \frac{1}{2}$ (otherwise $z \le \frac{1}{2}$ forces $S < G_D$, contradicting $S \ge 2G_D$), whence $\theta_1 = Sz > G_D \ge |\tilde{D}_i| \ge |g_1|$ and the first component stays positive. Clipping at 1 is symmetric ($\tilde{D}_i < 0$, $z < \frac{1}{2}$, second component stays positive).

(S0.3) Clipping criterion. By (33) and (34), clipping at 2 (component 2 of $\theta + g$ is negative) is equivalent to $\theta_2 + g_2 < 0$, equivalently $z_v^{t+1} > 1$; clipping at 1 iff $\theta_1 + g_1 < 0$ iff $z_v^{t+1} < 0$. When one component clips, the other remains positive by (S0.2), so the post-clip normalized value is 1 (resp. 0); when neither clips, it equals $z_v^{t+1} \in [0, 1]$. The post-update step is again a non-fallback step, so this normalized value is the coordinate actually played at $t + 1$: the energy never decreases on a non-fallback step ((21) in Lemma A.3, Step 3), hence $S_{i,t+1} \ge \|\theta^{t+1}\|_2 \ge \|\theta^t\|_2 \ge S/\sqrt{2} \ge \sqrt{2}\, G_D > G_D/4 \ge$ EPS. This is exactly the projection (31).

(S0.4) Overshoot magnitude. On a clip-at-2 step, $|\theta_2 + g_2| \le |g_2|$ (as $\theta_2 \ge 0$), so by (34)

$$0 < z_v^{t+1} - 1 = \frac{|\theta_2 + g_2|}{S + \tilde{D}_i(1 - 2z)} \;\le\; \frac{|g_2|}{S - G_D} \;\le\; \frac{G_D}{S/2} = 2G_D a_{i,t}, \tag{35}$$

and symmetrically $0 < -z_v^{t+1} \le 2G_D a_{i,t}$ on clip-at-1 steps; together with the no-clip bound $|z_v^{t+1} - z| \le 2G_D a_{i,t}$, the virtual iterate stays within $2G_D a_{i,t}$ of the box, which is the extent used in the extended-domain Taylor bound (Lemma A.12).

Domination (32)**.** If no clipping occurs, $z_i^{t+1} = z_v^{t+1}$ and (32) is an equality. If the update clips at 2, then $z_v^{t+1} > 1 \ge z_i^\star$ and $1 = z_i^{t+1} \le z_v^{t+1}$: both points lie in $[z_i^\star, \infty)$, where Φ is nondecreasing,

so $\Phi(z_i^{t+1}) \le \Phi(z_v^{t+1})$. If it clips at 1, then $z_v^{t+1} < 0 \le z_i^\star$ and $z_i^{t+1} = 0 > z_v^{t+1}$: both points lie in $(-\infty, z_i^\star]$, where Φ is nonincreasing, so again $\Phi(z_i^{t+1}) \le \Phi(z_v^{t+1})$. The claim for $V = \Phi_1 + \Phi_2$ follows because V is separable (diagonal Hessian): each coordinate is updated independently and dominates its own virtual value, so the two coordinates can be bounded separately and summed. □

Remark A.5 (Discharge of the step hypothesis; scope). **(i) Discharge.** Lemma A.4 is a statement *conditional* on the step being non-fallback with $S_{i,t} \ge 2G_D$; this must be discharged on every step counted by Lemma A.13. On $\{\varrho > t\}$ with $t \ge t_0$ and $t_0 \ge (4G_D/s_1)^2$ one has $S_{i,t} \ge s_1\sqrt{t} \ge s_1\sqrt{t_0} \ge 4G_D > 2G_D$, and $S_{i,t} \ge 4G_D > G_D/4 \ge$ EPS by (A-ε''), so no fallback occurs. Hence the projection lemma applies on every step on which (49)–(50) of Lemma A.13 run, and the inequality chain $\Phi_i(z_i^{t+1}) \le \Phi_i(z_{v,i}^{t+1})$ (no-clip or clipped) holds unconditionally: no clipping indicator enters any conditional expectation. **(ii)** $n = 2$ **specificity.** The mechanism is "clipping one of two components moves the normalized point onto a simplex vertex, and the virtual point lies beyond that vertex"; for $n \ge 3$ actions clipping a component does not land on a vertex and the projection argument does not automatically extend (§A.5). **(iii) Single-valley suffices.** Only the single-valley property is used, not convexity; indeed Φ_i'' can be negative on the extended domain $\mathcal{Z}^+$.

The drift-decomposition engine behind Lemma A.13 is assembled from the following facts, in the setting of §A.1 and notation of §A.2: $z^t = (z_1^t, z_2^t)$ with $z_i^t = x_{\mathrm{b},i,1}^t \in [0,1]$, executed policies $\hat{x}_{i,1} = (1-\gamma)z_i + \gamma/2$, $\hat{x}_{i,2} = (1-\gamma)(1-z_i) + \gamma/2$, and $\bar{a}_t = (a_{1,t} + a_{2,t})/2$.

Lemma A.6 (Exact one-step formula)**.** *Consider a* non-fallback *step of player i with $S_{i,t} \ge 2G_D$ (so the denominator below is positive, Lemma A.4). With regularized utility $\tilde{u}_i^t = u_i^t - \mu(\hat{x}_i^t - r_i)$ and realized difference $\tilde{D}_i^t = \tilde{u}_{i,1}^t - \tilde{u}_{i,2}^t$, the RM$^+$ increment is $g_i^t = \tilde{u}_i^t - \langle \tilde{u}_i^t, x_{\mathrm{b},i}^t \rangle \mathbf{1} = \tilde{D}_i^t(1 - z_i^t, -z_i^t)$, and since $\hat{x}_{i,1}^t - \hat{x}_{i,2}^t = (1-\gamma)(2z_i^t - 1)$ and $r_{i,1} - r_{i,2} = \rho_i$,*

$$\tilde{D}_i^t = D_i^t - \tilde{\mu}(2z_i^t - 1) + \mu\rho_i, \qquad D_i^t = u_{i,1}^t - u_{i,2}^t, \tag{36}$$

The lemma concerns the virtual *(unprojected) update $z_{v,i}^{t+1}$ of Lemma A.4 (under (A-ε'') the realized update is $z_i^{t+1} = \mathrm{proj}_{[0,1]}(z_{v,i}^{t+1})$, Lemma A.4; the identities below are purely virtual and need only $S_{i,t} \ge 2G_D$). Its exact identity and expansion are*

$$\Delta z_i^v := \; z_{v,i}^{t+1} - z_i^t = \frac{\tilde{D}_i^t\, \varphi(z_i^t)}{S_{i,t} + \tilde{D}_i^t(1 - 2z_i^t)} = a_{i,t}\tilde{D}_i^t \varphi(z_i^t) + r_i^t, \tag{37}$$

with $|r_i^t| \le 2G_D^2 a_{i,t}^2$ and $|\Delta z_i^v| \le 2G_D\, a_{i,t}$, where $z_i^t \in [0,1]$; the identity holds pointwise whether or not the realized step clips.

Proof. With $g_{i,1}^t = (1 - z_i^t)\tilde{D}_i^t$, $g_{i,2}^t = -z_i^t\tilde{D}_i^t$, the normalized first coordinate is $(S_{i,t}z_i^t + \tilde{D}_i^t(1 - z_i^t))/(S_{i,t} + \tilde{D}_i^t(1 - 2z_i^t))$; subtracting z_i^t times the denominator leaves $\tilde{D}_i^t(1 - 2z_i^t + 2(z_i^t)^2) = \tilde{D}_i^t\varphi(z_i^t)$. For the expansion use $|\tilde{D}_i^t| \le G_D$, $|\varphi| \le 1$ and the denominator bound $S_{i,t} - G_D \ge S_{i,t}/2$, with $|1/(1+x) - 1| \le 2|x|$ for $|x| \le \frac{1}{2}$, $x = \tilde{D}_i^t(1 - 2z_i^t)/S_{i,t}$; the amplitude bound is the same denominator bound. □

Lemma A.7 (Closed form of the drift field)**.** *Let $c_j = A_{1j} - A_{2j}$ ($j = 1, 2$), $d_1 = A_{12} - A_{11}$, $d_2 = A_{22} - A_{21}$, $\kappa_A = c_1 - c_2 = A_{11} - A_{12} - A_{21} + A_{22}$. With $f_i(z) := \mathbb{E}[\tilde{D}_i \mid \mathcal{F}]$ evaluated against the executed opponent policy at z, and $b = \binom{c_2 + \kappa_A\gamma/2 + \tilde{\mu} + \mu\rho_1}{d_2 - \kappa_A\gamma/2 + \tilde{\mu} + \mu\rho_2}$,*

$$f(z) = b + Mz, \quad M = \begin{pmatrix} -2\tilde{\mu} & \tilde{\kappa} \\ -\tilde{\kappa} & -2\tilde{\mu} \end{pmatrix}, \tag{38}$$

so that by (37)*, on a non-fallback step with $S_{i,t} \ge 2G_D$ (where the virtual increments are defined), the expected virtual increment of player 1 is $\mathbb{E}[\Delta z_1^v \mid \mathcal{F}_t] = a_{1,t}\varphi(z_1^t)f_1(z^t) + O(a_{1,t}^2)$, while for player 2, who moves after player 1's round-t update, the conditional mean at her pre-update filtration is $\mathbb{E}[\Delta z_2^v \mid \mathcal{F}_{t+1/2}] = a_{2,t}\varphi(z_2^t)f_2(z_1^{t+1}, z_2^t) + O(a_{2,t}^2)$ (Lemma A.11 converts the argument back to z^t). Moreover $|f_i| \le F_\infty = 2U + 2\mu$, and the centered sampling noise $\tilde{D}_i - f_i$ is bounded by $2G_D$. The field is affine with $\partial_{z_1} f_2 = -\tilde{\kappa}$ and $\partial_{z_2} f_1 = \tilde{\kappa}$.*

Proof. Player 1 samples her payoff difference against the executed policy of player 2, $p = \hat{x}_{2,1} = (1-\gamma)z_2 + \gamma/2$: $\mathbb{E}[D_1 \mid \mathcal{F}] = c_2 + \kappa_A p$ (dither mean zero), and the regularizer contributes $-\tilde{\mu}(2z_1-1)+\mu\rho_1$ ((36)), so $f_1(z) = b_1 - 2\tilde{\mu}z_1 + \tilde{\kappa}z_2$. Player 2 receives $-A$: $\mathbb{E}[D_2 \mid \cdot] = d_2 - \kappa_A q$, $q = \hat{x}_{1,1}$, giving $f_2(z) = b_2 - \tilde{\kappa}z_1 - 2\tilde{\mu}z_2$, i.e. (38). The drift claims follow by conditioning (37) at each player's pre-update filtration, against the opponent's executed strategy at that time. For the bounds, $|\mathbb{E}[D_i \mid \cdot]| \le 2U$ (a difference of two entries of A, the dither being centered), $|\tilde{\mu}(2z_i - 1)| \le \mu$ and $|\mu\rho_i| \le \mu$ give $|f_i| \le 2U + 2\mu = F_\infty$, and $|\tilde{D}_i| \le G_D$, $|f_i| \le F_\infty \le G_D$ give $|\tilde{D}_i - f_i| \le 2G_D$. □

Lemma A.8 (Fixed point and explicit interiority)**.** *Under (M) one has $\tilde{\mu} = \mu(1-\gamma) > 0$, so $\Delta := \det M = 4\tilde{\mu}^2 + \tilde{\kappa}^2 > 0$ (whether or not $\tilde{\kappa} = 0$), and the unique root of f is*

$$z_1^\star = \frac{2\tilde{\mu}b_1 + \tilde{\kappa}b_2}{\Delta}, \qquad z_2^\star = \frac{-\tilde{\kappa}b_1 + 2\tilde{\mu}b_2}{\Delta}; \tag{39}$$

whenever $z^\star \in (0,1)^2$ it is the base-coordinate representation of the perturbed equilibrium $\hat{x}^\star(\gamma, \mu, r)$. The fixed point sits next to the anchor: with $\bar{z}_i = (r_{i,1} - \gamma/2)/(1-\gamma)$ the base coordinate of the anchor,

$$\|z^\star - \bar{z}\|_2 \;\le\; \frac{\sqrt{2}\,U}{\tilde{\mu}}, \tag{40}$$

so a sufficient explicit condition for $z^\star \in [\delta_0, 1-\delta_0]^2$ is

$$\bar{z} \in \Big[\delta_0 + \frac{\sqrt{2}U}{\tilde{\mu}},\; 1 - \delta_0 - \frac{\sqrt{2}U}{\tilde{\mu}}\Big]^2, \tag{41}$$

i.e. the anchor lies inside the margin by $\sqrt{2}U/\tilde{\mu}$. For the uniform anchor $\bar{z} = (\frac{1}{2}, \frac{1}{2})$ this reads $\tilde{\mu} \ge \sqrt{2}U/(\frac{1}{2} - \delta_0)$.

Proof. $z^\star = -M^{-1}b$ with $M^{-1} = \Delta^{-1}\begin{pmatrix} -2\tilde{\mu} & -\tilde{\kappa} \\ \tilde{\kappa} & -2\tilde{\mu} \end{pmatrix}$. Both $z^\star$ and $\hat{x}^\star$ solve the equalized regularized first-order condition $\mathbb{E}[\tilde{D}_i] = 0$, i.e. $f(z) = 0$. For (40), evaluate the field at the anchor: since the executed strategy at $\bar{z}$ is r itself, the regularizer vanishes there and $f(\bar{z}) = \hat{b} := \binom{c_2 + \kappa_A r_{2,1}}{d_2 - \kappa_A r_{1,1}}$ is the pair of unregularized expected payoff differences against the anchor, each a difference of two entries of A, so $\|\hat{b}\|_2 \le \sqrt{2} \cdot 2U$. Then $z^\star - \bar{z} = -M^{-1}\hat{b}$ and $M^\top M = \Delta I$ gives $\|M^{-1}\|_2 = \Delta^{-1/2} \le (2\tilde{\mu})^{-1}$, whence $\|z^\star - \bar{z}\|_2 \le 2\sqrt{2}U/(2\tilde{\mu}) = \sqrt{2}U/\tilde{\mu}$. Condition (41) then places $z^\star$ coordinatewise in $[\delta_0, 1-\delta_0]$. □

Lemma A.9 (Lyapunov potential and weighted drift)**.** *Let Φ_i be as in* (5) *with fixed point $z^\star$ of Lemma A.8; write $e = z - z^\star$ and $V(z) = \sum_i \Phi_i(z_i)$.* ***(V1) On*** *$[0,1]$, $\frac{1}{2}e_i^2 \le \Phi_i(z_i) \le e_i^2$; hence $\frac{1}{2}\|e\|^2 \le V(z) \le \|e\|^2 \le 2$ on $[0,1]^2$ and $|\Phi_i'(z_i)| = |e_i|/\varphi(z_i) \le 2$.* ***(V2)*** (weighted drift.) *For any weights $\omega_1, \omega_2 \in [1/\sqrt{2}, 1]$ and every $z \in [0,1]^2$,*

$$\sum_{i=1}^{2} \omega_i\, e_i\, f_i(z) \;\le\; -m_w\|e\|^2, \qquad m_w := \sqrt{2}\,\tilde{\mu} - \tfrac{1}{2}\Big(1 - \tfrac{1}{\sqrt{2}}\Big)|\tilde{\kappa}|, \tag{42}$$

and under (A-push) $|\tilde{\kappa}| \le \tilde{\mu}/4$, so $m_w \ge \big(\sqrt{2} - \frac{1}{8}(1 - \frac{1}{\sqrt{2}})\big)\tilde{\mu} \ge 1.37\,\tilde{\mu} > 0$. ***(V3)*** *$\Phi_i'' = \frac{1}{\varphi} - \frac{(v - z_i^\star)\varphi'}{\varphi^2}$, $|\Phi_i''| \le 10 =: L_V$ on $[0,1]$.*

Proof. The integrand of Φ_i has the sign of e_i and $1/\varphi \in [1,2]$ on $[0,1]$, so $\frac{1}{2}e_i^2 \le \Phi_i \le e_i^2$; $|e_i| \le 1$ gives the rest of (V1). For (V2), $f(z) = M(z - z^\star) = Me$ by Lemma A.8, so $f_1 = -2\tilde{\mu}e_1 + \tilde{\kappa}e_2$, $f_2 = -\tilde{\kappa}e_1 - 2\tilde{\mu}e_2$ and

$$\sum_i \omega_i e_i f_i = -2\tilde{\mu}\big(\omega_1 e_1^2 + \omega_2 e_2^2\big) + \tilde{\kappa}\, e_1 e_2\, (\omega_1 - \omega_2). \tag{43}$$

The first term is at most $-2\tilde{\mu}\|e\|^2/\sqrt{2} = -\sqrt{2}\tilde{\mu}\|e\|^2$ since $\omega_i \ge 1/\sqrt{2}$; the second is at most $|\tilde{\kappa}| \cdot \frac{1}{2}\|e\|^2 \cdot (1 - 1/\sqrt{2})$ since $|e_1 e_2| \le \frac{1}{2}\|e\|^2$ and $|\omega_1 - \omega_2| \le 1 - 1/\sqrt{2}$. Under (A-push), $|\kappa_A| \le \mu\delta_0/2 < \mu/4$ (as $\delta_0 < \frac{1}{2}$), hence $|\tilde{\kappa}| \le \tilde{\mu}/4$ and $m_w \ge (1.4142 - 0.0366)\tilde{\mu} \ge 1.37\tilde{\mu}$. (V3) differentiates $\Phi_i' = (v - z_i^\star)/\varphi$ using $1/\varphi \le 2$, $|v - z_i^\star| \le 1$, $|\varphi'| \le 2$, $1/\varphi^2 \le 4$ on $[0,1]$. □

Remark A.10 (Why the potential Φ_i and the weights $N_{i,t}$). The plain-Euclidean inner product does not cancel the zero-sum coupling of the normalized dynamics: with the expected increment $a_i\varphi(z_i)f_i$,

$$\sum_i a_i\varphi(z_i)f_i(z)e_i = -2\tilde{\mu}\sum_i a_i\varphi(z_i)e_i^2 + \tilde{\kappa}\, e_1e_2\big[a_1\varphi(z_1) - a_2\varphi(z_2)\big], \tag{44}$$

and $\varphi(z_1) - \varphi(z_2) = 2(z_1 - z_2)(z_1 + z_2 - 1)$ does not vanish with $\|e\|$; no fixed (weighted) norm can kill it globally. Concretely, for $A = \left(\begin{smallmatrix} 3 & 0 \\ -1 & 1 \end{smallmatrix}\right)$ ($\kappa_A = 5$, interior mixed equilibrium $(2/5, 1/5)$), with $\gamma = 0$, the anchor placed at that equilibrium (so that $z^\star = (2/5, 1/5)$ is the regularized fixed point) and equal step sizes $a_1 = a_2$, the quadratic form (44) at $z^\star$ has diagonal $-2\mu(0.52, 0.68)$ and off-diagonal $\frac{1}{2}\cdot 5\cdot(-0.16)$, so it is indefinite for $\mu < \sqrt{0.16/1.4144} \approx 0.336$ and global strong monotonicity in the plain norm fails. The potential Φ_i removes the factor $\varphi(z_i)$ ($\Phi_i'\cdot a_i\varphi f_i = a_ie_if_i$), but the bracket $a_1 - a_2$ of the two step sizes remains. Controlling it through the ratio a_1/a_2 is unattractive: for fixed step sizes the form is negative definite iff $|\kappa_A|\,|a_1 - a_2| < 4\mu\sqrt{a_1a_2}$, so over the ratio window $a_1/a_2 \in [1/R, R]$, $R = s_2/s_1$, that the regret window $[s_1\sqrt{t}, s_2\sqrt{t}]$ of Lemma A.3 provides, the sharp sufficient condition is $|\kappa_A| < 4\mu\sqrt{R}/(R-1)$. Since $R \approx 10^3$ or more without dither (on matching pennies at $\gamma = 0.05$, $R \approx 1.2\times 10^3$ at $\mu = 20$ and 5.6×10^3 at $\mu = 100$), the condition fails at $\mu = 20$ ($2.3 < 4$) and holds only at much larger anchor strength ($5.4 > 4$ at $\mu = 100$), and it would tie the admissible coupling to μ through the window, a dependence absent from (A-push). The weights $N_{i,t} = \|\theta_{i,t}\|_2$ resolve this without any ratio condition: $N_{i,t}a_{i,t} = \sqrt{\varphi(z_i^t)} \in [1/\sqrt{2}, 1]$, so (V2) applies with $\omega_i = \sqrt{\varphi(z_i^t)}$. This is the continuous-time fact that $\dot{e} = D\,Me$ is stable for every *fixed* positive diagonal D, with $e^\top D^{-1}e$ as the Lyapunov function (for a time-varying D the derivative of D^{-1} contributes an extra term, which in the discrete proof is the separately controlled weight drift $(N_{i,t+1} - N_{i,t})_+$ of Lemma A.13); here $D = \mathrm{diag}(a_i\varphi_i)$ and D^{-1} is realized, up to the bounded factor $\sqrt{\varphi_i}$, by $N_{i,t}$.

Lemma A.11 (Alternating-update absorption). *Let $\mathcal{F}_{t+1/2} = \mathcal{F}_t \vee \sigma$(player 1's round-$t$ samples); player 2 samples against the already updated $\hat{x}_1^{t+1}$. With $\tilde{z}^t := (z_1^{t+1}, z_2^t)$, $\mathbb{E}[\tilde{D}_2 \mid \mathcal{F}_{t+1/2}] = f_2(\tilde{z}^t)$, so $\zeta_2^t := \tilde{D}_2 - f_2(\tilde{z}^t)$ has $\mathbb{E}[\zeta_2^t \mid \mathcal{F}_{t+1/2}] = 0$ and, for every $\mathcal{F}_t$-measurable bounded Λ, the tower property gives*

$$\mathbb{E}[\Lambda\,\zeta_2^t \mid \mathcal{F}_t] = \Lambda\,\mathbb{E}[\mathbb{E}[\zeta_2^t \mid \mathcal{F}_{t+1/2}] \mid \mathcal{F}_t] = 0. \tag{45}$$

The alternating deviation $\eta^t := f_2(\tilde{z}^t) - f_2(z^t)$ obeys, on a non-fallback step of player 1 with $S_{1,t} \geq 2G_D$ and under (A-ε'') (so that Lemma A.4 identifies the realized update),

$$|\eta^t| = |\tilde{\kappa}|\,|z_1^{t+1} - z_1^t| \leq 2|\tilde{\kappa}|G_D\,a_{1,t}, \tag{46}$$

because f_2 is affine with $\partial_{z_1}f_2 = -\tilde{\kappa}$ (Lemma A.7) and $|z_1^{t+1} - z_1^t| = |\mathrm{proj}_{[0,1]}(z_{v,1}^{t+1}) - \mathrm{proj}_{[0,1]}(z_1^t)| \leq |z_{v,1}^{t+1} - z_1^t| \leq 2G_Da_{1,t}$ by nonexpansivity of the projection (Lemma A.4) and the amplitude bound of (37).

Proof. Player 2's conditional mean is evaluated at $\tilde{z}^t$ because her opponent policy at her update time is the updated $\hat{x}_1^{t+1}$; $\mathcal{F}_t \subset \mathcal{F}_{t+1/2}$ gives (45). The bound (46) is the affine form of f_2 together with the nonexpansive step bound. □

Lemma A.12 (Extended-domain Taylor bound). *On $\{\varrho > t\}$ with $t \geq t_0$ and $t_0 \geq (4G_D/s_1)^2$, $S_{i,t} \geq s_1\sqrt{t} \geq 4G_D$, so $2G_Da_{i,t} \leq \frac{1}{2}$ and the virtual point and the segment $[z_i^t, z_{v,i}^{t+1}]$ lie in $\mathcal{Z}_1^+ = [-\frac{1}{2}, \frac{3}{2}]$ by* (37). *There, $\varphi \geq \frac{1}{2}$, $|\varphi'| \leq 4$, $|v - z_i^\star| \leq \frac{3}{2}$, $1/\varphi^2 \leq 4$, so by (V3)*

$$|\Phi_i''(v)| \leq \tfrac{1}{\varphi} + \tfrac{|v - z_i^\star||\varphi'|}{\varphi^2} \leq 2 + \tfrac{3}{2}\cdot 4\cdot 4 = 26 =: L_V^+, \tag{47}$$

and, for $\Delta z_i^v := z_{v,i}^{t+1} - z_i^t$,

$$\Phi_i(z_{v,i}^{t+1}) \leq \Phi_i(z_i^t) + \Phi_i'(z_i^t)\,\Delta z_i^v + \tfrac{L_V^+}{2}(\Delta z_i^v)^2 \qquad (i = 1, 2). \tag{48}$$

Proof. On the event the hypotheses of Lemma A.6 hold, giving the amplitude bound; $z_i^t \in [0, 1]$ and half-width $\leq \frac{1}{2}$ keep the segment in $\mathcal{Z}_1^+$. (47) is (V3) on the larger domain; (48) is the second-order Taylor expansion with Lagrange remainder along the segment. □

The last block assembles the one-step engine of Lemmas A.6–A.12 into the truncated recursion used by the theorem, on the weighted function V_t of (6).

Lemma A.13 (Truncated one-step recursion)**.** *Let (M), (A-var), (A-int), (A-push) and (A-ε'') hold (the window constants s_1, s_2 are those of Lemma A.3) and let $t \ge t_0$ with $t_0 \ge t_0^\dagger$ of* (54)*, so that on $\{\varrho > t\}$ one has $S_{i,t} \ge s_1\sqrt{t_0} \ge 4G_D$ for both players (the step hypotheses of Lemmas A.6–A.12) and, by (A-ε''), $S_{i,t} >$ EPS (no fallback); in particular $a_{i,t} = 1/S_{i,t}$ and the virtual increments Δz_i^v are well defined there. Then, pointwise on $\{\varrho > t\}$,*

$$W_{t+1} \le \sum_{i=1}^{2} \Big[N_{i,t} \Big(\Phi_i(z_i^t) + \Phi_i'(z_i^t)\Delta z_i^v + \tfrac{L_V^+}{2}(\Delta z_i^v)^2 \Big) + G_D^2 a_{i,t} \Big], \tag{49}$$

while $W_{t+1} = 0$ on $\{\varrho \le t\}$; and taking conditional expectations,

$$\mathbb{E}[W_{t+1} \mid \mathcal{F}_t] \le \Big(1 - \frac{m_w}{s_2\sqrt{t}}\Big) W_t + \frac{C_w}{s_1\sqrt{t}} \mathbb{1}\{\varrho > t\}, \qquad C_w = 114\, G_D^2 + 4|\tilde\kappa| G_D, \tag{50}$$

with $m_w \ge 1.37\tilde\mu > 0$ of Lemma A.9(V2); the factor $1 - m_w/(s_2\sqrt{t})$ lies in $[0,1)$ because $t \ge t_0^\dagger \ge (m_w/s_2)^2$.

Proof. Fix $t \ge t_0$. On $\{\varrho > t\}$ both $S_{i,t} \ge s_1\sqrt{t_0} \ge 4G_D$ (by $t_0 \ge (4G_D/s_1)^2$ and the window lower bound), so Lemma A.6, Lemma A.4 and Lemma A.12 apply to both players and no fallback occurs ($4G_D > G_D/4 \ge$ EPS). All quantities $N_{i,t}, a_{i,t}, z^t$ are $\mathcal{F}_t$-measurable.

(1) Pointwise bound (49). On $\{\varrho \le t\}, \mathbb{1}\{\varrho > t+1\} = 0$ and $W_{t+1} = 0$. On $\{\varrho > t\}, W_{t+1} \le V_{t+1}$ (as $\mathbb{1}\{\varrho > t+1\} \le \mathbb{1}\{\varrho > t\}$, $V_{t+1} \ge 0$), and for each player

$$N_{i,t+1}\Phi_i(z_i^{t+1}) \;\le\; N_{i,t}\,\Phi_i(z_i^{t+1}) + \big(N_{i,t+1} - N_{i,t}\big)_+ \Phi_i(z_i^{t+1}). \tag{51}$$

Weight drift. On a non-fallback step the energy inequality of Lemma A.3 (Step 2, orthogonality (17) and nonexpansive clipping) gives $Y_{i,t+1} \le Y_{i,t} + \|g\|_2^2 \le Y_{i,t} + G_D^2$, hence $N_{i,t+1} \le \sqrt{N_{i,t}^2 + G_D^2} \le N_{i,t} + G_D^2/(2N_{i,t})$; with $N_{i,t} \ge S_{i,t}/\sqrt{2}$ and $\Phi_i(z_i^{t+1}) \le 1$ on $[0,1]$ (V1), the second term of (51) is at most $G_D^2 a_{i,t}/\sqrt{2} \le G_D^2 a_{i,t}$. *Position.* Projection domination (Lemma A.4) and the extended Taylor bound (48) give $\Phi_i(z_i^{t+1}) \le \Phi_i(z_{v,i}^{t+1}) \le \Phi_i(z_i^t) + \Phi_i'(z_i^t)\Delta z_i^v + \frac{L_V^+}{2}(\Delta z_i^v)^2$, whether or not the step clips. Multiplying by $N_{i,t} \ge 0$ and summing over i gives (49).

(2) Conditional recursion (50). Since $\{\varrho > t\} \in \mathcal{F}_t$ and $W_{t+1} = 0$ off it, $\mathbb{E}[W_{t+1} \mid \mathcal{F}_t] = \mathbb{1}\{\varrho > t\}\mathbb{E}[W_{t+1} \mid \mathcal{F}_t]$, and it suffices to bound the conditional expectation of the right-hand side of (49) on $\{\varrho > t\}$ (all its terms are defined and bounded there). Decompose $\Delta z_i^v = a_{i,t}\varphi(z_i^t)\tilde D_i + r_i^t$ by (37); then, using $\Phi_i'(z_i^t) = e_i^t/\varphi(z_i^t)$ and $N_{i,t}a_{i,t} = \sqrt{\varphi(z_i^t)} =: \omega_i \in [1/\sqrt{2}, 1]$,

$$N_{i,t}\Phi_i'(z_i^t)\Delta z_i^v = \omega_i\, e_i^t\, \tilde D_i + N_{i,t}\Phi_i'(z_i^t)\, r_i^t. \tag{52}$$

Taking $\mathbb{E}[\cdot \mid \mathcal{F}_t]$ of the first term: for player 1, $\mathbb{E}[\tilde D_1 \mid \mathcal{F}_t] = f_1(z^t)$; for player 2, the tower property through $\mathcal{F}_{t+1/2}$ and Lemma A.11 give $\mathbb{E}[\tilde D_2 \mid \mathcal{F}_t] = f_2(z^t) + \mathbb{E}[\eta^t \mid \mathcal{F}_t]$ with $|\eta^t| \le 2|\tilde\kappa| G_D a_{1,t}$ (the martingale parts vanish by (45) with $\Lambda = \omega_2 e_2^t$). Bounding the groups:

drift $\sum_i \omega_i e_i^t f_i(z^t) \le -m_w\|e^t\|^2$ by Lemma A.9(V2), since $\omega_i \in [1/\sqrt{2}, 1]$;

alternating $|\omega_2 e_2^t\, \mathbb{E}[\eta^t \mid \mathcal{F}_t]| \le 1 \cdot 1 \cdot 2|\tilde\kappa| G_D a_{1,t} \le 4|\tilde\kappa| G_D\, \bar a_t$ ($a_{1,t} \le 2\bar a_t$);

residual $N_{i,t}|\Phi_i'(z_i^t)|\,|r_i^t| \le N_{i,t} \cdot 2 \cdot 2G_D^2 a_{i,t}^2 = 4G_D^2\, \omega_i a_{i,t} \le 4G_D^2 a_{i,t}$ ((37), (V1));

quadratic $N_{i,t}\frac{L_V^+}{2}(\Delta z_i^v)^2 \le N_{i,t} \cdot 13 \cdot 4G_D^2 a_{i,t}^2 = 52G_D^2\, \omega_i a_{i,t} \le 52G_D^2 a_{i,t}$ (amplitude (37));

weight $G_D^2 a_{i,t}$ from (49).

The three per-player groups sum to $57G_D^2(a_{1,t} + a_{2,t}) = 114G_D^2\bar a_t$. Assembling, on $\{\varrho > t\}$,

$$\mathbb{E}[V_{t+1} \mid \mathcal{F}_t] \le V_t - m_w\|e^t\|^2 + \big(114G_D^2 + 4|\tilde\kappa| G_D\big)\bar a_t = V_t - m_w\|e^t\|^2 + C_w\bar a_t. \tag{53}$$

Finally, on $\{\varrho > t\}$ the window gives $N_{i,t} \le S_{i,t} \le s_2\sqrt{t}$, so by (V1) $V_t = \sum_i N_{i,t}\Phi_i(z_i^t) \le s_2\sqrt{t}\,\|e^t\|^2$, i.e. $-m_w\|e^t\|^2 \le -m_w V_t/(s_2\sqrt{t})$; and $\bar{a}_t \le 1/(s_1\sqrt{t})$. Multiplying by $\mathbb{1}\{\varrho > t\}$ and using $W_t = V_t\mathbb{1}\{\varrho > t\}$ yields (50). □

The bound (50) is the engine of the theorem: the drift term of the transformed game cancels the sampling martingale and is absorbed by the weighted Lyapunov function at rate $-m_w/(s_2\sqrt{t})$, while the residual noise, the Taylor remainder and the motion of the weights are all first order in the step size and enter only through the forcing $C_w/(s_1\sqrt{t})$. The projection lemma is what makes the inequality branch-free (the same chain holds with or without clipping), which removes the trajectory-clipping tail from the argument entirely, and the weights $N_{i,t}$ are what remove any condition on the ratio of the two players' step sizes (Remark A.10).

A.4 Proof of Theorem 4.1

Fix the game (A, U, τ_d) and the parameters $(\mu, \gamma, \sigma_d, \delta_0)$; the bound below is pointwise in these, with the assembled K, T_0 of (64)–(56) carrying the dependence explicitly. The proof has four steps: *(i)* a supply argument that the regret norms stay in the two-sided window $[s_1\sqrt{t}, s_2\sqrt{t}]$ (Lemma A.3); *(ii)* a closed recursion on the weighted Lyapunov expectation (Lemma A.13 + Lemma A.1); *(iii)* a restoration step removing the truncation and converting the weighted value back to $\|e^T\|^2$; *(iv)* a conversion from e^T to the executed mixed strategy.

The K/T_0 chain. The burn-in must satisfy two requirements: $S_{i,t} \ge s_1\sqrt{t_0} \ge 4G_D$ on $\{\varrho > t\}$ (the denominator bound of Lemma A.6, the $\mathcal{Z}^+$-Taylor step and the no-fallback property are legal), and $c := m_w/s_2 \le \sqrt{t_0}$ (the factor of (50) is nonnegative and Lemma A.1 applies). Let $t_0 \ge t_0^\dagger$ with

$$t_0^\dagger := \left\lceil \max\left(\left(\tfrac{4G_D}{s_1}\right)^2,\ \left(\tfrac{m_w}{s_2}\right)^2 \right) \right\rceil, \tag{54}$$

an integer; since $s_1 < G_D/(16\sqrt{2})$ (from $c_Y \le c_\pi^2$, $\tilde{\mu} \le \mu \le G_D/2$ and $\delta_0 < \frac{1}{2}$, so $c_\pi < G_D/16$), $t_0^\dagger > 8192$, so in particular $t_0^\dagger \ge 1$. The supply tail (16) is then absorbed by scheduling

$$t_0(T) := \max\Big(t_0^\dagger,\ \Big\lceil \tfrac{1}{c_6}\big(\tfrac{1}{2}\log T + \log \tfrac{4}{1-e^{-c_6}}\big) \Big\rceil \Big) = \Theta(\log T), \tag{55}$$

an integer nondecreasing in T. Define T_0 so that the two scale conditions, (W1) $t_0(T) < T$, and (W2) homogeneous absorption

$$2s_2\sqrt{t_0(T)}\; e^{-2c(\sqrt{T}-\sqrt{t_0(T)})} \;\le\; 1,$$

hold for *every* $T \ge N$:

$$T_0 := \min\Big\{ N \in \mathbb{N} :\ \forall\, T \ge N,\ \ \text{(W1)–(W2)} \Big\}, \tag{56}$$

finite since $t_0(T) = \Theta(\log T)$: (W1) holds for all large T, and since $\sqrt{t_0(T)} = o(\sqrt{T})$ the exponent $2c(\sqrt{T} - \sqrt{t_0}) \sim 2c\sqrt{T}$ dominates the factor $\sqrt{t_0(T)} = O(\sqrt{\log T})$, which makes (W2). The theorem bound below holds for all $T \ge T_0$. Condition (W2) is exactly the piece that Lemma A.1 leaves at the T-scale (the homogeneous term started from the burn-in value $w_{t_0} \le 2s_2\sqrt{t_0}$); everything else is charged to the explicit constants K_w, K_e below.

Step 1 (supply: growth of the regret norms). Under (A-push)+(A-var)+(A-int)+(A-ε''), Lemma A.3 shows the two-sided growth: deterministically $S_{i,t} \le s_2\sqrt{t}$ (item (i)), and for the two-player union

$$\Pr\big(\exists\, t \ge t_0,\ i \in \{1,2\} :\ S_{i,t} < s_1\sqrt{t}\big) \;\le\; \frac{2\,e^{-c_6 t_0}}{1 - e^{-c_6}} \tag{57}$$

(item (ii)); the factor 2 is the union bound and the only multiplicity used anywhere below. Since ϱ first exits when some $S_{i,t}$ leaves the window, $\Pr(\varrho \le T) \le$ the same tail, independently of T. On $\{\varrho > t\}$ with $t \ge t_0 \ge t_0^\dagger$ every hypothesis of Lemma A.13 holds on every counted step.

Step 2 (recursion and closing). Take $t_0 = t_0(T)$ of (55); then $t_0 < T$ for $T \ge T_0$ (W1). With $w_t := \mathbb{E}W_t$, Lemma A.13 and the tower property give the scalar recursion

$$w_{t+1} \le \Big(1 - \frac{m_w/s_2}{\sqrt{t}}\Big) w_t + \frac{C_w/s_1}{\sqrt{t}}, \qquad t_0 \le t < T, \tag{58}$$

since $\mathbb{E}[\mathbb{1}\{\varrho > t\}] \le 1$. The burn-in value is bounded deterministically: $W_{t_0} \le V_{t_0} \le \max_i N_{i,t_0} \cdot \|e^{t_0}\|^2 \le s_2\sqrt{t_0} \cdot 2$ by (V1), $N_{i,t} \le S_{i,t} \le s_2\sqrt{t}$ (Lemma A.3(i)) and $\|e\|^2 \le 2$ on $[0,1]^2$, so $w_{t_0} \le 2s_2\sqrt{t_0}$. Lemma A.1 with $c = m_w/s_2 \in (0, \sqrt{t_0}]$ (positivity by Lemma A.9(V2) under (A-push); $c \le \sqrt{t_0}$ by (54)) and $C' = C_w/s_1 \ge 0$ gives, with $t_0 = t_0(T)$,

$$w_T \;\le\; 2s_2\sqrt{t_0}\, e^{-2c(\sqrt{T}-\sqrt{t_0})} + \frac{C_w}{c\, s_1} = 2s_2\sqrt{t_0}\, e^{-2c(\sqrt{T}-\sqrt{t_0})} + \frac{C_w\, s_2}{m_w\, s_1}. \tag{59}$$

For $T \ge T_0$ the homogeneous term of (59) is ≤ 1 by (W2); hence

$$w_T \;\le\; K_w, \qquad K_w := 1 + \frac{C_w}{c\, s_1} = 1 + \frac{C_w\, s_2}{m_w\, s_1}, \qquad c = \frac{m_w}{s_2}. \tag{60}$$

Dependence: $C_w = 114G_D^2 + 4|\tilde{\kappa}|G_D$ (Lemma A.13), $s_1 = \sqrt{c_Y/2}$, $s_2 = \sqrt{3}\,G_D$, m_w of Lemma A.9(V2).

Step 3 (restoration and un-weighting). Split $\|e^T\|^2 = \|e^T\|^2\mathbb{1}\{\varrho > T\} + \|e^T\|^2\mathbb{1}\{\varrho \le T\}$. On $\{\varrho > T\}$ both regret norms satisfy $S_{i,T} \ge s_1\sqrt{T}$, so $N_{i,T} \ge S_{i,T}/\sqrt{2} \ge s_1\sqrt{T}/\sqrt{2}$ and, by (V1),

$$W_T = \sum_i N_{i,T}\Phi_i(z_i^T)\mathbb{1}\{\varrho > T\} \;\ge\; \frac{s_1\sqrt{T}}{\sqrt{2}} \cdot \frac{1}{2}\|e^T\|^2\mathbb{1}\{\varrho > T\},$$

$$\text{i.e.} \qquad \|e^T\|^2\mathbb{1}\{\varrho > T\} \le \frac{2\sqrt{2}}{s_1\sqrt{T}}\, W_T. \tag{61}$$

On $\{\varrho \le T\}$, $\|e^T\|^2 \le 2$. Taking expectations,

$$\mathbb{E}\|e^T\|^2 \;\le\; \frac{2\sqrt{2}}{s_1\sqrt{T}}\, w_T + 2\Pr(\varrho \le T) \;\le\; \frac{2\sqrt{2}K_w}{s_1}\, T^{-1/2} + T^{-1/2} \;=\; K_e T^{-1/2}, \tag{62}$$

with $K_e := 2\sqrt{2}K_w/s_1 + 1$, where the middle inequality uses (60) and (55): $e^{-c_6 t_0} \le e^{-\frac{1}{2}\log T} \cdot \frac{1-e^{-c_6}}{4}$ makes the two-player tail $2\Pr(\varrho \le T) \le T^{-1/2}$. There is no fallback/clipping tail term: the projection domination of Lemma A.4 makes the recursion branch-free on the window.

Step 4 (conversion). With the executed mixture $\hat{x}_i = (1-\gamma)x_{\mathrm{b},i} + \frac{\gamma}{2}\mathbf{1}$, $\hat{x} - \hat{x}^\star = (1-\gamma)(x_{\mathrm{b}} - x_{\mathrm{b}}^\star)$ componentwise, and $x_{\mathrm{b}} - x_{\mathrm{b}}^\star = (e_1, -e_1, e_2, -e_2)$, the exact conversion is

$$\|\hat{x}^T - \hat{x}^\star\|^2 = 2(1-\gamma)^2\|e^T\|^2, \tag{63}$$

the factor 2 counting the two action coordinates per player. Hence $\mathbb{E}\|\hat{x}^T - \hat{x}^\star\|^2 = 2(1-\gamma)^2\mathbb{E}\|e^T\|^2 \le KT^{-1/2}$ for all $T \ge T_0$, with

$$K := 2(1-\gamma)^2 K_e, \qquad K_e = \frac{2\sqrt{2}}{s_1}\Big[1 + \frac{C_w\, s_2}{m_w\, s_1}\Big] + 1, \tag{64}$$

which is (3). The dependence chain of K is explicit, with the only absorbed constants being the unit homogeneous term of K_w ((W2), (56)) and the unit restoration tail of K_e (Step 3); beyond those the inputs are the leaf constants $A, U, \tau_d, \mu, \gamma, \sigma_d, \delta_0$: $K \leftarrow \gamma, K_e$; $K_e \leftarrow K_w, s_1$; $K_w \leftarrow C_w, s_1, m_w, s_2$; $C_w \leftarrow G_D, \tilde{\kappa}$; $s_1, s_2 \leftarrow c_Y, G_D$; $m_w \leftarrow \tilde{\mu}, \tilde{\kappa}$; $c_Y \leftarrow \sigma_{\min}^2, \delta_0, \tilde{\mu}, G_D$; $c_6 \leftarrow c_Y, G_D$ through $t_0(T)$; T_0 depends on $c_6, t_0^\dagger, s_2, c$. ■

Remark A.14 (What the hypotheses buy). (A-var) and (A-ε'') enter only through the regret-norm window: (A-var) through the supply lemma (the regret norms grow like $\sqrt{t}$, which is what makes the RM$^+$ step size $1/S_{i,t}$ behave as a $t^{-1/2}$ schedule), (A-ε'') through the supply lemma and through the certificate that no step counted on the window is a fallback step ($S_{i,t} \ge 4G_D > \mathrm{EPS}$, Remark A.5); (A-int) is used in the supply lemma (layer drift) and to keep $z^\star$ inside the box where Φ_i is compared; (A-push) is used in the supply lemma and to make $m_w > 0$ in Lemma A.9(V2). No hypothesis constrains the relative size of the two players' regret norms.

A.5 Diagnostics

No ratio hypothesis. A bound on the ratio of the two players' RM$^+$ regret norms (equivalently of their step sizes) is not available. On real game trees the measured bare ratio across information sets

grows approximately like $\sqrt{t}$ over the tested budgets, and the span appears stable there only after excluding a small set of *frozen* information sets where the winner action has been purified (Table 5). Within the single-set model the two-sided window of Lemma A.3 gives $a_{1,t}/a_{2,t} \in [s_1/s_2, s_2/s_1]$, but the window ratio s_2/s_1 is so large that the skew coupling can be dominated through it only at very large μ (Remark A.10). Theorem 4.1 therefore uses the weighted Lyapunov function (6) and needs no statement about the ratio at all.

Table 5: Ratio of RM$^+$ regret norms across information sets and, once the frozen information sets are excluded, the internal span of the remaining active set on Kuhn at $\mu = 1.0$. The ratio is read at the three budget checkpoints of the sweep on Kuhn and at the end of the run on Leduc.

	Kuhn	Leduc
Ratio, full set	$61 \to 394 \to 1999$	$\approx$ 530–560
Span, active set ($\mu = 1.0$)	$3.2 \to 3.5 \to 3.5$	—

Working range of the anchor. The theory restricts μ from below through two explicit conditions: (i) (A-int), for which (41) suffices: the window's anchor r sits $\sqrt{2}U/\tilde{\mu}$ inside the margin δ_0 in base coordinates (the fixed point is within $\sqrt{2}U/\tilde{\mu}$ of the anchor, (40)); (ii) (A-push), $|\kappa_A| \le \mu\delta_0/2$, i.e. $\mu \ge 2|\kappa_A|/\delta_0$, which is used in the supply lemma (the anchor drift must dominate the coupling in the boundary layers of width $\delta_0/2$) and gives $m_w > 0$. Neither condition involves the anchor beyond its interiority, so the theorem applies to a fresh-start run at any fixed anchor in the executed box, in particular at any window average of Algorithm 1, provided the anchor satisfies (A-int), for instance through (41), and the remaining hypotheses. The later windows of Algorithm 1 inherit the regret state of the previous window and are not covered by the statement. Each such fresh-start run contracts to its fixed point $\hat{x}^\star(\gamma, \mu, r)$, which by (40) moves with the anchor and tends to it as $\mu \to \infty$. Instantiated on matching pennies $A = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$ ($U = 1$, $\kappa_A = 4$) with the uniform anchor ($z^\star = (\frac{1}{2}, \frac{1}{2})$ for every μ, so (A-int) holds with any $\delta_0 < \frac{1}{2}$), (A-push) with $\delta_0 = 0.4$ requires $\mu \ge 20$, and (A-var) holds for every $\gamma \in (0, 1)$ with $\sigma^2_{\min} = 4\gamma(1 - \gamma/2)$ (the first mover's floor under the implemented sampler; the second mover's is $8\gamma(1 - \gamma/2)$), and (A-ε'') holds for the implemented $\mathrm{EPS} = 10^{-12}$ at every μ since $\sigma_{\min}/8 \approx 0.055$ at $\gamma = 0.05$. The proven regime is thus a strong-anchor regime, well above the $\mu = 1.0$ used in the experiments: the sufficient conditions are conservative, and the gap lies in the layer argument of Lemma A.3. Empirically the working range is much wider: a discretized scan on Kuhn, re-solving the reference fixed point per μ level, finds confirmed counterexamples to the monotone-potential condition for all tested $\mu \in [0.10, 0.22]$ and a sign flip between 0.22 and 0.24 (empirical threshold ≈ 0.23); the same scan on Leduc finds no counterexample down to $\mu = 0.02$, so the threshold is game-dependent. The main text therefore uses $\mu = 1.0$ and reports $\mu \ge 0.3$ as the working range (with margin above the Kuhn threshold); robustness at $\mu = 0.3$ is reported in §B.3.

Sampler variance check. A unit test enumerating the payoff-difference distribution of the implemented sampler on matching pennies gives variance $16p(1-p)$ for the second mover and $8p(1-p)$ for the first, p the opponent's first-action probability, matching the empirical variance over 4×10^4 sampled updates to within one percent.

Rate check on matching pennies. Table 6 reports the rate check of §5 on the fixed-anchor process of Theorem 4.1: matching pennies, $\gamma = 0.05$, a uniform anchor, no outer update, 5 seeds, 2×10^6 touches and 24 log-spaced checkpoints from 6.7×10^3 to 2×10^6. The target $\hat{x}^\star$, each player's executed strategy equal to $(\frac{1}{2}, \frac{1}{2})$, is exact by symmetry, and a 2×10^5-iteration full-feedback reference returns it with zero residual.

$n \ge 3$ projection margin (numerically probed). The projection lemma (Lemma A.4) is $n = 2$-specific: clipping one of two components lands on a simplex vertex and the virtual point lies beyond that vertex. For an $n \ge 3$-action information set, a random-event probe (20000 clipping events per (n, δ), targets stratified by $\min_i y_i \ge \delta$) tests the Euclidean domination $\|x^{t+1} - y\|_2 \le \|x_v^{t+1} - y\|_2$ between the clipped-and-renormalized RM$^+$ strategy x^{t+1} and the renormalized virtual point $x_v^{t+1} = (\theta + g)/\langle \mathbf{1}, \theta + g\rangle$, for targets $y \in \Delta(A)$ with $\min_i y_i \ge \delta$ and steps with $S \ge 2\|g\|_1$. It

Table 6: Matching-pennies rate check. Slope: seed-mean log-log slope of $\|\hat{x}^t - \hat{x}^\star\|^2$ over the 24 checkpoints; (3) bounds the expectation by a $T^{-1/2}$ rate, so $-1/2$ is the exponent of an upper bound, not a two-sided prediction. Constant: $\sqrt{T}\,\mathbb{E}\|\hat{x}^T - \hat{x}^\star\|^2$ at $T = 2 \times 10^6$, where T counts touches (9 per paired update round on this two-node tree: 5 for the first mover's traversal, 4 for the second's) and the squared distance is the mean over the two information sets of $\|\hat{x}_i^T - \hat{x}_i^\star\|_2^2$ (one half of the four-coordinate norm of (63)). The $\mu = 1$ rows lie outside the proven regime, whose conditions hold for $\mu \geq 20$; the intervals are computed for the two in-regime strengths.

Sampler	μ	Slope	95% interval	Constant
ES	20	-0.52	$[-0.59, -0.49]$	0.046
ES	50	-0.47	$[-0.54, -0.42]$	0.015
ES	1	-0.50	—	0.68
OS	20	-0.45	$[-0.61, -0.32]$	0.037
OS	50	-0.45	$[-0.58, -0.34]$	0.008
OS	1	-0.51	—	0.97

finds no violation above the threshold

$$\delta_\infty^*(n) = \frac{1}{n}\Big[1 - \frac{2(\sqrt{n-1}-1)}{n-2}\Big] \quad (n \geq 3), \qquad \sup_{n \geq 3} \delta_\infty^*(n) = \delta_\infty^*(4) = \frac{2-\sqrt{3}}{4} \approx 0.0670, \tag{65}$$

with $\delta_\infty^*(3) = \frac{3-2\sqrt{2}}{3} \approx 0.0572$; for $n = 2$ the threshold is 0 (the single-infoset projection lemma). Above the threshold the probe finds no positive gap; below it the worst gap is first-order in $\varepsilon = \|g\|_1/S \leq 2nG_D a_t$ (measured ratio 0.22 to 0.25). In the relative form $\rho(n) := n\,\delta_\infty^*(n)$ (monotone to 1: $\rho(3) \approx 0.172$, $\rho(4) \approx 0.268$, $\rho(10) = 0.5$), the condition is $\delta_0(I) \geq \delta_\infty^*(n_I)$ per information set, a condition on the interior margin, not on μ. If the margin came only from the uniform exploration component of the anchor, $\min_i y_i \geq \gamma/n$, one would need $\gamma \geq \rho(n) = n\delta_\infty^*(n)$ ($\approx 0.17, 0.27, 0.5$ for $n = 3, 4, 10$); margins from the regularization's own interiority ((A-int), larger μ) are the alternative source. The threshold (65) is conjectural: it summarizes the probe and is not backed by an analytic argument here. A tree-level $n \geq 3$ projection extension of Theorem 4.1 along this line is the main open theoretical direction.

B Additional Material

B.1 Full results and significance tests

Table 8 reports all three output senses (last, anchor, average) for all seven arms and all ten (game, sampler) cells; baselines report no anchor output, marked n/a. All p-values in this subsection and in §B.3 are paired one-sided t-tests over the 10 shared seeds without multiplicity correction (§B.14 and §B.9 report two-sided intervals); the main text states orderings of means and defers significance to these tests. Table 7 reports the paired one-sided t-test p-values for RT-MCCFR$^+$'s raw last iterate against each baseline's last iterate.

Against the RM-family baselines the only non-significant cells are Goofspiel(4)-OS vs. RM$^+$ ($p = 0.47$) and Goofspiel(4)-ES/OS vs. IREG-PRM$^+$ ($p = 0.08, 0.18$); against PFTRL-KL the three cells where PFTRL is numerically better (Goofspiel(4)-ES/OS, LD(4)-OS) have signed paired differences of $+0.2091 \pm 0.0764$, $+0.1234 \pm 0.2587$ and $+0.0226 \pm 0.0232$ (raw last iterate minus PFTRL-KL last iterate, two-sided 95% intervals), so the first is resolved against RT-MCCFR$^+$ and only the other two are unresolved. On Goofspiel(4)-ES the PFTRL-KL last iterate is lower than the RT-MCCFR$^+$ raw last iterate by a resolved margin; the anchor of the same run is lower than both (0.1113 against 0.2729 and 0.4820). The anchor output is stronger: RT-MCCFR$^+$'s windowed anchor is significantly lower than every baseline's last iterate in 10/10 cells against RM$^+$, PRM$^+$, IREG-PRM$^+$ and PFTRL-RKL, and in 9/10 against PFTRL-KL (only LD(4)-OS, $p = 0.31$, where the margin is small: 0.0451 vs. 0.0499).

Multi-budget sweep. To separate a stationary noise floor from merely slow convergence, we repeat the Kuhn/Leduc runs at three budgets $\{5 \times 10^6, 2 \times 10^7, 4 \times 10^7\}$ with a *budget-independent*

Table 7: Paired one-sided t-test p-values for RT-MCCFR$^+$'s raw last iterate against each baseline's last iterate, over the 10 shared seeds; lower is better. $n = 10$ in every cell. Abbreviations: 1d6 = Liar's Dice with one d6, G4 = Goofspiel(4), LD4 = Liar's Dice with one four-sided die per player.

Cell	RM$^+$	IREG-PRM$^+$	PFTRL-KL
Kuhn-ES	0.0004	0.0005	0.15
Kuhn-OS	5.2×10^{-6}	4.6×10^{-4}	0.010
Leduc-ES	7.4×10^{-6}	1.9×10^{-3}	0.006
Leduc-OS	0.014	3.4×10^{-3}	1.7×10^{-6}
1d6-ES	2.6×10^{-7}	1.9×10^{-5}	4.7×10^{-18}
1d6-OS	2.2×10^{-4}	3.1×10^{-4}	3.5×10^{-5}
G4-ES	0.004	0.08	1
G4-OS	0.47	0.18	0.85
LD4-ES	8.2×10^{-4}	3.2×10^{-5}	2.9×10^{-14}
LD4-OS	6.9×10^{-5}	1.3×10^{-4}	0.97

Table 8: **All outputs for all arms** (mean over 10 seeds at the final touch count). $last$ = current strategy, $anchor$ = windowed average, reported by RT-MCCFR$^+$ only; avg = full-trajectory average. Abbreviations: 1d6 = Liar's Dice with one d6, G4 = Goofspiel(4), LD4 = Liar's Dice with one four-sided die per player.

Arm / output	Kuhn ES	Kuhn OS	Leduc ES	Leduc OS	1d6 ES	1d6 OS	G4 ES	G4 OS	LD4 ES	LD4 OS
RM$^+$ last	0.0995	0.1107	0.4841	0.4580	0.2253	0.3206	0.7322	0.8390	0.0870	0.1052
RM$^+$ anchor	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
RM$^+$ avg	0.0022	0.0079	0.0544	0.1498	0.0988	0.2999	0.0345	0.1394	0.0289	0.0475
PRM$^+$ last	0.1086	0.1124	0.4337	0.5102	0.2259	0.3286	0.8623	0.9044	0.0940	0.1012
PRM$^+$ anchor	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
PRM$^+$ avg	0.0021	0.0079	0.0552	0.1534	0.1012	0.3023	0.0308	0.0805	0.0297	0.0488
IREG-PRM$^+$ last	0.1169	0.1338	0.3980	0.4475	0.2220	0.3114	0.6220	0.9003	0.0842	0.1013
IREG-PRM$^+$ anchor	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
IREG-PRM$^+$ avg	0.0018	0.0075	0.0495	0.1457	0.0930	0.2891	0.0287	0.0986	0.0274	0.0440
PFTRL-KL (loc.) last	0.0191	0.0534	0.3990	1.4770	1.1589	0.3329	0.2729	0.7096	0.2372	0.0499
PFTRL-KL (loc.) anchor	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
PFTRL-KL (loc.) avg	0.0056	0.0047	0.0932	0.1412	1.3250	0.4757	0.1625	0.0369	0.4730	0.0810
PFTRL-RKL last	0.0264	0.0590	0.3266	0.8867	1.1849	0.3463	0.2615	0.7229	0.3119	0.0640
PFTRL-RKL anchor	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
PFTRL-RKL avg	0.0085	0.0074	0.1506	0.1081	1.3260	0.4759	0.2274	0.0670	0.5108	0.0912
RT-MCCFR$^+$-L last	0.0206	0.0353	0.3029	0.3578	0.1601	0.2881	0.4396	0.7484	0.0659	0.0699
RT-MCCFR$^+$-L anchor	0.0207	0.0355	0.3029	0.3579	0.1601	0.2881	0.4396	0.7480	0.0659	0.0699
RT-MCCFR$^+$-L avg	0.0255	0.0322	0.1077	0.1866	0.1497	0.3376	0.1506	0.1508	0.1060	0.1138
RT-MCCFR$^+$-W last	0.0153	0.0325	0.2728	0.3762	0.1519	0.2761	0.4820	0.8330	0.0555	0.0725
RT-MCCFR$^+$-W anchor	0.0041	0.0085	0.0708	0.1857	0.1002	0.2542	0.1113	0.1226	0.0376	0.0451
RT-MCCFR$^+$-W avg	0.0251	0.0298	0.1044	0.1830	0.1473	0.3362	0.1499	0.1443	0.1079	0.1170

schedule ($T_u = 2 \times 10^7/12$ for every budget), 10 seeds each; Table 2 in the main text reports the final NashConv and Table 9 the paired tests. RM$^+$'s last iterate improves only modestly with $8\times$ more budget: it decreases in all four cells but significantly in only two, and the RM$^+$ window's anchor is essentially flat, with an inconsistent direction. RT-MCCFR$^+$'s last iterate and anchor instead both decrease from 5×10^6 to 4×10^7, significantly in all four cells, the anchor by a fold ratio of 0.04 to 0.61 against 0.70 to 0.92 for RM$^+$'s last iterate. That improvement is concentrated in the first half of the range: between 2×10^7 and 4×10^7 none of the four anchor changes is resolved at 10 seeds, and the raw last iterate has an interval excluding zero only on Kuhn-OS. The smallest-budget rows are the transient regime: with only three T_u-windows the perturbation has not reached its floor, so RT-MCCFR$^+$'s anchor starts above the plain window average and overtakes it as the budget grows. The baselines' slow, only-partly-significant improvement under the same budget increase does not reproduce the anchor's advantage, so the effect is not explained by budget alone.

Window-average control on all five games. The RM$^+$ window column of Table 1 is the plain window average of RM$^+$ iterates at the same budget and seeds as the main grid (10 seeds, $T_u = \text{budget}/12$). Table 10 adds the paired one-sided t-test: RT-MCCFR$^+$'s windowed anchor is lower in

Table 9: Multi-budget sweep, 10 shared seeds per cell, budget-independent schedule. p: paired one-sided t-test for the decrease from 5×10^6 to 4×10^7 touches (lower is better); Δ: mean $\pm$ 95% half-width of the paired change from 2×10^7 to 4×10^7; a p near 1 means the output rises over the range.

	p ($5 \to 40$M)				Δ ($20 \to 40$M)	
Cell	RM$^+$ last	RT-MCCFR$^+$ last	RT-MCCFR$^+$ anch	RM$^+$ window	anch	last
Kuhn-ES	0.04	$< 10^{-6}$	$< 10^{-6}$	0.98	-0.0007 ± 0.0019	-0.0042 ± 0.0049
Kuhn-OS	0.17	$< 10^{-6}$	$< 10^{-6}$	0.02	-0.0019 ± 0.0032	-0.0115 ± 0.0070
Leduc-ES	0.17	7×10^{-4}	$< 10^{-6}$	0.94	-0.0045 ± 0.0197	-0.0376 ± 0.0652
Leduc-OS	0.004	2×10^{-4}	1.2×10^{-5}	0.02	$+0.0102 \pm 0.0483$	-0.0120 ± 0.0831

Table 10: Windowed-anchor control: RT-MCCFR$^+$'s anchor vs. the plain recent-window average of RM$^+$ iterates, mean over 10 shared seeds at the operating budget. Paired one-sided t-test; lower anchor is better.

Cell	n	RT-MCCFR$^+$ anch	w. RM$^+$ anch	paired p
Kuhn-ES	10	**0.0041**	0.0056	0.022
Kuhn-OS	10	**0.0085**	0.0128	0.001
Leduc-ES	10	**0.0708**	0.0908	0.006
Leduc-OS	10	**0.1857**	0.2291	0.014
1d6-ES	10	**0.1002**	0.1399	< 0.001
1d6-OS	10	**0.2542**	0.2698	0.033
G4-ES	10	**0.1113**	0.1124	0.47
G4-OS	10	**0.1226**	0.1985	0.0003
LD4-ES	10	**0.0376**	0.0391	0.33
LD4-OS	10	**0.0451**	0.0623	0.0001

all 10 cells, significantly in 8/10; the two non-significant cells (Goofspiel(4)-ES, LD(4)-ES) differ by less than 0.002. Window averaging alone therefore does not reproduce the anchor's advantage.

B.2 Average-policy comparison

The perturbation and anchor pull the trajectory toward a perturbed-NE solution, so the full-trajectory average of RT-MCCFR$^+$ is inflated relative to the plain baselines (Kuhn-ES 0.025 vs. 0.002 for IREG-PRM$^+$; Goofspiel(4)-ES 0.150 vs. 0.029). The windowed anchor mitigates this by averaging only the most recent window, and we report it as the method's output. The residual gap to the baselines' average is the price of the last-iterate regime.

B.3 Robustness to the anchor strength

The main text fixes $\mu = 1.0$ on every game and does not tune it per game. A robustness check at $\mu = 0.3$ is provided by the noise-floor scan (§B.4), which sweeps μ at fixed $\sigma = 0.5$. The floor increases as μ decreases: the ratio $D_{\text{floor}}(\mu{=}0.3)/D_{\text{floor}}(\mu{=}1.0)$ is 2.9 (ES) and 2.3 (OS), i.e. a local log-log exponent of -0.87 ± 0.11 (ES) and -0.70 ± 0.07 (OS), written as mean $\pm$ half-width of a bootstrap 95% interval that resamples seeds within each μ, against $1/0.3 = 3.3$ predicted by the $1/\mu$ heuristic of §5; both intervals exclude an exact exponent of -1. Over the full μ-axis $\{0.01, 0.03, 0.1, 0.3, 1\}$ the fitted exponent is -0.48 (ES) and -0.35 (OS), because the floor flattens below $\mu \approx 0.1$ (exponent -0.30 on $\{0.01, 0.03, 0.1\}$ under ES), the freeze regime of §A.5. The mechanism is unchanged at $\mu = 0.3$: the last iterate still settles, at a higher level consistent with the μ dependence above.

B.4 Noise-injection protocol

To probe the noise floor directly we run RT-MCCFR$^+$ on Kuhn with a fixed perturbation $\gamma = 0.05$, a fixed uniform anchor and no outer updates, inject additive zero-mean noise of variance σ^2 into the sampled counterfactual values (on top of the natural ES/OS variance), and measure D_{floor} as the mean over information sets of the squared ℓ_2 distance between the last iterate and the regularized

fixed point $\hat{x}^\star(\gamma, \mu)$ (computed by a 2×10^5-iteration full-feedback run at the same μ), averaged over the last eight of 24 log-spaced checkpoints of a 2×10^6-touch run. Grid: $\sigma \in \{0, 0.5, 1, 2, 4\} \times \mu \in \{1.0, 0.3\}$, plus a μ-axis sweep $\mu \in \{1.0, 0.3, 0.1, 0.03, 0.01\}$ at $\sigma = 0.5$, under both ES and OS, 5 seeds; the two grids share the cells $\sigma = 0.5$, $\mu \in \{1, 0.3\}$, so there are 130 distinct runs and every fit uses each run once. The reference solver feeds each minimizer the counterfactual value divided by the summed external reach of its information set, so its regularized fixed point is the fixed point of the expected per-visit ES dynamics, in which the regularizer $-\mu(\hat{x}_I - r_I)$ enters once per visit. Under OS the per-visit estimate is importance weighted and its conditional scaling differs, so the OS distances may carry a sampler-dependent offset; §B.5 measures it. On each σ-axis we fit $\log D_{\text{floor}} = c + p \log(\sigma^2 + \sigma^2_{\text{nat}})$ with the natural variance σ^2_{nat} as a free parameter: under ES, $p = 0.47 \pm 0.10$ with $\hat{\sigma}^2_{\text{nat}} = 0.26$ at $\mu = 1$ and $p = 0.45 \pm 0.15$ with $\hat{\sigma}^2_{\text{nat}} = 0.36$ at $\mu = 0.3$ (mean $\pm$ half-width of a bootstrap 95% interval); under OS the fitted natural variance is 4–6, in the middle of the injected range $\sigma^2 \in [0, 16]$, so the injected noise exceeds the natural one by at most a factor of four and does not provide a strongly noise-dominated range, and the exponent is not identified (the bootstrap interval reaches the fitting bound). The μ-axis exponents are in §B.3. The heuristic $D \propto \sigma_{\text{tot}}/(\mu\sqrt{t})$ predicts a slope of $1/2$ against the composite $(\sigma^2 + \hat{\sigma}^2_{\text{nat}})/\mu^2$; with $\hat{\sigma}^2_{\text{nat}}$ the mean of the two σ-axis estimates, the least-squares slope over the working range $\mu \geq 0.3$ is 0.44 ± 0.03 under ES and 0.35 ± 0.03 under OS (dashed line in Figure 2; bootstrap resamples seeds within each cell). Including the freeze-regime cells $\mu \leq 0.1$ lowers it to 0.28 (ES) and 0.19 (OS), since the floor stops responding to μ there.

B.5 Sampler dependence of the reference fixed point

The reference solver behind $\hat{x}^\star(\gamma, \mu)$ divides each counterfactual value by the summed external reach of its information set, which is the visit-conditional expectation of the ES estimator. Under OS the expected per-visit estimate is the value divided by the summed OS sampling reach. We measure how large the two effects are.

The ES expectation is exact. Freezing the executed strategy at $\hat{x}^\star$ and running 2×10^5 sampled iterations at $\gamma = 0.05$, the visit-weighted mean transformed update $g_I = \hat{u}_I - \mu(\hat{x}_I - r_I)$ has a positive-part instantaneous regret, $\max_a[\, g_I(a) - \langle g_I, \hat{x}_I\rangle]_+$ over the actions of positive base probability, of 0.0059 ± 0.0024 at the fixed point under ES with $\mu = 1$ and 0.0075 ± 0.0025 with $\mu = 0.3$, against full-feedback reference residuals of 0.0047 and 0.0064. The sampled residual is within one interval half-width of the reference residual at every μ of the grid (Table 11). The ratio of the norm of the sampled mean update to the normalized full-feedback value is 1.000 at the median information set, with range 0.98 to 1.20 at $\mu = 1$ and 0.99 to 1.00 at $\mu = 0.3$, so the ES visit-conditional expectation reproduces the quantity the reference solver divides by.

The OS offset sits below the floor. Under OS the same ratio reaches 2.4 at the least visited information sets, where the summed OS sampling reach is smallest relative to the external reach, and the sampled residual exceeds the reference one at every μ, by a factor between 1.7 and 3.6 and beyond one interval half-width at $\mu = 1$, of the same order as the OS sampling variance. The two reference fixed points are nevertheless close. Solving the OS-scaled game with the same protocol gives a fixed point at a mean squared distance of 3.1×10^{-8} from the ES reference at $\mu = 0.01$, 5.5×10^{-7} at $\mu = 0.03$, 6.8×10^{-6} at $\mu = 0.1$, 7.8×10^{-6} at $\mu = 0.3$ and 3.8×10^{-4} at $\mu = 1$. Re-measuring the 65 distinct OS runs of the grid (75 grid entries; the two scans share ten) against this reference changes their final distance by 2.3% in median absolute value, at most by a factor of 1.6 at $\mu = 1$, where the anchor is strongest; on these final distances the μ-axis log-log slope is -0.39 against the ES reference and -0.41 against the OS-scaled one, and the σ-axis exponent is not identified under either. The smallest reported OS floor at $\mu = 1$ is a factor of 12 above the reference gap, and at $\mu \leq 0.3$ the gap lies three to six orders of magnitude below the floor. The ES reference is therefore the right object for the ES axis that carries the fitted exponent, and the OS axis, which the fit does not identify, carries this offset at the strongest anchor.

B.6 Medium-scale cross-check

The LD(4) game is Liar's Dice in which each player rolls one private, uniformly random four-sided die; the highest face is wild and counts toward every other face, and a bid on the wild face counts exact matches only. A bid is a pair (quantity, face) with quantity in $\{1, 2\}$, ordered lexicographically,

Table 11: Reference-point residual at $\gamma = 0.05$: positive-part instantaneous regret of the visit-weighted mean transformed update with the executed strategy frozen at $\hat{x}^\star$ (2×10^5 sampled iterations), against the full-feedback residual of the 10^6-iteration reference. Mean $\pm$ bootstrap 95% half-width over visits.

	ES		OS	
μ	reference	sampled	reference	sampled
1	0.0047	0.0059±0.0024	0.0039	0.0141±0.0086
0.3	0.0064	0.0075±0.0025	0.0054	0.0111±0.0091
0.1	0.0069	0.0078±0.0019	0.0056	0.0104±0.0050
0.03	0.0083	0.0087±0.0022	0.0068	0.0117±0.0070
0.01	0.0086	0.0090±0.0018	0.0071	0.0123±0.0063

so there are eight bids. The first player must bid; afterwards a player either raises to a strictly higher bid or calls "Liar", and after the highest bid the only legal move is "Liar". At the call, the bidder wins if at least the bid quantity of matching dice is present, else the caller wins; the payoff is ± 1. The tree has $1{,}024$ information sets and $4{,}096$ decision nodes. We validate the implementation against OpenSpiel's Liar's Dice (Lanctot et al., 2019) in three ways: the legal action sets and terminal payoffs agree at all $4{,}096$ decision nodes; on 100 random behavioural profiles the expected utility agrees to 5.6×10^{-17} and NashConv to 4.4×10^{-16}, the latter also against a second best-response routine of ours written on the public bidding tree; and the game values of the two solvers agree to 4.4×10^{-16}. Cost per run: $2{\times}10^6$ touches in 31s, so $2{\times}10^7$ touches in $\approx$ 312s on one CPU core, well within the budget for the 20 LD(4) cells of the main grid.

B.7 OS VARIANCE BOUND

Under OS the estimator of a counterfactual value carries the importance weight $w = \pi^\sigma_{-i}(z)\, \pi^\sigma(h \to z)/q(z)$, where the traverser samples its own actions from the exploration mixture $q_I^{\mathrm{OS}} = (1 - \varepsilon)\hat{x}_I + \varepsilon\, \mathrm{Unif}$ (distinct from the base output q_I of Algorithm 1) with $\varepsilon = 0.6$ and the opponent's and chance's factors cancel between numerator and denominator. The weight is therefore bounded by $(|A|_{\max}/\varepsilon)^{\mathrm{depth}}$ by the sampler's exploration alone, for any $\gamma \geq 0$; the perturbation does not tighten this bound. What $\gamma > 0$ bounds is the played strategy, $\hat{x}_a \geq \gamma/|A|$, so the opponent's reach of every information set is at least $\pi_c\, (\gamma/|A|_{\max})^{\mathrm{depth}}$ under both samplers, with π_c the chance probability of the information set. This gives every information set with $\pi_c > 0$ a positive conditional visitation probability at each traversal (a bound on the probability, not on the realized frequency of a finite run) and keeps the played strategy in the interior. It does not by itself give the variance floor (A-var), which needs a payoff difference that varies with the sampled opponent action, nor the interiority (A-int) of the regularized target in base coordinates; both are assumptions of the one-infoset theorem and are checked numerically in §A.5. A quantitative variance bound in terms of γ, ε, the game depth and the reach distribution is left for the tree-level theory.

B.8 PFTRL VARIANTS

PFTRL (Masaka et al., 2025) applies a perturbation to the counterfactual values of an FTRL base. Its two named variants differ in the direction of the KL penalty: the reversed-KL variant uses $d(h, a) = (\sigma(a|x) - \pi(a|x))/\pi(a|x)$, which satisfies $\mathbb{E}_{a\sim\pi}[\delta(h, a)] = 0$ and therefore makes the perturbation local in expectation, whereas the KL variant uses $d(h, a) = \log(\sigma(a|x)/\pi(a|x))$ and carries non-vanishing downstream terms. We run both. Our KL arm localizes the perturbation to the information set and drops the downstream terms, so it is an approximation of the published KL variant and is labelled PFTRL-KL (loc.) throughout; the reversed-KL variant is exact under ES, where the implementation's numerical guard does not fire in any of our runs; under OS the guard fires and both variants receive the same per-visit scaling as RT-MCCFR$^+$ (§B.5). The reversed-KL arm is the stable one in our grid. Both arms use $\mu = 0.1$, uniform initial reference σ, and refresh σ to the current strategy every $T_\sigma = \mathrm{budget}/1000$ visits of the information set; the utilities they observe come from the same ES/OS estimator as every other arm. The comparison to RT-MCCFR$^+$ in §2 is against the reversed-KL variant.

B.9 CONTINUATION TO 8×10^7 TOUCHES

To test whether the anchor keeps improving past the budgets of Table 2, we rerun the four sweep cells to 8×10^7 touches with the same fixed window $T_u = 2 \times 10^7/12$, 10 seeds, and record the same run at 5, 10, 20, 40 and 80M touches, so the $40 \to 80$M change is a within-run paired difference. The matched-window outputs of RM$^+$, PRM$^+$ and IREG-PRM$^+$ (§B.14) run alongside. The ES cells and the OS matched-window arms reproduce across the batches we ran, to 2×10^{-15} for the ES cells and bit for bit at all five checkpoints for the OS matched-window arms. The OS RT-MCCFR$^+$ rows did not reproduce, so we report them from the local batch alone. Two batches of the same code and seeds separate at the tenth significant digit of the 5M checkpoint and reach the size of the seed spread by 10^7 touches, with no configuration field differing between them. The baseline utilities entering the minimizer have a single nonzero entry, so their inner product with the strategy is exactly rounded on any build; the regularized utility of RT-MCCFR$^+$ is dense, and its inner product is a two-term reduction whose rounding can depend on whether the build fuses multiply and add. The sampler draws the acting player's own action from a distribution that depends on the current policy, so a difference at that scale changes which action is drawn and the two trajectories separate. Tracing the seed-11 run, the first divergent operation is the inner product $\langle \tilde{u}, \hat{x} \rangle$ of the dense regularized utility in the RM$^+$ update, at the sixth iteration; re-running the tracer with an exactly rounded inner product, a compensated summation, makes the two environments agree bit for bit over the traced prefix. No anchor change from 40 to 80M is resolved at 10 seeds in any cell, so the eightfold ratios of Table 2 are not a rate: over the wider range from 5×10^6 to 4×10^7 the anchor falls in every cell and the improvement comes out of the exploration transient. The raw last iterate falls in all four cells, none resolved at 10 seeds. The windowed baselines do not hold their level: on Leduc-ES all three rise by 0.027–0.030 ($p \leq 0.053$), so at 8×10^7 touches the anchor is below the best windowed baseline in all four cells.

Table 12: Continuation runs, mean NashConv over 10 seeds at five touch counts (M $= 10^6$) and the paired $40 \to 80$M change (mean $\pm$ 95% half-width). Best windowed: the matched-window baseline with the lowest mean at 80M.

Cell	Output	5M	10M	20M	40M	80M	$40 \to 80$M
Kuhn-ES	RT-MCCFR$^+$ anchor	0.0767	0.0348	0.0041	0.0034	0.0042	$+0.0009 \pm 0.0017$
	RT-MCCFR$^+$ last	0.0812	0.0414	0.0153	0.0111	0.0100	-0.0011 ± 0.0029
	RM$^+$ last	0.1255	0.1356	0.0995	0.0882	0.0927	$+0.0045 \pm 0.0195$
	best windowed (IREG-PRM$^+$)	0.0036	0.0031	0.0048	0.0087	0.0056	-0.0031 ± 0.0027
Kuhn-OS	RT-MCCFR$^+$ anchor	0.0817	0.0333	0.0085	0.0066	0.0080	$+0.0014 \pm 0.0020$
	RT-MCCFR$^+$ last	0.0812	0.0406	0.0325	0.0192	0.0181	-0.0011 ± 0.0057
	RM$^+$ last	0.1251	0.1094	0.1107	0.1006	0.1172	$+0.0166 \pm 0.0295$
	best windowed (PRM$^+$)	0.0121	0.0097	0.0117	0.0110	0.0103	-0.0006 ± 0.0034
Leduc-ES	RT-MCCFR$^+$ anchor	0.1900	0.1172	0.0945	0.0899	0.0895	-0.0004 ± 0.0163
	RT-MCCFR$^+$ last	0.3560	0.3241	0.3073	0.2698	0.2483	-0.0215 ± 0.0421
	RM$^+$ last	0.5269	0.4622	0.4456	0.4841	0.3831	-0.1010 ± 0.0663
	best windowed (IREG-PRM$^+$)	0.1232	0.1282	0.1270	0.1274	0.1548	$+0.0274 \pm 0.0207$
Leduc-OS	RT-MCCFR$^+$ anchor	0.4052	0.3128	0.2359	0.2461	0.2666	$+0.0206 \pm 0.0538$
	RT-MCCFR$^+$ last	0.5618	0.4987	0.4023	0.3903	0.3557	-0.0345 ± 0.0595
	RM$^+$ last	0.6384	0.5840	0.4844	0.4580	0.4873	$+0.0293 \pm 0.0833$
	best windowed (RM$^+$)	0.3675	0.3020	0.2970	0.3098	0.3509	$+0.0411 \pm 0.0482$

B.10 FEEDBACK MODELS

Last-iterate guarantees in the literature fall into three feedback models with increasing variance. Under exact gradient feedback, optimistic variants and reward-transformed CFR converge in the last iterate. Under external sampling the counterfactual values are conditionally unbiased at the visited information sets and carry the variance of the sampled opponent and chance. Under outcome sampling the estimates are importance-weighted along a single trajectory and carry the largest variance. The telescoping identity behind average-to-last reductions (Cai et al., 2025b) is exact only under exact gradient feedback; under sampling it accumulates the per-round estimation noise. The reward transformation acts on the observed counterfactual value, so it applies unchanged under both samplers. The noise-floor scaling of §5 is the empirical face of this hierarchy: the RT-MCCFR$^+$ anchor

is higher under OS than under ES on every game in Table 1, and the measured floor grows with the variance.

B.11 Executable schedule semantics

The touch counter n of Algorithm 1 increases by one at every tree node entered by a traversal, terminal and chance nodes included, and one iteration is one traversal per player. After each iteration, if $n \geq \tau$, every anchor is refreshed once and γ is halved once, then $\tau \leftarrow \tau + T_u$; a single iteration never triggers two refreshes. The number of refreshes in a run is therefore not exactly $\lfloor B/T_u \rfloor$: with $n_{\text{fin}} \geq B$ the final touch count (the last iteration may overshoot B), it equals $\lfloor n_{\text{fin}}/T_u \rfloor$, i.e. $\lfloor B/T_u \rfloor$ or $\lfloor B/T_u \rfloor + 1$, whenever one iteration costs fewer than T_u touches (always the case in our runs, $T_u = B/12$); if T_u were smaller than one iteration, at most one refresh per iteration would occur. Strategies are queried at the moment an information set is visited, so a γ change takes effect at the next visit of every information set, for the traverser and the sampled opponent alike. The reported raw last iterate is the mixture (1) with the current γ at the final touch; the reported anchor is the one produced by the last completed refresh. The exponential-smoothing anchors of Table 4 are updated at the same refreshes, $r_I \leftarrow \beta r_I + (1-\beta)\,\mathrm{acc}_I/c_I$.

Window length. The window rows of Table 4 change T_u, which also changes the number of γ halvings. Table 13 holds the halving clock fixed at $B/12$ and varies only the anchor window, so every run ends at $\gamma = 2.4 \times 10^{-5}$.

Table 13: Window-length sweep with the halving clock fixed at $B/12$: last-iterate and anchor NashConv at the final touch count, mean over 10 seeds. The $B/12$ column is the anchor of Table 4.

	window $B/24$		window $B/6$		window $B/12$
Cell	last	anchor	last	anchor	anchor
Kuhn-ES	0.0136	0.0039	0.0263	0.0305	0.0041
Kuhn-OS	0.0276	0.0092	0.0388	0.0240	0.0085
Leduc-ES	0.3083	0.1026	0.2573	0.0736	0.0708
Leduc-OS	0.4058	0.2464	0.3257	0.1472	0.1857

B.12 Hyperparameters and computational fairness

The baselines are published algorithms with their own hyperparameters. We state our settings and the tuning we performed below. Payoffs are in the games' native units, so $U = 2$ on Kuhn, 13 on Leduc, 8 on Goofspiel(4) and 1 on the two Liar's Dice games, and NashConv is reported in the same units; μ and η are not rescaled across games. RT-MCCFR$^+$ uses $\gamma_0 = 0.1$ initial perturbation (halved at each anchor refresh), $\mu = 1.0$ anchor strength on every game, and $T_u = \mathrm{budget}/12$ window refreshes. RM$^+$, PRM$^+$, and IREG-PRM$^+$ use their published settings. PFTRL (Masaka et al., 2025) was tuned on Kuhn: we swept $\eta \in \{10^{-3}, 3{\times}10^{-3}, 10^{-2}, 3{\times}10^{-2}, 10^{-1}\}$ for both the KL and reversed-KL variants (2 seeds each) and selected $\eta = 10^{-3}$, the smallest value of the grid and the best or tied-best step size for both variants; larger step sizes were 2–3.7$\times$ worse. The same $\eta = 10^{-3}$ is used on all other games without further tuning, with perturbation strength $\mu = 0.1$ and a reference refresh every $\mathrm{budget}/1000$ visits (§B.8), so the two PFTRL cells where the method reaches NashConv ≥ 1 (Leduc-OS, Liar's Dice 1d6-ES) inherit the Kuhn-tuned step size. On the two cells where the implemented guard fires, Leduc-OS and Goofspiel(4)-OS, we reran the reversed-KL variant at $\eta = 10^{-4}$ and $\eta = 10^{-5}$ with 10 seeds: at $\eta = 10^{-4}$ the last outputs fall to 0.6022 and 0.5420 from 0.8867 and 0.7229 while the matched-window outputs rise to 0.4828 and 0.4399 from 0.2069 and 0.1307, and at $\eta = 10^{-5}$ the last outputs return to 1.0891 and 1.5413. The ordering against the anchor on these two cells holds at every step size tested. All arms share the identical touch budget and the same implementation and machines. We report touch counts, since the arms have different per-touch costs: relative to RM$^+$, the median solve time per run is $1.11\times$ for PRM$^+$, $2.8\times$ for IREG-PRM$^+$, $1.3\times$ for both PFTRL variants and $1.25\times$ for RT-MCCFR$^+$ (range 1.14–1.35 across the ten cells), so at equal touches IREG-PRM$^+$ and PFTRL consume more wall-clock than RT-MCCFR$^+$. Reported solver time excludes the anchor refresh, one vector copy per information

set every T_u touches, and the anchor adds one accumulate and one counter increment at each visited information set.

B.13 INITIALIZATION-INDEPENDENT SELECTION

We check whether the last iterate depends on the starting point on two textbook games with off-path information sets, solved with full feedback (4×10^4 iterations, $T_u = 2500$, $\mu = 1$, $\gamma_0 = 0.1$, $\gamma_{\min} = 10^{-6}$). In G1, player 1 chooses T (payoff 0) or B, after which player 2 chooses L (payoff 0) or R (payoff -1 to player 1); every profile with player 1 at T is an equilibrium and R weakly dominates L for player 2. In G2 the R branch leads to a second player-1 information set with payoffs $+2/-2$. The starts are three distinct regret states: zero regret, and a cumulative regret of 5 on the first or on the second action at every information set. Five initializations are run, and each realizes one of these states. On G1 the last iterate of the unregularized RM$^+$ leaves player 2's probability of R at 0.0, 0.5 or 1.0 depending on the start, while RT-MCCFR$^+$ reaches $P(R) = 1.0$ from all five starts with last-iterate NashConv 7.6×10^{-7}, which is $\gamma/2$ after the sixteen halvings of this run ($\gamma = 0.1/2^{16}$) and is the exploration floor itself; on G2 both variants reach the refined profile from every start. The anchor regularization removes the initialization dependence of last-iterate play where it appears.

B.14 MATCHED-WINDOW BASELINES

Table 1 compares RT-MCCFR$^+$'s windowed anchor against baseline last iterates, and the control column (RM$^+$ window) applies the same recent-window average to RM$^+$ only. To separate the window from the reward transformation we apply the identical window to the other minimizers (γ=μ=0 in the windowed rule, 10 seeds, $T_u = \text{budget}/12$): PRM$^+$, IREG-PRM$^+$ and the two PFTRL variants. Table 14 lists the ten cells with the signed paired difference between the RT-MCCFR$^+$ anchor and the best of the five windowed baselines of the cell, the RM$^+$ window included. The anchor is below the best windowed baseline in six cells and above it in four; at 10 seeds the two-sided paired test resolves one cell in each direction: 1d6-ES in favour of the anchor, and LD(4)-OS in favour of the windowed PFTRL-KL, the cell where the PFTRL-KL last iterate is already the lowest baseline and still decreasing at the budget. On 1d6-OS the best windowed baseline is the RM$^+$ window itself and the difference is not resolved. Against the three RM-family windows alone, the anchor is lower in eight cells and higher by at most 0.0016 in two (Kuhn-ES, LD(4)-ES). Selecting the best baseline on the same seeds leaves these differences as uncorrected paired two-sided tests, so we read them as descriptive.

Table 14: Matched-window outputs: NashConv of the window average (mean over 10 seeds) for each baseline given the RT-MCCFR$^+$ output rule (w. RM$^+$ is the RM$^+$ window column of Table 1), the RT-MCCFR$^+$ anchor, and the paired difference anchor minus best windowed baseline (mean $\pm$ 95% half-width, two-sided paired t-test). Bold: best windowed baseline. 1d6 = Liar's Dice 1d6, G4 = Goofspiel(4), LD4 = Liar's Dice 1d4.

Cell	RT-MCCFR$^+$ anchor	w. RM$^+$	w. PRM$^+$	w. IREG-PRM$^+$	w. PFTRL-KL (loc.)	w. PFTRL-RKL	anchor $-$ best	p
Kuhn-ES	0.0041	0.0056	**0.0038**	0.0048	0.0047	0.0044	$+0.0003 \pm 0.0015$	0.668
Kuhn-OS	0.0085	0.0128	0.0117	0.0107	0.0098	**0.0076**	$+0.0008 \pm 0.0036$	0.619
Leduc-ES	0.0708	0.0908	0.0882	**0.0832**	0.2075	0.1590	-0.0125 ± 0.0143	0.079
Leduc-OS	0.1857	0.2291	0.2501	0.2508	0.7821	**0.2069**	-0.0212 ± 0.0319	0.167
1d6-ES	0.1002	0.1399	0.1366	**0.1335**	1.1647	1.1907	-0.0333 ± 0.0063	< 0.001
1d6-OS	0.2542	**0.2698**	0.2811	0.2801	0.3156	0.3231	-0.0156 ± 0.0170	0.067
G4-ES	0.1113	**0.1124**	0.1258	0.1244	0.2171	0.1939	-0.0011 ± 0.0310	0.936
G4-OS	0.1226	0.1985	0.1944	0.1827	0.1624	**0.1307**	-0.0081 ± 0.0281	0.533
LD4-ES	0.0376	0.0391	0.0369	**0.0360**	0.2458	0.3148	$+0.0016 \pm 0.0030$	0.259
LD4-OS	0.0451	0.0623	0.0596	0.0603	**0.0301**	0.0430	$+0.0150 \pm 0.0112$	0.014

C DISCUSSION

What the last-iterate regime costs. RT-MCCFR$^+$ targets the last iterate, and its full-trajectory average is worse than the baselines' (Kuhn-ES: 0.025 against 0.002), because the perturbation pulls the trajectory toward a perturbed solution. The recent-window output is below the RM$^+$ last iterate in every cell and sweep budget and below every baseline last iterate at the budgets of Table 1; from

40 to 80M no change is resolved at 10 seeds (§B.9), in the regime where the acting policy must come from the recent trajectory.

Using the noise-floor scaling. The measured dependence on the ES noise axis, $D_{\text{floor}} \propto (\sigma^2 + \sigma^2_{\text{nat}})^{0.47\pm0.10}$, decreasing roughly as $1/\mu$, is an additive-noise result on Kuhn and we use it as a heuristic. It makes the sampling variance the primary lever: a lower-noise sampler (Li et al., 2026b) or a training-side control variate (Schmid et al., 2019) should lower the floor at about the square-root rate, at the price of per-touch cost. The anchor strength is a bounded lever: below the freeze threshold the floor stops responding, and raising it shrinks the floor but moves the regularized fixed point away from the equilibrium. Our operating point, $\mu = 1.0$ and $T_u = \text{budget}/12$, sits inside the responsive range; the floor at $\mu = 0.3$ stays within a factor of three of it (§B.3).

What is new. RT-MCCFR$^+$ combines a perturbation keeping every information set visited, an anchor regularization making local subproblems strongly concave, and truncated averaging that reduces output variance, all in an RM-based sampler where one constant-memory object per information set is both the regularizer and the output. RTCFR$^+$ (Meng et al., 2025a) proves last-iterate convergence of CFR$^+$ under full feedback, and the sampled last-iterate behaviour of the RM family has remained at the noise floor we measure. PFTRL (Masaka et al., 2025) is FTRL-based with a step size and reports its raw iterate; average-to-last reductions (Cai et al., 2025b) require exact feedback, and RT-MCCFR$^+$ is the base minimizer itself. §B.10 places RT-MCCFR$^+$ and the average-to-last reductions among the three feedback models of last-iterate guarantees.