WHEN LESS IS MORE: POLYLOGARITHMIC NASH-VALUE REGRET IN BANDIT MATRIX GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

In an unknown zero-sum matrix game, an adaptive opponent can choose which payoff column the learner observes. Under informed-bandit feedback, the opponent sees the learner's mixed row strategy before choosing a column; the learner then observes that column and one sampled payoff. Existing polylogarithmic Nash-value regret guarantees for this feedback model cover 2×2 games. We obtain such a guarantee for square games of any dimension with a unique, fully mixed equilibrium. Our algorithm, Instance-Certified Restricted-Face Learning (IC-RF), plays against the columns it has learned well enough. If the opponent withholds other columns, this restricted game yields a value surplus; if those columns appear, the learner gains samples. A multiscale ledger uses the surplus to finance learning across accuracy levels without knowing the game's separation parameters. An observable saddle certificate then starts local refinement. At any fixed confidence level, the resulting high-probability Nash-value regret is $O(\log T \log\log T)$, with a constant polynomial in the instance difficulty parameters. Experiments under best-response opponents show lower finite-horizon regret than established bandit baselines in most tested settings, while revealing slow certification in some higher-dimensional noisy games.

1 INTRODUCTION

Consider a fixed zero-sum game whose payoffs are initially unknown to one player. That player announces a distribution over rows; an opponent who sees the distribution chooses a column; only the payoff of the sampled row is observed. The opponent's choices determine both the learner's reward and which parts of the game can be estimated. An unplayed column may remain unknown for a long time, even while the learner collects thousands of observations elsewhere. This is the informed-bandit model of matrix-game learning (O'Donoghue et al., 2021).

We ask how quickly the learner can secure the value of the *mixed* game against such an opponent. Mixing matters: in games such as rock–paper–scissors, no single row protects the game value. Regret relative to the best pure row therefore answers a different question, and regret relative to the best row in hindsight can charge the learner for favorable choices made by the opponent. Nash-value regret instead compares play with the fixed value that a player knowing the matrix could guarantee. It can be negative when the opponent repeatedly offers columns on which the learner can earn more than that value.

Existing guarantees depend on what the learner observes and which value defines regret. Matrix–UCB handles arbitrary opponents but has square-root-order Nash-value regret (O'Donoghue et al., 2021). A logarithmic guarantee under bandit feedback uses the weaker pure-row security value (Ito et al., 2026). The related regimes of adversarial bandits, external regret, randomized optimism, and self-play are discussed below.

Maiti et al. (2026) obtain polylogarithmic Nash-value regret at arbitrary dimension with a noisy copy of the *entire* matrix each round, and for 2×2 games with bandit feedback. Their bandit argument exploits a tension: choosing a column reveals data; withholding it can raise the remaining game's value. Here the opponent controls which column appears, even during exploration.

The two-column argument becomes more intricate as the action set grows. If one column is withheld in a 2×2 game, only a single column remains. If the third column of the 3×3 identity game is

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.

withheld, the remaining two columns still form a game: their restricted value is $1/2$, versus $1/3$ for the full game, and obtaining the surplus requires a mixed row policy. Several hidden columns may subsequently arrive at different rates. The learner must use the surplus available today without spending more than it can earn when these columns return. This is the high-dimensional obstacle behind our question.

We develop Instance-Certified Restricted-Face Learning (IC-RF) to resolve this accounting problem. It tracks the value of observed column faces across accuracy scales, then enters local refinement only after an observable saddle certificate succeeds. For every fixed square game with a unique, fully mixed equilibrium, it attains high-probability $O(\log T \log\log T)$ Nash-value regret at fixed confidence against an adaptive informed opponent, with polynomial instance dependence. Synthetic and game-derived experiments show finite-horizon improvements in several settings, as well as slow certification in some noisy games.

The paper makes three contributions.

1. *A fast informed-bandit guarantee.* Against adaptive opponents, IC-RF achieves high-probability $O(\log T \log\log T)$ Nash-value regret on every fixed regular fully mixed $d \times d$ game, with polynomial instance and computational dependence and no known difficulty parameters.
2. *A higher-dimensional learn-or-earn argument.* We show how the opponent's column choices yield either information about the game or extra value on the observed columns. The analysis accounts for several columns being withheld and revealed at different rates, the obstacle to extending the two-column argument directly.
3. *Finite-horizon evidence.* Experiments on synthetic and game-derived matrices compare IC-RF with established bandit methods, test its components, and track when certification becomes effective. They identify settings with lower regret as well as noisy settings where certification remains slow.

2 Related Work

Bandit learning against adaptive opponents. EXP3 and implicit exploration protect against arbitrary reward sequences without estimating a fixed matrix (Auer et al., 2002; Neu, 2015). Matrix–UCB handles any opponent, including a best response to the current policy, with square-root Nash-value regret (O'Donoghue et al., 2021); randomized optimism has the same worst-case order (Lin, 2025). Gaussian-process learning exploits correlated unknown payoffs but controls external regret (Sessa et al., 2019). None establishes a fixed-instance polylogarithmic mixed-value rate in our protocol.

Ito et al. (2026) obtain logarithmic pure-strategy maximin regret under simultaneous action selection. Both their comparator and timing differ from ours. Bayesian Stackelberg learning also has a leader committing first, but its unknown follower types and optimal- commitment comparator define another problem (Bollini et al., 2026). In stochastic games, Jafarnia Jahromi et al. (2024) allow history-dependent opponents but learn transitions and bound Bayesian average-reward regret.

Instance-dependent fast Nash-value regret. On well-separated 2×2 games, myopic updates, repeated empirical equilibria, and Matrix–UCB can still incur square-root Nash-value regret (Maiti et al., 2026). Their fast result uses the surplus created when an opponent restricts play: it covers general $n \times m$ games with full-matrix observations, and 2×2 unique-equilibrium games with bandit observations, including cases outside our fully mixed class. Other fast approaches coordinate both players' exploration to identify a pure equilibrium (Yılmaz & Dimitrakakis, 2025), or impose minimum-play fairness and obtain $\widetilde{O}(T^{2/3})$ for the resulting mixed game (Akash & Gajane, 2026). Neither controls the sampled columns under the informed opponent used here.

Self-play and equilibrium identification. Equilibrium identification actively samples entries to output an accurate equilibrium (Maiti et al., 2023; 2024; Jiang & Zhang, 2026). Bandit self-play studies instance-dependent regret or last-iterate convergence when both players learn (Ito et al., 2025; Fiegel et al., 2025); other last-iterate results use full or noisy gradients (Abe et al., 2023). Here the row learner earns rewards online while an unrestricted opponent chooses the observed column. Time-varying games use changing or averaged equilibria (Cardoso et al., 2019); opponent shaping

accesses another agent's differentiable loss (Qiao et al., 2025). Both differ from a fixed unknown matrix.

3 SETTING

Interaction and feedback. The game has an unknown, fixed payoff matrix $A \in [0,1]^{d\times d}$; the learner chooses rows and the opponent chooses columns, with $d \geq 2$ actions each. Write $[d] = \{1,\dots,d\}$ and let Δ_d be the set of probability distributions on $[d]$. In round t, the learner first publishes a mixed row strategy $p_t \in \Delta_d$. After seeing p_t and the past interaction, the opponent chooses a column $j_t \in [d]$. The learner then draws $i_t \sim p_t$ and observes j_t and one payoff $Y_t \in [0,1]$; it does not observe the other entries of column j_t.

Let $\mathcal{F}_{t-1}$ be the history available before round t. The opponent and payoff noise may depend on this history. We assume

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}, p_t, j_t, i_t] = A_{i_t j_t}, \qquad 0 \leq Y_t \leq 1. \tag{1}$$

The learner knows the selected column after the round, but sees only one payoff in it. This is the informed-bandit feedback model.

Performance measure. For a mixed row strategy p, its expected payoff against column j is $p^\top A_{:j}$, where $A_{:j}$ is column j of A. The mixed game value is $v(A) = \max_{p\in\Delta_d} \min_{j\in[d]} p^\top A_{:j}$. We measure the learner's cumulative shortfall from this value:

$$R_T = \sum_{t=1}^{T} \{v(A) - p_t^\top A_{:j_t}\}. \tag{2}$$

Our goal is a small upper bound on R_T, not a small absolute value. Against a best-response opponent, $j_t \in \arg\min_j p_t^\top A_{:j}$, each summand is nonnegative, and zero is ideal. Against other opponents, negative regret means that cumulative expected payoff against the selected columns exceeds $Tv(A)$; it does not by itself establish a stronger worst-case security guarantee. The reference value remains that of the full game, even when the opponent repeatedly selects only a few columns.

Ito et al. use a different reference, the best pure-row security value $v_{\mathrm{pure}}(A) = \max_{i\in[d]} \min_{j\in[d]} A_{ij}$. Denote by PSMR_T the same cumulative shortfall as in (2) with $v_{\mathrm{pure}}(A)$ in place of $v(A)$. The two quantities differ on every interaction path by

$$R_T - \mathrm{PSMR}_T = T\{v(A) - v_{\mathrm{pure}}(A)\}. \tag{3}$$

Instance class. The fast-rate theorem concerns the following class of fixed square games.

Assumption 1 (Regular fully mixed game)**.** *The matrix A is square and has a unique Nash equilibrium $(x^\star, y^\star)$ with $x_i^\star > 0$ and $y_j^\star > 0$ for all i, j.*

Every row and column has positive equilibrium probability. We write $\rho = \min_i x_i^\star$ and $\tau = \min_j y_j^\star$ for their smallest row and column probabilities. For a nonempty set of columns $S \subseteq [d]$, let $v_S(A) = \max_{p\in\Delta_d} \min_{j\in S} p^\top A_{:j}$ be the value when the opponent can use only columns in S. The smallest increase in value after one column is removed is

$$\Delta_{\mathrm{del}} = \min_j \{v_{[d]\setminus\{j\}}(A) - v(A)\}.$$

Under Assumption 1, this gap is positive: omitting any column makes the row player's restricted game strictly easier. The algorithm does not know ρ, τ, or Δ_{del}. They enter the bound as instance parameters, along with the saddle system's condition number defined in Section 4.4.

4 METHOD

IC-RF first learns globally on well-estimated columns while sampling unresolved entries of columns the opponent has revealed. Withholding a column can raise the restricted game's value; revealing it supplies data. A certificate of saddle-point stability then permits local learning, where an anchor carries the regret guarantee and a gate controls the extra loss of an adaptive proposal. Algorithm 1 gives the policy; the multiscale ledger tracks credits and liabilities in the proof.

For each payoff entry (i, j), IC-RF maintains nested, anytime confidence bounds $[L_{t,ij}, U_{t,ij}]$ with half-width $r_{t,ij}$. Their adaptive-sampling construction uses time-uniform martingale bounds (Howard et al., 2021) and is given in the appendix.

4.1 Global restricted-face learning

While the opponent plays from columns S, the learner can secure $v_S(A)$. For example, two columns of the 3×3 identity game give value $1/2$, compared with $1/3$ for the full game. Exploiting this surplus in higher dimensions requires learning the restricted mixed strategy, even as new columns appear.

At global scale k, let $\varepsilon_k = 2^{-k}$ and define the active face

$$S_k(t) = \{j : \max_i r_{t,ij} \le \varepsilon_k\}.$$

Thus every entry of an active column has radius at most ε_k; counts and intervals persist across scales. A proper face has value at least $v(A) + \Delta_{\mathrm{del}}$. The learner must exploit this surplus while continuing to sample revealed but inactive columns.

Before round t, let $N_{ij}(t)$ and $\widehat{A}_{ij}(t)$ be the count and empirical mean of entry (i, j). At the start of each scale, a bounded prelude plays a row-maximin strategy of the Matrix–UCB matrix

$$U^{\mathrm{MUCB}}_{t,ij} = \widehat{A}_{ij}(t) + \sqrt{\frac{2\log(2d^2T^2)}{N_{ij}(t) \vee 1}}.$$

After the prelude, let $\mathcal{B}_k(t)$ contain each row with an entry of radius above ε_k in a previously seen column outside $S_k(t)$. The learner assigns a sampling floor only to these blocking rows. Its feasible set is $\mathcal{P}_k(t) = \{p \in \Delta_d : p_i \ge \varepsilon_k/d \text{ for } i \in \mathcal{B}_k(t)\}$. If the active face is nonempty, the global policy is

$$p_t \in \arg\max_{p \in \mathcal{P}_k(t)} \min_{j \in S_k(t)} p^\top L_{t,:,j}. \tag{4}$$

Here L_t contains the lower confidence endpoints. The maximization protects value on the active face; the row floor lets repeatedly played inactive columns become estimable. A column never seen before creates one first-encounter cost, after which its unresolved rows can enter $\mathcal{B}_k(t)$. If $S_k(t)$ is empty, the learner plays uniformly.

The scale schedule fixes how long the learner can pay for exploration. Set $K_T = 1 + \lceil \log_2(T \vee 2) \rceil$, where $a \vee b = \max\{a, b\}$. The factor $\mathcal{L}$ is defined in (12); it depends only on d, T, δ and is $O(\log(d/\delta) + \log(1 + \log(T \vee 2)))$. We use

$$N_k = \left\lceil \frac{9\mathcal{L}}{\varepsilon_k^2} \right\rceil, \qquad Q_k = 8d^2 N_k,$$
$$W_k = \frac{160 d^2 \mathcal{L}}{\varepsilon_k^3}, \quad H_k = \left\lceil \frac{8W_k}{\varepsilon_k} \right\rceil.$$

The prelude lasts at most Q_k actions, H_k is the scale clock, and W_k is an upper bound on liability used only in the analysis. The proof charges the cost of a finer scale against the surplus earned on a completed proper face. Since $W_{k+1} = 8W_k$, these charges form a geometric ledger across scales. The policy uses d, T, δ and its observed history; the game value and deletion gap enter only the bound. The prelude enters this ledger argument through its length Q_k.

4.2 Observable certification and local play

When local learning is justified. A small confidence radius alone does not tell the learner whether the estimated saddle point remains stable for every game still compatible with the observations. IC-RF therefore tests the entire rectangle $[L, U]$ when the active face contains every column. Let $\varepsilon = \max_{i,j} r_{t,ij}$. Interval resolvents certify positive equilibrium row and column floors $\underline{\rho}, \underline{\tau}$. Deleting each column j from the lower matrix supplies a restricted maximin policy r_j and a conservative surplus $\delta_j = v_{-j}(L) - v(U)$. The certificate also checks that the local strategy set has row mass at least $\underline{\rho}/2$, and that interpolating toward each r_j stays in that set. The resulting reserve $S_{\mathrm{res}} = \min_j \alpha_j \delta_j$ must dominate the second-order error $\mathcal{G}^{(2)}\varepsilon^2$ by a fixed margin. Acceptance freezes a local model; rejection advances the global scale. Anytime validity covers this data-dependent test, whose thresholds are in Appendix F.

The local anchor. For each j, the restricted policy r_j serves as a comparator in the proof. It exposes the local learn-or-earn step. If some column appears rarely during a local epoch, an interpolated restricted policy can earn enough surplus to cover the epoch's estimation error. If every column appears frequently, the observations shrink the confidence rectangle and end the epoch. The local regret bound is carried by an *anchor* that combines the certified center with projected FTRL. Its expert weights persist across confidence refinements and restart only when their dyadic cap grows.

Let $\|B\|_{\max}$ denote the largest absolute entry of a matrix B. For an accepted epoch e, write M_e and R_e for the frozen midpoint and radius matrices, $\widehat{x}_e$ for the certified center, $\varepsilon_e = \|R_e\|_{\max}$, and $\mathcal{R}_e$ for the certified local-radius multiplier. Set $a_e = \mathcal{R}_e\varepsilon_e$ and define the certified local set $\mathcal{X}_e = \{x \in \Delta_d : \|x - \widehat{x}_e\|_2 \le a_e\}$. It contains the full-game equilibrium strategy and the interpolated restricted comparators needed in the two cases above.

The local action uses two experts: the frozen certified center $c_{t,1} = \widehat{x}_e$ and centered FTRL

$$c_{t,2} = f_t = \arg\max_{x \in \mathcal{X}_e} \left\{ x^\top \sum_{s<t,\, s\in e} M_{e,:,j_s} - \frac{\|x - \widehat{x}_e\|_2^2}{2\varepsilon_e^2} \right\}.$$

Writing $\pi_{t,a} = w_{t,a}/(w_{t,1} + w_{t,2})$, the anchor is $h_t = \sum_{a=1}^{2} \pi_{t,a} c_{t,a}$. Its exponential weights use the learning rate λ stored for the current dyadic-cap segment:

$$w_{t+1,a} = w_{t,a} \exp\{\lambda c_{t,a}^\top M_{e,:,j_t}\}, \qquad a \in \{1, 2\}.$$

When a new cap segment begins, its initial FTRL weight is proportional to $\max\{\varepsilon_e, 1/[4(T \vee 2)]\}$. Keeping the weights between certified refinements avoids paying a fresh expert-comparison cost at each epoch.

Adaptive proposal and publication gate. Two projected, scale-free two-expert controllers form an adaptive proposal: $s_t = \sum_a \omega_{t,a} c_{t,a}$ and $q_t = \nu_{t,1} h_t + \nu_{t,2} s_t$. For either controller with weights $\xi \in \Delta_2$, cumulative squared variation V, and observed proxy payoffs $u \in \mathbb{R}^2$, its exact update is

$$\widetilde{u} = u - \tfrac{1}{2}(u_1 + u_2)\mathbf{1}, \qquad V^+ = V + \|\widetilde{u}\|_2^2, \qquad \xi^+ = \Pi_{\Delta_2}\big(\xi + \sqrt{2/V^+}\,\widetilde{u}\big),$$

with $\xi^+ = \xi$ when $V^+ = 0$. The inner controller uses $u_a = c_{t,a}^\top M_{e,:,j_t}$; the outer one uses $u = (h_t^\top M_{e,:,j_t}, s_t^\top M_{e,:,j_t})$. Here Π_{Δ_2} is Euclidean projection onto Δ_2. With $P = I - \mathbf{1}\mathbf{1}^\top/d$, define the observable one-epoch publication liability

$$H_e^{\mathrm{pay}} = \max_j \|P M_{e,:,j}\|_2, \qquad J_e^{\mathrm{rad}} = \max_j \|R_{e,:,j}\|_2, \qquad B_e = 2a_e(H_e^{\mathrm{pay}} + J_e^{\mathrm{rad}}).$$

At the start of an accepted epoch, the gate sets $Z_e = B_e$. After observing column j_t, it computes a lower bound on the proposal's payoff advantage over the anchor:

$$\underline{g}_t = (q_t - h_t)^\top M_{e,:,j_t} - |q_t - h_t|^\top R_{e,:,j_t}.$$

The gate publishes $p_t = (1 - \alpha_t)h_t + \alpha_t q_t$ with $\alpha_t = \min\{1, Z_t/B_e\}$ (taking $\alpha_t = 1$ if $B_e = 0$), then updates $Z_{t+1} = Z_t + \alpha_t \underline{g}_t$. Each published local prefix adds at most B_e to the anchor's regret.

4.3 Complete algorithm

Transitions are checked before the next policy is announced and consume no round. In global mode, a complete face is certified before the scale clock is checked; rejection or expiration moves to a finer scale. In local mode, the first halving of the frozen radius triggers recertification. Acceptance refreshes the local model, while rejection returns to the first global scale finer than the current radius. If no such scale remains, IC-RF uses its terminal policy. The counter ℓ records actions at the current global scale and resets on advancement.

Algorithm 1 IC-RF (Instance-Certified Restricted-Face Learning)

Require: horizon T and confidence level δ.

Initialize entry estimates and intervals; $k \leftarrow 0$, $\ell \leftarrow 0$, mode $\leftarrow$ Global.
for $t = 1, \dots, T$ **do**
  Apply the pre-action transitions above; reset $\ell \leftarrow 0$ on a new global scale.
  if the uniform or exact-matrix terminal branch has been reached **then**
    Publish its fixed terminal policy; observe (j_t, i_t, Y_t); **continue**.
  end if
  if mode $=$ Global **then**
    if $\ell < Q_k$ **then**
      $p_t \in \arg\max_{p \in \Delta_d} \min_{j \in [d]} p^\top U^{\mathrm{MUCB}}_{t,:,j}$.
    else if $S_k(t) = \varnothing$ **then**
      $p_t \leftarrow \mathbf{1}/d$.
    else
      $p_t \in \arg\max_{p \in \mathcal{P}_k(t)} \min_{j \in S_k(t)} p^\top L_{t,:,j}$.
    end if
  else
    Compute (f_t, h_t, s_t, q_t) from the local updates above.
    $p_t \leftarrow (1 - \alpha_t) h_t + \alpha_t q_t$.
  end if
  Announce p_t; opponent chooses j_t; sample $i_t \sim p_t$; observe (j_t, Y_t).
  $N_{i_t j_t} \leftarrow N_{i_t j_t} + 1$; update $(\widehat{A}, L, U)$.
  if mode $=$ Local **then**
    $Z_{t+1} \leftarrow Z_t + \alpha_t \underline{g}_t$.
    $w_{t+1,a} \leftarrow w_{t,a} e^{\lambda c_{t,a}^\top M_{e,:,j_t}}$; update (ω, ν) by the projected rule above.
  else
    $\ell \leftarrow \ell + 1$; recompute $S_k(t+1)$.
  end if
end for

Appendix B specifies initialization, cap changes and terminal branches. Maximin ties are resolved lexicographically.

4.4 Theoretical results and scope

The regret bound depends on how much value is gained by deleting a column, how much equilibrium probability each action receives, and the conditioning of the saddle equations. The first two quantities were defined in Section 3. For a square matrix B, define

$$\mathsf{B}_x(B) = \begin{bmatrix} B^\top & -\mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix}, \qquad \mathsf{B}_y(B) = \begin{bmatrix} B & -\mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix}.$$

Under Assumption 1, both augmented systems are invertible (Proposition 1 below). Define $\kappa = \max_{q \in \{x,y\}} \|\mathsf{B}_q(A)^{-1}\|_\infty$, where the matrix infinity norm is the maximum absolute row sum. The quantities $(\rho, \tau, \Delta_{\mathrm{del}}, \kappa)$ measure the difficulty of a fixed instance; none is supplied to IC-RF.

Let $\overline{\kappa} = \max\{1, \kappa\}$ and let Γ_{cert} be the sufficient entrywise radius for certificate acceptance. Up to universal constants it is controlled by

$$\min\left\{ (\overline{\kappa} d)^{-1}, \frac{\rho}{\overline{\kappa}\sqrt{d}}, \frac{\tau}{\overline{\kappa}}, \frac{\Delta_{\mathrm{del}}}{\overline{\kappa} d} \right\}.$$

Theorem 1 (Instance-dependent regret)**.** *Let A be any fixed square $d \times d$ game with a unique Nash equilibrium that is fully mixed for both players (Assumption 1), and suppose the feedback satisfies Equation* (1)*. IC-RF has polylogarithmic Nash-value regret against every adaptive informed-bandit opponent. More precisely, for every $T \geq 2$ and $\delta \in (0, 1)$, there is a constant C_A, polynomial in the fixed game's difficulty parameters and independent of T and δ, such that, with probability at least $1 - \delta$,*

$$R_T \leq C_A \log(T \vee 2) \left\{ \log \frac{d}{\delta} + \log\big(1 + \log(T \vee 2)\big) \right\}.$$

In particular, at fixed confidence this is $O(C_A \log T \log\log T)$. An explicit bound is

$$R_T \le C\left[\frac{d^2\mathcal{L}}{\Gamma_{\mathrm{cert}}^3} + \frac{\sqrt{d}\log(T\vee 2)}{\Gamma_{\mathrm{cert}}} + (K_T+1)\left\{\frac{1}{\Gamma_{\mathrm{cert}}^2} + \frac{d\mathcal{L}}{\Gamma_{\mathrm{cert}}\rho\Delta_{\mathrm{del}}} + \frac{\sqrt{d}}{\Gamma_{\mathrm{cert}}}\right\}\right], \tag{5}$$

where C is universal and $K_T = O(\log T)$. No instance parameter in the bound is given to the learner.

Fixing A keeps C_A constant as T grows; substituting the bound on $\mathcal{L}$ gives the simplified rate. The leading cubic term combines $O(\mathcal{L}/\varepsilon^2)$ samples per blocking entry with its ε/d sampling floor: each inactive column needs at most $O(d\mathcal{L}/\varepsilon^3)$ appearances. The remaining terms pay for local comparison, refresh, and publication. A small Γ_{cert} can therefore delay certification despite the fast asymptotic rate. Appendix F.2 calibrates these quantities for identity games.

Proposition 1 (Consequences of the structural assumption)**.** *Under Assumption 1, $\Delta_{\mathrm{del}} > 0$ and both augmented saddle systems are nonsingular.*

Deleting any column strictly raises the row player's guaranteed value, and the saddle equations are locally invertible. Hence the deletion gap and condition number in Theorem 1 describe the difficulty of the game; they are not extra assumptions. The proof is in Appendix J.

This class contains identity games and shifted-and-scaled rock–paper–scissors, and is open: Proposition 9 gives explicit stability neighborhoods. For identity games, a certificate obstruction at radius $1/d$ and a sufficient radius of order d^{-2} leave a factor-d resolution gap (Proposition 5); this concerns certification, not regret. Allowing the game to vary with the horizon gives the following worst-case limit.

Theorem 2 (Dimension-dependent worst-case lower bound)**.** *For every $d \ge 2$, $T \ge 1$, and possibly randomized learner, there is a $d \times d$ game $A_{d,T}$ with a unique, fully mixed equilibrium and a deterministic opponent for which the learner's expected regret is at least $\frac{1}{64}\min\{T, \sqrt{dT}\}$.*

The hard game depends on d, T, so this problem-class lower bound is compatible with fixed-instance polylogarithmic regret. It does not establish IC-RF's worst-case upper bound. See Appendix K for the proof.

Under the finite-input model, a rationalized IC-RF controller retains the same regret orders with changed universal constants. Its expected running time and memory are polynomial in T, d, the observation bit length, and $\log(1/\delta)$; the construction and rounding analysis are in Appendix N.

5 Experiments

5.1 Games, feedback, and baselines

We evaluate IC-RF on four games: rock–paper–scissors (RPS) and empirical OpenSpiel meta-games from Colonel Blotto, Goofspiel, and Kuhn poker (Lanctot et al., 2019). Each game has a deterministic mean-payoff and a noisy-feedback condition, giving eight fixed settings. The noisy observations are Bernoulli rewards or resampled episode returns, as appropriate to the game. The opponent chooses a best response to the learner's announced mixed strategy on each round. We run $640{,}000$ rounds with eight paired seeds per setting; Kuhn is a near-boundary stress case.

Matrix–UCB (O'Donoghue et al., 2021) is the closest baseline: it uses an optimistic payoff matrix and has the same informed timing and mixed-value objective. CoEBL (Lin, 2025) uses randomized optimism, while EXP3 (Auer et al., 2002) is an adversarial bandit method that does not exploit a fixed matrix. We also plot Maximin–UCB (Ito et al., 2026), whose theoretical comparator is the best pure row. A Nash-security oracle that knows A and plays an equilibrium strategy has zero regret against the best-response opponent; this reference is omitted from the plots. Maximin–UCB is shown for context but excluded from the ranking of methods designed for mixed-value regret. Whenever a method solves a maximin linear program, we use the same backend and tie rule.

Our primary measure is signed Nash-value regret R_T from Equation (2); lower is better. We also record the cumulative security gap $G_T = \sum_{t=1}^{T}\{v(A) - \min_j p_t^\top A_{\cdot j}\}$ and maximum prefix debt $D_T = \max_{s \le T} R_s$. Under current-policy best response, $R_T = G_T$; D_T reveals losses that a final signed sum could hide. The main runs use the time-uniform confidence sequence and proof schedule. We report an $\mathcal{L} = 1$ schedule only as an appendix diagnostic.

5.2 Learning curves and final regret

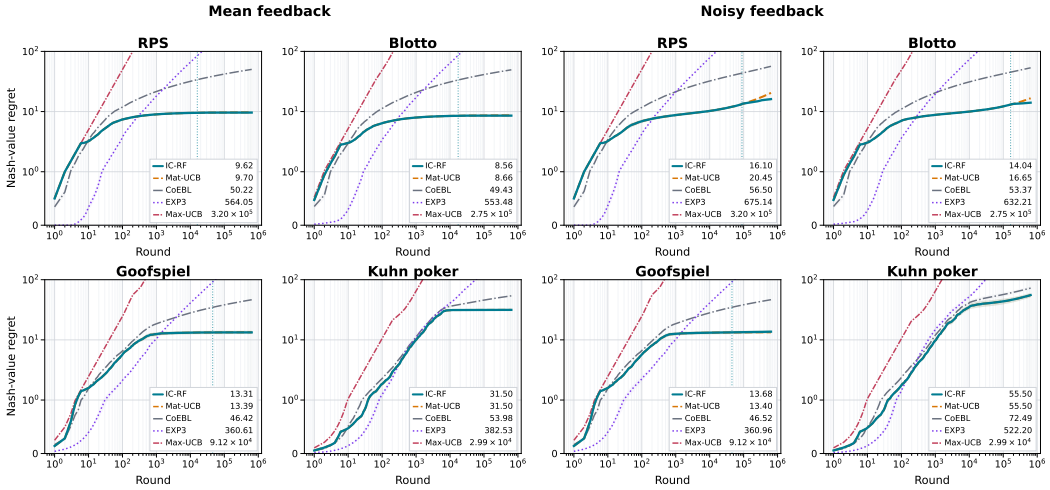


Figure 1: Best-response regret under mean (left) and noisy (right) feedback. The ordinate ends at 100, so higher trajectories are clipped. The key inside each panel reports every method's full final regret beside its name. Mat-UCB abbreviates Matrix–UCB, and Max-UCB abbreviates Maximin–UCB. Shading is one standard error; blue dotted lines mark median local entry.

Figure 1 and Table 1 show that IC-RF beats Matrix–UCB in five settings, ties twice on Kuhn without certification, and loses on resampled Goofspiel. It beats CoEBL and EXP3 in all eight. Gains are largest on noisy RPS and resampled Blotto, and smaller on the deterministic games.

Table 1 reports means (standard errors). For IC-RF minus Matrix–UCB, unadjusted paired 95% confidence intervals across matched seeds, expressed as mean difference $\pm$ half-width, are -4.347 ± 1.254 on Bernoulli RPS, -2.609 ± 0.530 on resampled Blotto, and 0.284 ± 0.189 on resampled Goofspiel.

Table 1: **Final cumulative signed Nash-value regret** at $T = 640{,}000$ (eight paired seeds; entries are mean (standard error); lower is better). Bold marks the lowest mean, including ties, and does not denote statistical significance. Mean and episode denote deterministic mean-payoff and resampled-return feedback. Maximin–UCB targets the pure-row value and is excluded from the ranking.

Game	Feedback	IC-RF (ours)	Matrix–UCB	CoEBL	EXP3	Maximin–UCB†
RPS	mean	**9.62 (0.31)**	9.70 (0.32)	50.22 (0.31)	564.05 (1.14)	3.20×10^5
RPS	Bernoulli	**16.10 (0.90)**	20.45 (0.39)	56.50 (0.36)	675.14 (1.71)	3.20×10^5
Blotto	mean	**8.56 (0.29)**	8.66 (0.29)	49.43 (0.43)	553.48 (0.89)	2.75×10^5
Blotto	episode	**14.04 (0.35)**	16.65 (0.27)	53.37 (0.47)	632.21 (0.86)	2.75×10^5
Goofspiel	mean	**13.31 (0.86)**	13.39 (0.86)	46.42 (0.59)	360.61 (1.01)	9.12×10^4
Goofspiel	episode	13.68 (0.83)	**13.40 (0.83)**	46.52 (0.59)	360.96 (1.20)	9.12×10^4
Kuhn poker	mean	**31.50 (0.85)**	**31.50 (0.85)**	53.98 (1.20)	382.53 (11.93)	2.99×10^4
Kuhn poker	episode	**55.50 (4.77)**	**55.50 (4.77)**	72.49 (4.77)	522.20 (11.78)	2.99×10^4

5.3 COMPONENT ABLATIONS

On the same eight settings, eight paired seeds, and $T = 640{,}000$, nine of ten component or controller controls have higher fixed-suite mean regret than IC-RF under best response by simultaneous 95% paired intervals; anchor-only is inconclusive. Removing the bounded prelude hurts most, while replacing a local expert or removing the publication gate also raises regret. Table 2 reports the average increase in R_T caused by each change, across the eight settings. Positive values favor complete IC-RF.

Table 2: **Best-response ablations** at $T = 640{,}000$. Entries are the mean increase in $R_T \pm$ the paired 95% confidence-interval half-width, Bonferroni-adjusted across ten comparisons. Positive values favor full IC-RF.

Change to IC-RF	**Increase in R_T**	**Change to IC-RF**	**Increase in R_T**
No bounded prelude	$135\,000 \pm 13\,000$	Center only	173.241 ± 85.504
Hoeffding intervals	0.444 ± 0.217	FTRL only	12.239 ± 0.276
Fixed cert. radius	0.297 ± 0.249	Reset local master	0.077 ± 0.071
Sharp-only cert.	0.868 ± 0.275	Anchor only	0.012 ± 0.052
Matrix–UCB after cert.	0.867 ± 0.274	Ungated proposal	3.031 ± 2.063

Both switching-face schedules favor IC-RF in signed regret across the four tested controller comparisons. Uniform-row exploration instead lowers the security gap, though it raises signed regret. Complete paired results and intervention details are in Appendix A.5.

6 CONCLUSION

We studied how a learner can secure the value of an unknown zero-sum game when an adaptive opponent controls which payoff columns are observed. Our method, IC-RF, uses the extra payoff available on restricted faces to fund exploration, then enters local refinement through an observable certificate. This approach yields polylogarithmic Nash-value regret at fixed confidence in square games of arbitrary dimension with a unique, fully mixed equilibrium, with polynomial dependence on instance parameters that the learner need not know. A dimension-dependent minimax lower bound characterizes the problem's worst-case difficulty. Experiments on synthetic and OpenSpiel-derived games show that IC-RF attains the lowest or tied-lowest final mean regret among the compared methods in most tested best-response settings.

Limitations. Rectangular games, boundary equilibria, and multiple equilibria remain outside the fast theorem. Large instance constants can delay certification, especially under noise, and some experiments show little finite-horizon gain. The optimal dependence on regularity parameters and IC-RF's uniform worst-case upper bound remain open. Support-adaptive certification and earlier local entry are natural next steps.

ETHICS STATEMENT

This theoretical study uses no human subjects or personal data. OpenSpiel experiments use simulated trajectories.

REPRODUCIBILITY STATEMENT

The accompanying artifact contains proofs, IC-RF and baseline code, frozen manifests, per-seed trajectories, plots, and OpenSpiel meta-game construction.

AI USE STATEMENT

Generative AI tools were used to improve the clarity and grammar of the manuscript. The authors reviewed the edited text and take responsibility for the final content.

REFERENCES

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, Kentaro Toyoshima, and Atsushi Iwasaki. Last-iterate convergence with full and noisy feedback in two-player zero-sum games. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7999–8028. PMLR, 2023. URL https://proceedings.mlr.press/v206/abe23a.html.

S. Akash and Pratik Gajane. Fairness in two-player zero-sum games with bandit feedback. *arXiv preprint arXiv:2606.01159*, 2026. URL https://arxiv.org/abs/2606.01159.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

H. Beeck. Zur Problematik der Hüllenbestimmung von Intervallgleichungssystemen. In *Interval Mathematics: Proceedings of the International Symposium, Karlsruhe, 1975*, volume 29 of *Lecture Notes in Computer Science*, pp. 150–159. Springer, 1975. doi: 10.1007/3-540-07170-9_12. URL https://doi.org/10.1007/3-540-07170-9_12.

Matteo Bollini, Francesco Bacchiocchi, Samuel Coutts, Matteo Castiglioni, and Alberto Marchesi. Learning in Bayesian Stackelberg games with unknown follower's types. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.00771.

Adrian Rivera Cardoso, Jacob Abernethy, He Wang, and Huan Xu. Competing against Nash equilibria in adversarially changing zero-sum games. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 921–930. PMLR, 2019.

Côme Fiegel, Pierre Menard, Tadashi Kozuno, Michal Valko, and Vianney Perchet. The harder path: Last iterate convergence for uncoupled learning in zero-sum games with bandit feedback. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17131–17152. PMLR, 2025. URL https://proceedings.mlr.press/v267/fiegel25a.html.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

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

Shinji Ito, Haipeng Luo, Arnab Maiti, Taira Tsuchiya, and Yue Wu. Adversarial learning in games with bandit feedback: Logarithmic pure-strategy maximin regret. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 3666–3692. PMLR, 2026.

Mehdi Jafarnia Jahromi, Rahul Jain, and Ashutosh Nayyar. A Bayesian learning algorithm for unknown zero-sum stochastic games with an arbitrary opponent. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3880–3888. PMLR, 2024. URL https://proceedings.mlr.press/v238/jafarnia-jahromi24a.html.

Jiashuo Jiang and Mengxiao Zhang. An LP-based approach for bilinear saddle point problem with instance-dependent guarantee and noisy feedback. *arXiv preprint arXiv:2602.12513*, 2026. URL https://arxiv.org/abs/2602.12513.

Irving Kaplansky. A contribution to von Neumann's theory of games. *Annals of Mathematics*, 46(3): 474–479, 1945. doi: 10.2307/1969164.

Leonid G. Khachiyan. A polynomial algorithm in linear programming. *Doklady Akademii Nauk SSSR*, 244(5):1093–1096, 1979.

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

Shishen Lin. Randomised optimism via competitive co-evolution for matrix games with bandit feedback. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 8894–8902, 2025. doi: 10.24963/ijcai.2025/989. URL https://www.ijcai.org/proceedings/2025/989.

Arnab Maiti, Kevin Jamieson, and Lillian Ratliff. Instance-dependent sample complexity bounds for zero-sum matrix games. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9429–9469. PMLR, 2023. URL https://proceedings.mlr.press/v206/maiti23a.html.

Arnab Maiti, Ross Boczar, Kevin Jamieson, and Lillian Ratliff. Near-optimal pure exploration in matrix games: A generalization of stochastic bandits & dueling bandits. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2602–2610. PMLR, 2024. URL https://proceedings.mlr.press/v238/maiti24a.html.

Arnab Maiti, Kevin Jamieson, and Lillian J. Ratliff. On the limitations and possibilities of Nash regret minimization in zero-sum matrix games under noisy feedback. In Ron Lavi and Jie Zhang (eds.), *Algorithmic Game Theory*, volume 15953 of *Lecture Notes in Computer Science*, pp. 3–20, Cham, 2026. Springer Nature Switzerland. doi: 10.1007/978-3-032-03639-1_1. URL https://doi.org/10.1007/978-3-032-03639-1_1.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Arnold Neumaier. *Interval Methods for Systems of Equations*, volume 37 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1990.

Brendan O'Donoghue, Tor Lattimore, and Ian Osband. Matrix games with bandit feedback. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 279–289. PMLR, 2021. URL https://proceedings.mlr.press/v161/o-donoghue21a.html.

Xinyu Qiao, Yudong Hu, Congying Han, Weiyan Wu, and Tiande Guo. Preference-based opponent shaping in differentiable games. *Machine Learning*, 114(12):292, 2025. doi: 10.1007/s10994-025-06893-z.

Siegfried M. Rump. Solving algebraic problems with high accuracy. In Ulrich W. Kulisch and Willard L. Miranker (eds.), *A New Approach to Scientific Computation*, pp. 51–120. Academic Press, 1983. doi: 10.1016/B978-0-12-428660-3.50010-0.

Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, and Andreas Krause. No-regret learning in unknown games with correlated payoffs. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/685217557383cd194b4f10ae4b39eebf-Abstract.html.

Lloyd S. Shapley and R. N. Snow. Basic solutions of discrete games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume I*, volume 24 of *Annals of Mathematics Studies*, pp. 27–35. Princeton University Press, 1950.

Elif Yılmaz and Christos Dimitrakakis. Two-player zero-sum games with bandit feedback. *arXiv preprint arXiv:2506.14518*, 2025. doi: 10.48550/arXiv.2506.14518. URL https://arxiv.org/abs/2506.14518.

APPENDIX

A EXPERIMENTAL DETAILS

The eight best-response trajectories are shown in Figure 1. Unless stated otherwise, all results use the theorem-faithful schedule. The diagnostic schedule sets the multiplier $\mathcal{L} = 1$, changing scale clocks, promotion budgets, and local liability thresholds but not the empirical-Bernstein confidence sequence. It has no regret guarantee, and its results are reported separately.

In descriptive results, an entry $m\,(s)$ reports the mean m across seed runs and its standard error s, computed as the sample standard deviation divided by the square root of the number of seeds. For method comparisons, we first form within-seed regret differences and use a Student-t confidence interval for their mean. In the ablation tables, $\pm$ instead gives that paired mean difference plus or minus the confidence- interval half-width; the captions specify any multiple-comparison adjustment.

A.1 SYNTHETIC REGULAR FAMILY

The scaling family is

$$A_{d,a} = \frac{1}{2}\mathbf{1}\mathbf{1}^{\mathsf{T}} + aI_d,$$

where $d \in \{2, 3, 4, 5\}$ and $a \in \{0.1, 0.2, 0.4\}$ in the combinations shown in Figure 2. Its equilibrium is uniform and its value is $1/2 + a/d$. Varying d changes the number of entries that must be controlled, while varying a changes the separation of the deleted faces. Deterministic feedback returns the selected matrix entry. Bernoulli feedback has that entry as its mean.

These games permit an exact calibration of the sufficient certificate radius. Their equilibrium masses are $\rho = \tau = 1/d$ and their deletion gap is $\Delta_{\mathrm{del}} = a/[d(d-1)]$. For either augmented saddle system in (22), direct block elimination gives

$$\kappa = \max\left\{\frac{2(d-1)}{da} + \frac{1}{d}, \frac{3}{2} + \frac{a}{d}\right\}.$$

Substituting these quantities into (33) gives a worst-case sufficient certificate radius for the entire confidence box. The algorithm tests its sharper observable certificate, so entry may occur while the current maximum radius exceeds Γ_{cert}.

At $T = 640{,}000$, every deterministic run enters the local regime. The mean entry round rises from approximately 6.6×10^3 at $d = 2$ to 3.9×10^5 at $d = 5$ when $a = 0.4$. At $d = 3$, reducing a from 0.4 to 0.1 delays mean entry from approximately 3.2×10^4 to 1.27×10^5. Under Bernoulli feedback, the $d = 2, a = 0.4$ runs enter after approximately 1.44×10^5 rounds; all higher-dimensional points in the left panel of Figure 2 are right-censored, rather than observed entry times. This is the main finite-sample limitation visible in the study.

A.2 GAME-DERIVED EMPIRICAL META-GAMES

We construct empirical normal-form games from OpenSpiel (Lanctot et al., 2019). For each underlying game, a pool of 15 behavior policies contains a uniform policy, current and average policy-iteration checkpoints, and 12 seeded Dirichlet policies. We first estimate the resulting 15×15 payoff table. A deterministic search then selects a 3×3 regular fully mixed subgame by maximizing

$$\frac{\min\{\min_i x_i^\star, \min_j y_j^\star\}\,\Delta_{\mathrm{del}}}{\kappa_{\mathrm{aug}}},$$

where $x^\star, y^\star$ are the selected subgame's equilibrium strategies, Δ_{del} is its minimum deletion gap, and κ_{aug} is the augmented saddle-system condition number. This selection rule deliberately constructs an in-class stress test; it is not a claim that arbitrary OpenSpiel meta-games satisfy the paper's assumptions. The online learner receives neither the selected payoff matrix nor these regularity parameters.

For every selected entry we pre-generate a bank of 20,000 episode returns. The deterministic condition uses the exact selected mean matrix. In the episode condition, a run samples returns from the fixed bank. The policy pool, subgame indices, episode banks, and random seeds are stored with the released run manifests. Hence every method in a paired comparison sees the same empirical game and uses the same seed index.

Kuhn has the smallest deletion gap and largest condition number in the selected suite. None of its sixteen IC-RF runs enters local mode by $T = 640{,}000$, so IC-RF agrees with Matrix–UCB there. On resampled Goofspiel, the two methods separate only after local entry at roughly 4.5×10^4 rounds. At the horizon, the paired IC-RF-minus-Matrix–UCB regret difference is $0.284\ (0.080)$ (mean and standard error over eight matched seed pairs). This localizes the observed loss to finite-horizon local control.

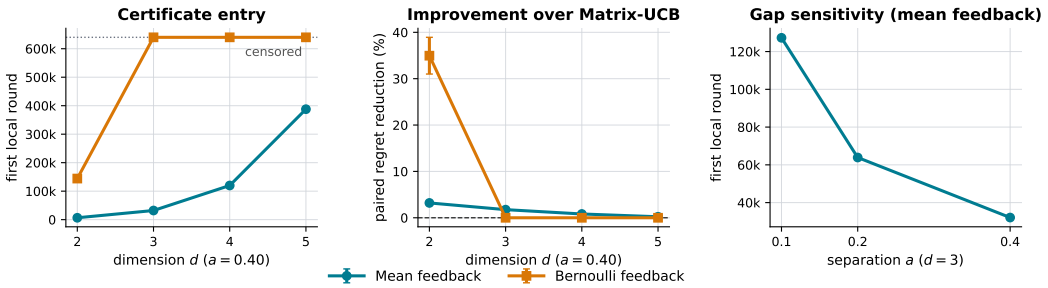


Figure 2: Finite-horizon scaling diagnostics on $A_{d,a} = \frac{1}{2}\mathbf{1}\mathbf{1}^\top + aI_d$ at $T = 640{,}000$. Left: first certificate entry; censored Bernoulli runs are plotted at the horizon. Middle: paired final-regret change relative to Matrix–UCB. Right: deterministic entry time as a varies. Error bars are one standard error over four paired seeds.

Table 3: Geometry of the selected 3×3 empirical meta-games. Here $\rho_{\min}$ is the smaller of the two minimum equilibrium masses.

game	value	$\rho_{\min}$	Δ_{del}	κ_{aug}
Colonel Blotto	0.5343	0.2909	0.1286	3.94
Goofspiel	0.4886	0.2349	0.0730	8.91
Kuhn poker	0.5423	0.1465	0.0083	42.11

A.3 Baselines and parameter handling

The main baselines are defined in Section 5.1. We also include EXP3-IX (Neu, 2015) as an implicit-exploration variant and, in a separate 2×2 camouflage–then–best-response panel, the official experimental rule of Maiti et al. (2026). The latter is not applied to higher-dimensional games. All methods in a panel use the same payoff normalization, opponent rule, horizon, and paired seed indices. Whenever a method solves a maximin linear program, we use the same deterministic vertex solver and lexicographically least optimum. The eight-setting main comparison has eight paired seeds at $T = 640{,}000$; the twelve-point scaling study has four. Run manifests retain the solver, reward streams, schedules, seeds, and source and output hashes.

A.4 Opponent-strategy stress tests

The five additional opponent rules test balanced column visits (round robin), permanent censoring (fixed column), a changing visible face (switching faces), one-step delayed adaptation (lagged best response), and a switch from equilibrium play to current-policy best response (camouflage). They all follow the informed timing of the theorem.

The *P2E anchor* (persistent two-expert anchor, abbreviated P2E in the figures) is an internal comparator. It retains the bounded prelude and selective restricted-face global policy. After its aggregate-interval certificate accepts, it publishes the persistent center–FTRL mixture $h_t = \pi_{t,1}\widehat{x}_e + \pi_{t,2}f_t$ defined in Section 4. It does not use IC-RF's hierarchical proposal or self-financing gate. Its certificate also differs from the final IC-RF configuration, so this column is not a single-component ablation; the anchor-only control in Table 2 holds the other IC-RF components fixed.

Table 4 reports all five rules and their numerical endpoints. Figures 3 and 4 retain the two trajectories in which the opponent's changing behavior is central to interpretation. The round-robin, fixed-column, and lagged-best-response results remain in the table; their curves add little beyond those endpoints. Curves average four paired seeds with one-standard-error bands. Table boldface is descriptive, not a significance claim. Negative signed regret means the realized restricted opponent was more favorable than the full game; it does not imply negative counterfactual security gap.

The fixed-column case checks the learn-or-earn alternative. On $A = I_2$, with column 1 chosen throughout, four theorem-schedule runs of length $50{,}000$ never accepted a local certificate. Their mean terminal signed regret per round was -0.4991, and their maximum positive prefix debt was one payoff unit. Proper-face surplus therefore financed the global phase despite the hidden column; unconditional eventual certification would be false in this setting.

Table 4: **Signed Nash-value regret under five opponent rules** at $T = 200{,}000$ (four paired runs; lower is better). Bold marks the lowest displayed mean, including ties. Noisy denotes Bernoulli feedback for RPS and resampled returns otherwise. Negative regret reflects favorable opponent play; the ranking does not measure counterfactual security.

Opponent	Game	Feedback	IC-RF	P2E anchor	M-UCB	CoEBL	EXP3	Max-UCB
Round robin	RPS	Mean	3.610	3.631	3.731	**3.374**	52.16	9.97×10^4
		Noisy	6.514	6.431	4.223	**3.777**	52.33	9.97×10^4
	Blotto	Mean	-329.5	-164.6	4.013	3.081	$\mathbf{-1.17\times10^4}$	-2126.3
		Noisy	-184.4	-59.00	-5.287	-4.068	$\mathbf{-1.17\times10^4}$	-4578.9
	Goofspiel	Mean	-106.3	1.165	2.715	3.066	**-3195.1**	-1736.1
		Noisy	-137.7	0.7643	2.736	3.387	**-3194.5**	-1738.2
	Kuhn	Mean	8.285	8.285	8.285	13.07	**-1638.7**	1999.5
		Noisy	138.4	138.4	138.4	137.5	**-1648.6**	114.9
Fixed column	RPS	Mean	$\mathbf{-9.99\times10^4}$	$\mathbf{-9.99\times10^4}$	-9.98×10^4	-9.98×10^4	-9.92×10^4	$\mathbf{-9.99\times10^4}$
		Noisy	-9.97×10^4	-9.97×10^4	-9.98×10^4	-9.98×10^4	-9.92×10^4	$\mathbf{-9.99\times10^4}$
	Blotto	Mean	$\mathbf{-6.54\times10^4}$	$\mathbf{-6.54\times10^4}$	-6.53×10^4	-6.52×10^4	-6.50×10^4	-6.52×10^4
		Noisy	$\mathbf{-6.54\times10^4}$	$\mathbf{-6.54\times10^4}$	-6.53×10^4	-6.53×10^4	-6.50×10^4	-6.52×10^4
	Goofspiel	Mean	$\mathbf{-3.99\times10^4}$	$\mathbf{-3.99\times10^4}$	-3.96×10^4	-3.96×10^4	$\mathbf{-3.99\times10^4}$	-3.96×10^4
		Noisy	$\mathbf{-3.99\times10^4}$	$\mathbf{-3.99\times10^4}$	-3.96×10^4	-3.96×10^4	$\mathbf{-3.99\times10^4}$	-3.96×10^4
	Kuhn	Mean	-5959.7	-5959.7	-5198.5	-5191.4	**-6007.3**	-5047.0
		Noisy	**-5955.0**	**-5955.0**	-5176.3	-5171.4	-5943.0	-5039.4
Switching faces	RPS	Mean	$\mathbf{-1.11\times10^4}$	-1.10×10^4	-9464.3	-9396.1	-8229.9	318.0
		Noisy	-8015.6	-7991.5	**-9448.6**	-9390.0	-7658.1	319.1
	Blotto	Mean	-1.52×10^4	-1.50×10^4	-1.37×10^4	-1.33×10^4	-2877.4	$\mathbf{-2.19\times10^4}$
		Noisy	-1.34×10^4	-1.34×10^4	-1.37×10^4	-1.33×10^4	-3935.3	$\mathbf{-2.19\times10^4}$
	Goofspiel	Mean	-5358.0	-5074.9	-3965.4	-3933.0	-1.04×10^4	$\mathbf{-1.69\times10^4}$
		Noisy	-5320.8	-4975.6	-3967.1	-3930.5	-1.04×10^4	$\mathbf{-1.69\times10^4}$
	Kuhn	Mean	-68.95	-68.95	683.2	679.0	**-300.9**	1996.2
		Noisy	433.6	433.6	783.6	791.6	**-889.4**	1965.2
Lagged BR	RPS	Mean	7.442	7.393	7.445	**3.829**	312.6	5.00×10^4
		Noisy	13.69	13.58	14.81	**11.89**	379.0	5.00×10^4
	Blotto	Mean	7.764	7.766	7.815	**3.615**	305.8	6.09×10^4
		Noisy	12.31	12.44	12.23	**9.182**	347.1	7.24×10^4
	Goofspiel	Mean	7.110	7.140	7.165	**6.159**	198.1	2.62×10^4
		Noisy	8.619	8.799	8.524	**5.494**	198.8	2.62×10^4
	Kuhn	Mean	24.60	24.60	24.60	**13.38**	208.2	4674.0
		Noisy	37.67	37.67	37.66	**26.25**	304.8	9101.0
Eq. camouflage	RPS	Mean	**9.266**	9.277	9.377	22.67	314.7	1.00×10^5
		Noisy	10.98	**10.83**	13.66	27.46	373.2	1.00×10^5
	Blotto	Mean	**19.01**	22.77	64.41	78.36	307.7	8.67×10^4
		Noisy	57.30	57.43	**57.22**	70.72	347.8	8.66×10^4
	Goofspiel	Mean	**23.36**	30.90	41.78	54.86	198.2	2.90×10^4
		Noisy	**23.79**	38.47	41.81	53.91	198.8	2.90×10^4
	Kuhn	Mean	**45.42**	**45.42**	**45.42**	56.17	207.6	9881.7
		Noisy	**55.29**	**55.29**	**55.29**	64.17	310.2	9911.1

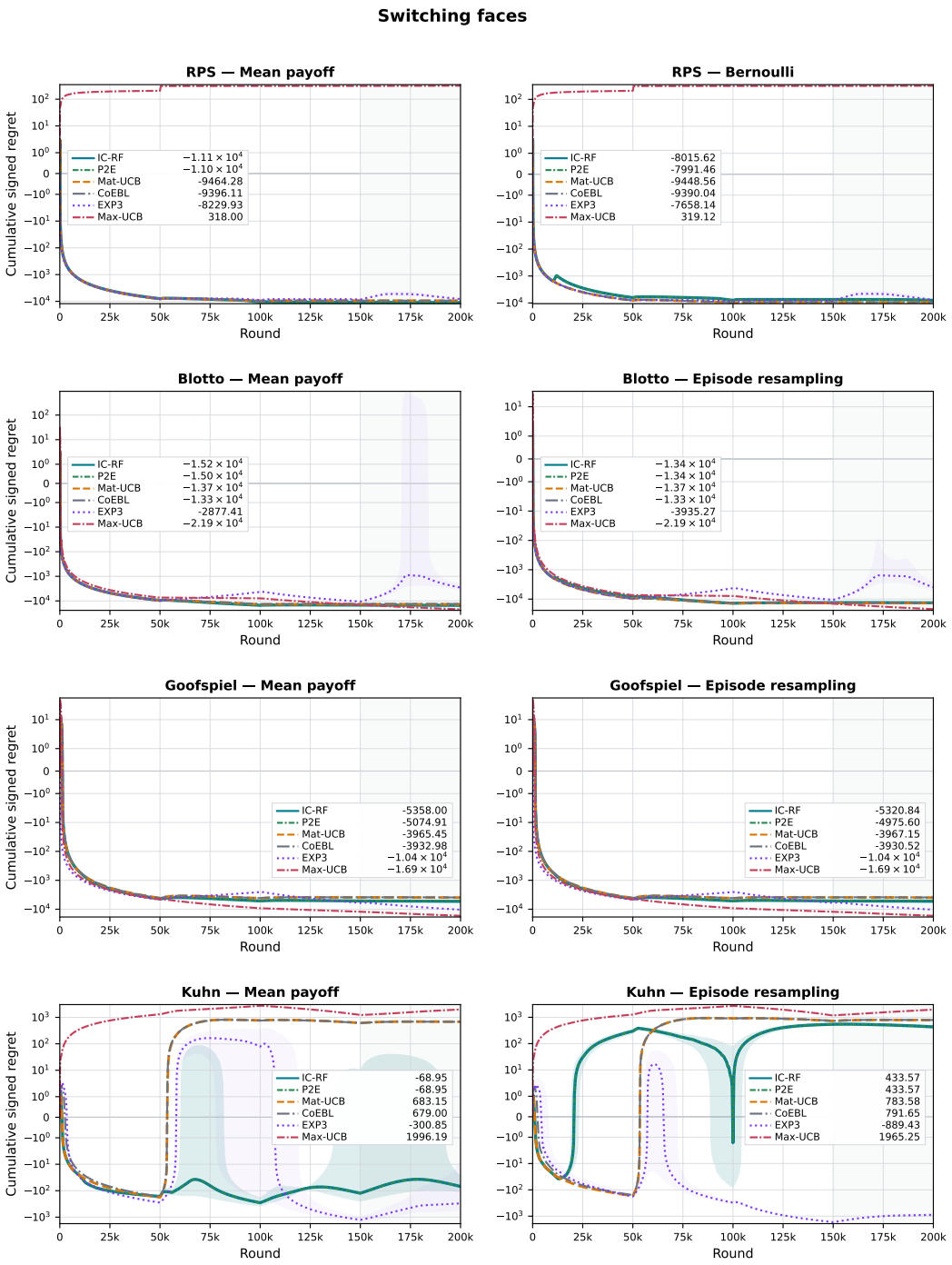


Figure 3: Switching-face opponent across four games and two feedback models. The omitted column changes over time, stressing restricted-face bookkeeping. Each panel lists method names and terminal regret; the same endpoints appear in Table 4.

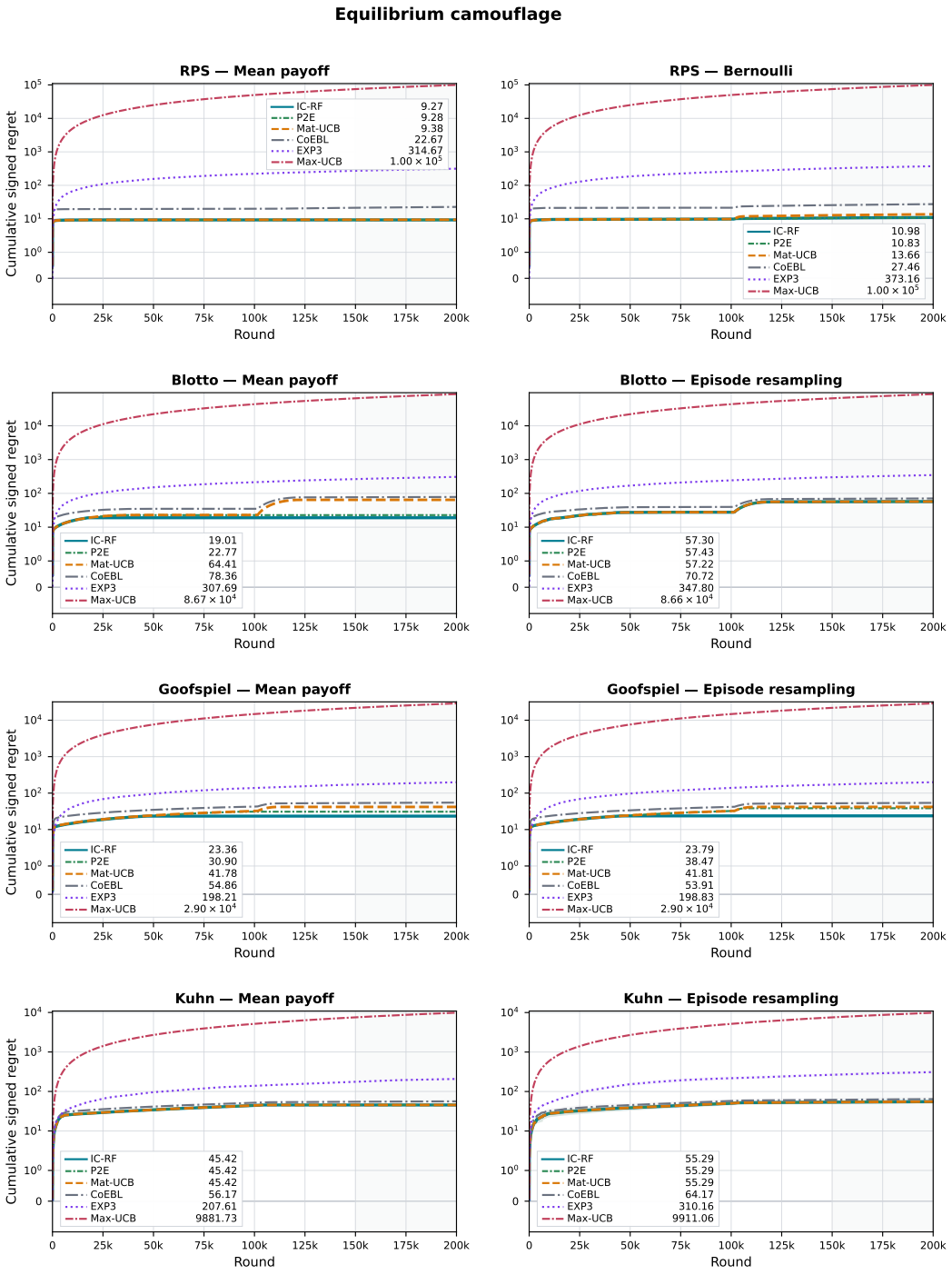


Figure 4: Equilibrium camouflage followed by a current-policy best response. The switch tests whether a benign-looking prefix causes persistent trust in the local controller. Each panel lists method names and terminal regret; the same endpoints appear in Table 4.

A.5 Component ablation

The best-response ablations are reported in Table 2. We use the same eight game–feedback settings, eight paired seeds per setting, and horizon $T = 640{,}000$ for the switching-face controls below. Each entry is the paired increase in terminal signed regret R_T when IC-RF is replaced by the indicated control, averaged equally across

the eight settings. Positive values favor IC-RF. The intervals describe seed variation on these fixed settings, not uncertainty over unseen games.

Table 5: **Switching-face ablations** at $T = 640{,}000$: paired mean increase in $R_T \pm 95\%$ confidence-interval half-width (positive favors IC-RF). Intervals are Bonferroni-adjusted by schedule except † (nominal); the diagnostic schedule sets $\mathcal{L} = 1$.

Change to IC-RF	**Proof schedule**	**Diagnostic schedule**
Old fixed-floor controller	334.15 ± 135.93	$8\,201.93 \pm 1\,131.59$
Matrix–UCB global	$1\,525.92 \pm 182.33$	$3\,322.40 \pm 301.76$
No bounded prelude	$55\,065.07 \pm 10\,337.36$	$56\,861.55 \pm 10\,400.58$
All-row floor, same LP†	696.50 ± 4.67	$1\,067.51 \pm 113.31$

The old fixed-floor control changes both midpoint and policy, so it does not isolate the exploration floor. All-row floor keeps the lower-confidence restricted-game LP and changes only the exploration rows. Its signed-regret advantage for IC-RF under the proof schedule is concentrated in one setting, while it lowers mean cumulative security gap by 1473.28 (proof schedule) and 2359.64 (diagnostic). It therefore does not establish a counterfactual-security advantage. Separately, removing confidence coverage from face and entry decisions raises mean best-response regret from 20.29 to 11818.86; this extra control and per-setting results remain in the artifact.

Algorithm 2 Complete state machine of IC-RF

Require: horizon T, confidence level δ

Initialize the lifelong confidence state, $t \leftarrow 1$, $k \leftarrow 0$, and mode Global; retain counts and intervals across all modes
while $t \leq T$ **do**
  if a nested entry interval is empty **then**
    play the uniform row through T; **break**
  end if
  if mode is Global **then**
    if $S_k(t) = [d]$ **then** ▷ this test precedes the scale timeout
      if $\max_{i,j} r_{t,ij} = 0$ **then**
        play an exact maximin strategy of the identified matrix through T; **break**
      else
        evaluate the observable saddle certificate and radius grid
        if accepted **then**
          call INITIALIZEEPOCH in Algorithm 3
          set mode to Local without consuming a round; **continue**
        else
          find the least scale whose target is strictly below the current maximum radius
          if none exists **then**
            play the uniform row through T; **break**
          end if
          advance to that scale; reset both scale clocks; **continue**
        end if
      end if
    end if
    if the proper-face scale clock equals H_k **then**
      $k \leftarrow k + 1$; reset the scale and prelude clocks; **continue**
    end if
    if fewer than Q_k actions have been played at scale k **then**
      publish the lexicographically selected maximin policy of the Matrix–UCB optimistic matrix
    else if $S_k(t) = \varnothing$ **then**
      publish the uniform row distribution
    else
      publish the lexicographically selected selective-floor restricted maximin policy
    end if
    observe (j_t, i_t, Y_t), update the lifelong intervals and clocks, and promote every column meeting the radius test

Algorithm 2 Complete state machine of IC-RF (continued)

else
  call LOCALSTEP in Algorithm 3
  if a nested entry interval is empty **then**
    play the uniform row on the remaining rounds; **break**
  end if
  if the current radius first falls to or below $\varepsilon_e/2$ and fewer than $K_T + 1$ epochs have been accepted **then**
    if the current radius is zero **then**
      play an exact maximin strategy of the identified matrix on the remaining rounds; **break**
    end if
    recertify before publishing another action
    if accepted **then**
      call INITIALIZEEPOCH
    else
      find the least global scale strictly finer than the current radius
      if none exists **then**
        play the uniform row on the remaining rounds; **break**
      end if
      return to that scale in Global mode; reset both scale clocks
    end if
  else if $K_T + 1$ epochs have been accepted **then**
    disable further refreshes; retain the frozen certificate and continue LOCALSTEP
  end if
end if
$t \leftarrow t + 1$
end while

This state machine defines IC-RF throughout the paper.

Algorithm 3 Local-epoch initialization and single-round update

procedure INITIALIZEEPOCH ▷ called once per accepted epoch
  Freeze $(M_e, R_e, \widehat{x}_e, \mathcal{X}_e)$ returned by the certificate
  Let $\overline{C}_e$ be the smallest dyadic number no smaller than $\max\{1, \mathcal{R}_e\}$
  if this is the first epoch or $\overline{C}_e$ exceeds the stored cap C **then**
    set $C \leftarrow \overline{C}_e$ and $\lambda = (\sqrt{d}C)^{-1}$
    start a cap segment and initialize the center and FTRL weights in proportion to 1 and $\widetilde{\varepsilon}_e = \max\{\varepsilon_e, 1/[4(T \vee 2)]\}$
  else
    retain C, λ, and the two anchor weights
  end if
  Reset the epoch FTRL state and observable wealth $Z_e \leftarrow B_e$
  Initialize proposal states only at first entry; retain them thereafter
end procedure
procedure LOCALSTEP ▷ no initialization or reset
  Form the center–FTRL anchor $h_t \in \mathcal{X}_e$ and predictable proposal $q_t \in \mathcal{X}_e$
  Set $\alpha_t = \min\{1, Z_t/B_e\}$ (use $\alpha_t = 1$ if $B_e = 0$) and publish $p_t = (1 - \alpha_t)h_t + \alpha_t q_t$
  Observe (j_t, i_t, Y_t) and set
  $$\underline{g}_t = (q_t - h_t)^\top M_{e,:,j_t} - |q_t - h_t|^\top R_{e,:,j_t}, \qquad Z_{t+1} = Z_t + \alpha_t \underline{g}_t.$$
  Update the lifelong intervals, cumulative FTRL gradient, center–FTRL weights, and proposal controllers using the same frozen midpoint column $M_{e,:,j_t}$
end procedure

B THE IC-RF STATE MACHINE

Algorithms 2 and 3 give the transition order for Theorem 3; this section specifies the remaining optimization rules. All tests precede publication and transitions consume no round. Matrix-game and certificate ties are resolved lexicographically; radius-grid ties use the smallest index.

Persistent state. Counts, payoff sums, nested intervals, and local cap, weights, and proposal states persist through transitions, including returns to global mode. The multiscale liabilities are analysis bookkeeping, not policy state. The empty-intersection branch of Algorithm 2 is unreachable on the lifelong confidence event but makes the policy total. INITIALIZEEPOCH resets the FTRL accumulator and wealth only when a certificate is accepted. Every subsequent action calls LOCALSTEP, preserving these states. Terminal local mode freezes the certificate but continues the same updates; the uniform and exact-matrix terminal branches instead use fixed policies.

Global mode. The scale index belongs to $\{0, \dots, K_T\}$, and the scale-round and prelude counters are reset to zero whenever a new scale starts. Let $N_{ij}(t)$ and $\widehat{A}_{ij}(t)$ be the count and empirical mean of entry (i, j) before round t, and define

$$U_{ij}^{\mathrm{MUCB}}(t) = \widehat{A}_{ij}(t) + \sqrt{\frac{2\log(2d^2T^2)}{N_{ij}(t) \vee 1}}.$$

The active face $S_k(t)$ is defined in (15). The full-face branch precedes a simultaneous timeout; at scale K_T, $H_{K_T} > T$ makes timeout unreachable. The first Q_k actions use a row-maximin strategy of $U^{\mathrm{MUCB}}(t)$. For later actions, let $C_j(t)$ count prior appearances of column j and define

$$\begin{aligned}\mathcal{B}_k(t) = \{i \in [d] : \text{some previously revealed } j \notin S_k(t),\\ C_j(t) > 0, \text{ and } r_{t,ij} > \varepsilon_k\}.\end{aligned} \tag{6}$$

be the rows blocking promotion of an inactive column. For a nonempty active face, the policy solves

$$\begin{aligned}\max_{p \in \Delta_d, z} \quad & z \\ \text{subject to} \quad & p_i \geq \varepsilon_k/d \qquad i \in \mathcal{B}_k(t), \\ & p^\top L_{t,:,j} \geq z \quad j \in S_k(t).\end{aligned} \tag{7}$$

The empty-face policy is uniform. A previously unseen column is exempt from the floor on its first appearance; its one-round liability is booked once. Subsequent inaccurate entries enter $\mathcal{B}_k(t)$.

Proposition 2 (First-completion certification)**.** *Suppose the certificate is evaluated immediately after an update makes the active face full, and again before the next global action whenever the face is already full. Give a full face priority over the phase timeout. This inspection rule preserves the confidence level and the regret guarantee of Theorem 3.*

Proof. The active face is measurable after the preceding observation, so every inspection occurs before the next action is published. The entrywise confidence event is time-uniform; evaluating a deterministic certificate at an additional stopping time spends no additional failure probability. If the certificate accepts, its soundness theorem supplies all local hypotheses. If it rejects, the scale advances before another global action. The actions already played at that scale form a prefix covered by Lemma 4. Giving the full-face branch priority also prevents a completed face from being discarded by a simultaneous timeout. Hence every realized round belongs to exactly one certified global or local prefix. □

Certification and cap selection. An accepted candidate supplies $(M_e, R_e, \widehat{x}_e, \mathcal{R}_e)$, the local set $\mathcal{X}_e$ of Section 4, a row floor, and deletion-surplus bounds. The exact-zero global branch precedes certification, so $\varepsilon_e > 0$ in every accepted epoch. Before scoring a candidate, define

$$C_e = \max\left\{C, 2^{\lceil \log_2 \max\{1, \mathcal{R}_e\} \rceil}\right\}, \qquad \lambda_e = (\sqrt{d} C_e)^{-1}.$$

The certificate evaluates curvature using λ_e, which is fixed before radius selection and has $C_e \geq \mathcal{R}_e$.

At the first accepted epoch or when $C_e > C$, set $C \leftarrow C_e$, $\lambda \leftarrow \lambda_e$ and initialize a cap segment with

$$\widetilde{\varepsilon}_e = \max\{\varepsilon_e, 1/[4(T \vee 2)]\}, \qquad (w_1, w_2) = \frac{(1, \widetilde{\varepsilon}_e)}{1 + \widetilde{\varepsilon}_e}.$$

Otherwise retain the cap, learning rate, and weights. The center–FTRL anchor, proposal, publication gate, and updates are defined in Section 4 and Algorithm 3; their regret properties are proved in Sections G.3 and H.

C PROOF OF THEOREM 1

Theorem 3 (IC-RF). *Suppose $d \geq 2$ and $A \in [0,1]^{d\times d}$ has a unique equilibrium and both equilibrium strategies have full support. Against every adaptive informed-bandit opponent, IC-RF satisfies, with probability at least $1-\delta$,*

$$\mathrm{Reg}_T \leq \frac{38400d^2\mathcal{L}}{\Gamma_{\mathrm{cert}}^3} + \frac{6\sqrt{d}\,\Lambda_T}{\Gamma_{\mathrm{cert}}} + (K_T+1)\left\{\frac{9}{8\Gamma_{\mathrm{cert}}^2} + \frac{42496d\mathcal{L}}{\Gamma_{\mathrm{cert}}\rho\Delta_{\mathrm{del}}} + \frac{6\sqrt{d}}{\Gamma_{\mathrm{cert}}}\right\}. \tag{8}$$

The learner is not given an instance parameter in this bound. For every fixed regular game, the dependence on T is polylogarithmic.

Proof. Let $\mathcal{E}$ be the intersection of the lifelong confidence event in Theorem 8, the global row-sampling event in Lemma 3, and the local row-sampling event in Lemma 10. The failure allocations in their statements give $\mathbb{P}(\mathcal{E}) \geq 1-\delta$. We work on $\mathcal{E}$.

Let k_Γ be the first dyadic scale for which $\varepsilon_{k_\Gamma} \leq \Gamma_{\mathrm{cert}}$. Lemmas 2 and 3 establish the payoff and promotion properties of the actual selective-floor policy. Proposition 6 returns every rejected local certificate to a strictly finer scale, so no global scale is revisited. All prefixes before k_Γ therefore form a single geometric sum and cost at most $30W_{k_\Gamma}$.

At scale k_Γ and every finer scale, a full confidence face is accepted by Theorem 6. A fine scale that does not reach a full face must finish its clock; its credit is at least $2W_{k+1}$ by Lemma 4, whereas the complete positive prefix liability of its immediate successor is at most $7W_{k+1}/10$. The explicit sum in Proposition 3 bounds the fine-scale suffix by $7W_{k_\Gamma}/10$. Coarse prefixes together cost less than $(26/7)W_{k_\Gamma}$, so the common $30W_{k_\Gamma}$ envelope covers both. Since $\varepsilon_{k_\Gamma} > \Gamma_{\mathrm{cert}}/2$ by minimality,

$$\mathrm{Reg}_{\mathrm{global}} \leq 30W_{k_\Gamma} \leq \frac{38400d^2\mathcal{L}}{\Gamma_{\mathrm{cert}}^3}.$$

If the horizon ends earlier, the executed scales are a prefix of the same sum. If the terminal branch is reached at scale K_T, rejection and Theorem 6 imply $\Gamma_{\mathrm{cert}} < \varepsilon_{K_T} \leq 1/[2(T\vee 2)]$; the remaining regret is at most T. In this case $k_\Gamma > K_T$, the preceding global prefixes cost less than $30W_{K_T} \leq (30/8)W_{k_\Gamma}$, and $T < 1/(2\Gamma_{\mathrm{cert}}) < W_{k_\Gamma}$. Both pieces are therefore below $30W_{k_\Gamma}$. Thus the first term of (8) covers every global round.

It remains to bound the accepted local epochs. Every accepted certificate has maximum radius ε_e. If $\varepsilon_e > \Gamma_{\mathrm{cert}}$, the certified row-floor test and $\underline{\rho}_e \leq 1$ imply $\mathcal{R}_e < 1/(2\varepsilon_e) < 1/(2\Gamma_{\mathrm{cert}})$. If $\varepsilon_e \leq \Gamma_{\mathrm{cert}}$, then $k_{2,e} \leq 4\overline{\kappa}\sqrt{d}$ and the adaptive bonus is at most $8\sqrt{d}$, so $\mathcal{R}_e \leq 12\overline{\kappa}\sqrt{d} \leq 3/(2\Gamma_{\mathrm{cert}})$ by the first component of (33). Moreover, (45) gives $\underline{\rho}_e \geq \rho/2$ and $\underline{\Delta}_e \geq \Delta_{\mathrm{del}}/2$. The published local row floor is $\rho_e = \underline{\rho}_e/2$, while $\beta_e = \underline{\Delta}_e/[8(1+\underline{\Delta}_e)]$. Because every payoff interval is contained in $[0,1]$, the certified deletion surplus satisfies $\underline{\Delta}_e \leq 1$. Consequently every accepted epoch satisfies

$$\mathcal{R}_e \leq \frac{3}{2\Gamma_{\mathrm{cert}}}, \qquad \rho_e \geq \frac{\rho}{4}, \qquad \beta_e \geq \frac{\Delta_{\mathrm{del}}}{32}.$$

Also $H_e^2 \leq d$, $D_e^{\mathrm{norm}} \leq \sqrt{d}$, and $C_{r(e)} \geq \mathcal{R}_e$. The two-expert curvature therefore satisfies

$$\mathcal{G}_e^{(2)} \leq \frac{2d}{\Gamma_{\mathrm{cert}}}.$$

For every nonterminal epoch, substitution into Theorem 7 gives

$$\frac{\mathcal{R}_e^2}{2} \leq \frac{9}{8\Gamma_{\mathrm{cert}}^2}, \qquad \frac{166\mathcal{G}_e^{(2)}\mathcal{L}}{\rho_e\beta_e} \leq \frac{42496d\mathcal{L}}{\Gamma_{\mathrm{cert}}\rho\Delta_{\mathrm{del}}}.$$

If a terminal certified epoch exists, (60) gives

$$2\mathcal{G}_{e_\dagger}^{(2)}T\varepsilon_{e_\dagger}^2 \leq \frac{d}{4\Gamma_{\mathrm{cert}}(T\vee 2)}.$$

Since $\rho \leq 1$, $\Delta_{\mathrm{del}} \leq 1$, and $\mathcal{L} \geq 1$, this is smaller than the nonterminal per-epoch envelope in the preceding display. Thus the common (K_T+1) multiplier in (8) covers all accepted epochs, including the terminal one.

Lemma 13 with $\mathcal{R}_{\max} = 3/(2\Gamma_{\mathrm{cert}})$ gives the second term in (8); the horizon-clipped prior in the state machine supplies its logarithmic bound. There are at most K_T+1 accepted local epochs.

It remains to account for the published self-financing policy. In epoch e, Lemma 11 adds at most B_e relative to the anchor. Confidence half-widths are at most $1/2$, $H_e \le \sqrt{d}$, and $J_e \le \sqrt{d}\varepsilon_e$. Hence

$$B_e \le 4\sqrt{d}\mathcal{R}_e\varepsilon_e \le \frac{6\sqrt{d}}{\Gamma_{\mathrm{cert}}}.$$

Adding this conservative per-epoch liability proves the display. □

D CONFIDENCE-RADIUS PROMOTION AND THE GLOBAL LEDGER

This section proves the global explore-or-earn argument. We first show that deleting any column from a regular fully mixed game creates a strict value surplus. We then show that an outside-face round either promotes an inaccurate column or can occur only a bounded number of times. The final ledger carries the surplus from one accuracy scale to the next.

Let $v_S(B) = \max_{x\in\Delta_d} \min_{j\in S} x^\top B_{:j}$ denote the value of the game with payoff matrix B and nonempty column set S. Write $v(B) = v_{[d]}(B)$.

Lemma 1 (A unique fully mixed equilibrium has strict deletion surplus)**.** *If A has a unique equilibrium $(x^\star, y^\star)$ and $y^\star$ is fully mixed, then*

$$v_{[d]\setminus\{j\}}(A) > v(A) \qquad \textit{for every } j \in [d]. \tag{9}$$

Proof. Removing a column cannot decrease the row value. Suppose equality held for some j, and let $\widetilde{y}$ be a column minimax strategy of the restricted game. Extend it to $[d]$ by assigning mass zero to column j. Its full-game upper value remains $\max_i e_i^\top A\widetilde{y} = v(A)$. Hence $\widetilde{y}$ is a column minimax strategy of the original game. It differs from $y^\star$, whose jth coordinate is positive, contradicting uniqueness. □

For the true regular game, define the deletion surplus

$$\Delta_{\mathrm{del}} = \min_{j\in[d]} \{v_{[d]\setminus\{j\}}(A) - v(A)\} > 0. \tag{10}$$

Positivity follows from Lemma 1 and finiteness of the column set.

Use the allocation in Section I, whose per-tail error is $\alpha_\star = \delta/(4d^2)$. Let

$$K_T = 1 + \lceil \log_2(T \vee 2) \rceil \tag{11}$$

and define the common logarithmic factor

$$\mathcal{L} = \max\left\{1, \mathcal{L}_T(\alpha_\star), \log\frac{8d^2(K_T+1)}{\delta}\right\}. \tag{12}$$

At global scale k, set $\varepsilon_k = 2^{-k}$ and define

$$N_k = \left\lceil \frac{9\mathcal{L}}{\varepsilon_k^2} \right\rceil, \qquad V_k = \left\lceil \frac{2d(N_k + 4\mathcal{L})}{\varepsilon_k} \right\rceil, \tag{13}$$

$$Q_k = 8d^2 N_k, \qquad W_k = \frac{160 d^2 \mathcal{L}}{\varepsilon_k^3}, \qquad H_k = \left\lceil \frac{8W_k}{\varepsilon_k} \right\rceil. \tag{14}$$

The quantities N_k and V_k are proof envelopes. The algorithm does not use them to promote a column. It uses Q_k only as the deterministic length cap of the Matrix-UCB prelude, and it uses H_k as the deterministic scale cap.

Let L_t, U_t, M_t, and $r_{ij}(t)$ denote the nested confidence rectangle, its midpoint, and its entrywise radii at calendar time t. The active face is

$$S_k(t) = \{j \in [d] : \max_i r_{ij}(t) \le \varepsilon_k\}. \tag{15}$$

After the Matrix-UCB prelude, the global policy is uniform when $S_k(t)$ is empty. Otherwise it is the lexicographically selected solution of (7). The blocking set in that program is observable before the action. Every row whose entry prevents promotion of some previously revealed inactive column receives probability at least ε_k/d. A never-revealed column is exempt until its first appearance; that single reveal is charged directly below.

Lemma 2 (Proper faces have true surplus)**.** *Every nonempty proper set $S \subsetneq [d]$ satisfies*

$$v_S(A) \ge v(A) + \Delta_{\mathrm{del}}. \tag{16}$$

After the prelude, on the confidence event, if $S_k(t)$ is nonempty and proper and the opponent chooses $j_t \in S_k(t)$, then the one-round signed Nash-value regret obeys

$$v(A) - p_t^\top A_{:j_t} \le -\Delta_{\mathrm{del}} + 3\varepsilon_k. \tag{17}$$

Proof. Choose any $j \notin S$. Since removing columns can only increase the row player's value, $v_S(A) \geq v_{[d]\setminus\{j\}}(A) \geq v(A) + \Delta_{\mathrm{del}}$.

Let p_L be a maximin strategy of the lower matrix L_t on $S_k(t)$. The distribution $\bar{p} = (1-\varepsilon_k)p_L + \varepsilon_k \mathbf{1}/d$ is feasible for (7); hence the selected policy p_t satisfies

$$\min_{j \in S_k(t)} p_t^\top L_{t,:,j} \geq v_{S_k(t)}(L_t) - \varepsilon_k.$$

On an active column, confidence coverage gives $0 \leq A_{:j} - L_{t,:,j} \leq 2\varepsilon_k \mathbf{1}$, so value Lipschitzness implies $v_{S_k(t)}(L_t) \geq v_{S_k(t)}(A) - 2\varepsilon_k$. Therefore $p_t^\top A_{:,j_t} \geq v_{S_k(t)}(A) - 3\varepsilon_k$. Combining this inequality with (16) proves (17). □

Lemma 3 (Observable promotion after finitely many outside appearances)**.** *There is an event $\mathcal{E}_{\mathrm{row}}$ of probability at least $1 - \delta/4$ on which the following holds simultaneously at every scale and for every column. After the prelude, a column that has already appeared can be selected at most V_k times while remaining outside $S_k(t)$. A column that has never appeared before the post-prelude block can be selected at most $V_k + 1$ times: its first appearance is exempt from the row floor, and every subsequent appearance is covered by the same bound.*

Proof. Fix a scale k, a column j, and a row i. If j has not previously appeared, discard its first post-prelude appearance from the following indexing. Thereafter $C_j(t) > 0$. Index the next post-prelude rounds on which column j is selected, and let τ_i be the first such appearance before which $r_{ij} \leq \varepsilon_k$. Before τ_i, row i belongs to the blocking set in (7); hence its conditional sampling probability is at least $p_k = \varepsilon_k/d$. Define

$$Z_s = \mathbf{1}\{s \geq \tau_i\} + \mathbf{1}\{s < \tau_i,\ i_s = i\}.$$

This is a $\{0,1\}$-valued adapted process with conditional mean at least p_k. If the scale or horizon ends before the required appearances, complete the process with ones. Its conditional lower-mean bound is preserved, and any failure at an actually reached appearance is a failure of this completed process. A multiplicative martingale Chernoff bound therefore gives

$$\mathbb{P}\left(\sum_{s=1}^{V_k} Z_s < \frac{p_k V_k}{2} \;\middle|\; \text{the scale-start history} \right) \leq \exp(-p_k V_k/8). \tag{18}$$

The definitions give $p_k V_k/2 \geq N_k + 4\mathcal{L} \geq N_k$ and $p_k V_k/8 \geq \mathcal{L}$. Taking a union bound over at most $d^2(K_T+1)$ row–column–scale triples gives failure probability at most $d^2(K_T+1)e^{-\mathcal{L}} \leq \delta/8$. Conditioning on the random scale-start history is removed by the tower property.

On the complementary event, either $\tau_i \leq V_k$, or row i was still blocking throughout the first V_k appearances and was sampled at least N_k times. In the second case the constants in (63) satisfy $4a_1^2 < 9$ and $4a_2 < 9$, so Corollary 2 and (12) imply $r_{ij} \leq \varepsilon_k$ as well. Thus every row has ceased blocking by the V_k-th appearance. Applying the argument simultaneously to all rows shows $j \in S_k(t)$ by (15). Restoring the discarded first reveal proves the $V_k + 1$ bound for a previously unseen column. Since the appearance counter persists across scales, a given column incurs this exception at most once over the entire run. □

Lemma 4 (One-scale signed-regret ledger)**.** *On the intersection of the confidence and row-sampling events, let O_k be the number of post-prelude rounds at scale k on which $j_t \notin S_k(t)$. Then*

$$O_k \leq d(V_k+1) \leq \frac{W_k}{5}, \qquad Q_k \leq \frac{W_k}{2}. \tag{19}$$

Every prefix of the scale has regret at most $26W_k$. If $\varepsilon_k \leq \Delta_{\mathrm{del}}/6$, every prefix has regret at most $7W_k/10$. If such a scale completes H_k rounds without entering local mode, its total regret is at most

$$-2W_{k+1}. \tag{20}$$

Proof. Lemma 3 allows at most $V_k + 1$ outside appearances per column, hence $O_k \leq d(V_k+1)$. Since $\mathcal{L} \geq 1$ and $\varepsilon_k \leq 1$,

$$N_k \leq \frac{10\mathcal{L}}{\varepsilon_k^2}, \qquad Q_k \leq \frac{80d^2\mathcal{L}}{\varepsilon_k^2} \leq \frac{W_k}{2},$$

and

$$d(V_k+1) \leq 2d + \frac{2d^2}{\varepsilon_k}\left(\frac{10\mathcal{L}}{\varepsilon_k^2} + 4\mathcal{L}\right) \leq \frac{30d^2\mathcal{L}}{\varepsilon_k^3} < \frac{W_k}{5}.$$

Thus the total number of prelude and outside-face rounds is at most $7W_k/10$. Each such round has regret at most one. On every remaining round, Lemma 2 gives regret at most $3\varepsilon_k$ after dropping the negative surplus. Since $H_k \leq 8W_k/\varepsilon_k + 1$, every prefix costs at most

$$\frac{7W_k}{10} + 3\varepsilon_k H_k \leq 26W_k.$$

If $\varepsilon_k \le \Delta_{\mathrm{del}}/6$, every inside-face round has nonpositive regret by (17); this proves the sharper prefix bound. A completed scale contains at least $H_k - 7W_k/10$ inside-face rounds. Their regret is at most $-\Delta_{\mathrm{del}}/2 \le -3\varepsilon_k$. Consequently,

$$\begin{aligned} R_k &\le \frac{7W_k}{10} - 3\varepsilon_k \left(H_k - \frac{7W_k}{10} \right) \\ &\le \frac{14W_k}{5} - 24W_k \le -16W_k = -2W_{k+1}, \end{aligned}$$

where $W_{k+1} = 8W_k$. □

Proposition 3 (One uninterrupted global block)**.** *Fix a deterministic analysis threshold Γ such that $\Gamma \le \Delta_{\mathrm{del}}/6$, and suppose every full-face certificate with maximum entrywise radius at most Γ accepts. Let $k_0 = \min\{k \ge 0 : \varepsilon_k \le \Gamma\}$, and consider a global block starting at a scale no larger than k_0. Up to the horizon, local entry, or a later local-to-global return, the cumulative regret of the block is at most*

$$30W_{k_0} \le \frac{38400 d^2 \mathcal{L}}{\Gamma^3}. \tag{21}$$

After scale k_0, every completed scale pays for the next scale's entire prefix without reusing credit.

Proof. Before k_0, distinct executed scales cost at most $26 \sum_{k<k_0} W_k < (26/7)W_{k_0}$. At and after k_0, a full face exits to local mode. Otherwise completed scales are consecutive. If $q > k_0$ is the last executed scale, their regret including its final prefix obeys

$$\sum_{k=k_0}^{q-1} R_k + R_q(\mathrm{prefix}) \le -2 \sum_{k=k_0}^{q-1} W_{k+1} + \frac{7}{10} W_q \le 0.$$

If $q = k_0$, the remaining prefix costs at most $7W_{k_0}/10$. These bounds are below $30W_{k_0}$, including when the horizon ends before scale k_0 is reached. Finally, minimality of k_0 gives $\varepsilon_{k_0} > \Gamma/2$, which implies the displayed explicit bound. □

D.1 The bounded Matrix–UCB prelude

At scale k, let σ_k denote the stopping time at which the scale begins and let Q_k be the deterministic prelude cap. A prelude rule is a sequence of row distributions p_t^{pre} that is measurable with respect to the learner's information immediately before publishing p_t. Every observation collected during the prelude is incorporated into the single lifelong confidence rectangle.

Proposition 4 (Bounded-prelude ledger)**.** *Suppose the payoff observations lie in $[0, 1]$. Any predictable prelude of length at most Q_k, followed by the selective-floor policy in* (7)*, satisfies the promotion and one-scale ledger bounds in Lemmas 3 and 4.*

Proof. Fix a scale k and condition on the history at its starting time σ_k. The proof of the promotion lemma indexes only rounds after the prelude on which a column outside the active face is selected. On each such round after a column's first reveal, the post-prelude policy assigns every still-blocking row probability at least ε_k/d. Hence its conditional multiplicative-martingale bound, and therefore its upper bound $O_k \le d(V_k + 1)$ on post-prelude outside-face rounds, is unaffected by the actions played before that index starts. The added one is the single unfloored first reveal of a previously unseen column. Prelude observations can only shrink the nested confidence rectangle and therefore cannot delay promotion.

The proper-face surplus lemma is a pointwise statement about the post-prelude restricted policy. The one-scale ledger treats the prelude only through the deterministic estimate

$$\sum_{t \in \mathcal{P}_k} \left(v(A) - p_t^\top A_{:,j_t} \right) \le |\mathcal{P}_k| \le Q_k,$$

where $\mathcal{P}_k$ is the set of prelude rounds at scale k. The first inequality follows from $A \in [0, 1]^{d \times d}$ and does not depend on the prelude decision rule. Since the budget choice satisfies $Q_k \le W_k/2$, every prefix and completed-scale calculation in the one-scale ledger follows. Conditioning on the random scale-start history is removed by the tower property. □

The state machine uses the predictable Matrix–UCB distribution defined in Section B for exactly this bounded prelude. The theorem charges these rounds only through Q_k; no optimism property is required for the regret bound.

E INTERVAL CERTIFICATION AND SEQUENTIAL RE-ENTRY

The local algorithm needs more than a point estimate of the equilibrium. It needs guarantees that every matrix still allowed by the confidence rectangle has the same stable saddle geometry. This section gives an observable test for those guarantees. Soundness is deterministic; the confidence event is used later to ensure that the true matrix lies in the tested rectangle.

For a matrix $B \in \mathbb{R}^{d\times d}$, define the augmented saddle systems

$$\mathsf{B}_x(B) = \begin{bmatrix} B^\top & -\mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix}, \qquad \mathsf{B}_y(B) = \begin{bmatrix} B & -\mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix}, \qquad b = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}. \tag{22}$$

Whenever the systems are invertible, write $z_x(B) = \mathsf{B}_x(B)^{-1}b = (x(B), v_x(B))$ and $z_y(B) = \mathsf{B}_y(B)^{-1}b = (y(B), v_y(B))$. The notation does not assert that the first d coordinates are positive; positivity is one of the certificate's tests.

Lemma 5 (Uniqueness and full mixing imply nonsingularity)**.** *If a square game $A \in \mathbb{R}^{d\times d}$ has a unique fully mixed equilibrium, then both augmented systems in* (22) *are invertible.*

This is the augmented-system form of the classical completely mixed-game characterization (Kaplansky, 1945; Shapley & Snow, 1950). We include the short proof because the precise system is used throughout the certificate.

Proof. Let $(x^\star, y^\star)$ be the unique fully mixed equilibrium and suppose $\mathsf{B}_x(A)$ had a nonzero kernel vector (h, c). Then

$$A^\top h = c\mathbf{1}, \qquad \mathbf{1}^\top h = 0. \tag{23}$$

Because $x^\star$ is interior, $x^\star + th$ lies in the simplex for all sufficiently small positive and negative t. Its payoff against every column equals $v(A) + tc$. If $c \neq 0$, choosing the sign of t contradicts the definition of the game value. If $c = 0$, every sufficiently small $x^\star + th$ is a distinct row equilibrium, contradicting uniqueness. Thus the row augmented system is nonsingular. Applying the same argument to $Ay^\star = v(A)\mathbf{1}$ proves nonsingularity of the column system. □

Fix a nonempty confidence rectangle $[L, U] \subseteq [0, 1]^{d\times d}$ and put

$$M = \frac{L+U}{2}, \qquad R = \frac{U-L}{2}, \qquad \varepsilon = \max_{i,j} R_{ij} > 0. \tag{24}$$

For $q \in \{x, y\}$, let $\widehat{B}_q = \mathsf{B}_q(M)$ and let D_q be the nonnegative augmented matrix whose upper-left block is $R^\top$ for $q = x$ and R for $q = y$, with all remaining entries zero. When $\widehat{B}_q$ is invertible, define

$$C_q = |\widehat{B}_q^{-1}|D_q, \qquad \widehat{z}_q = \widehat{B}_q^{-1}b, \tag{25}$$

$$e_q = (I - C_q)^{-1}C_q|\widehat{z}_q|, \qquad W_q = (I - C_q)^{-1}|\widehat{B}_q^{-1}|, \tag{26}$$

provided the spectral radius $\rho_{\mathrm{sp}}(C_q)$ is smaller than one. Absolute values and inequalities between matrices and vectors are understood componentwise.

Lemma 6 (Componentwise resolvent enclosure)**.** *Suppose $\rho_{\mathrm{sp}}(C_q) < 1$. Every matrix B in the rectangle $[L, U]$ has an invertible q-augmented system and satisfies*

$$|\mathsf{B}_q(B)^{-1}| \le W_q, \qquad |z_q(B) - \widehat{z}_q| \le e_q. \tag{27}$$

The spectral-radius test and inverse enclosure are standard strong-regularity tools for interval linear systems (Beeck, 1975; Rump, 1983; Neumaier, 1990). The proof is included to record the exact componentwise envelope passed to the subsequent game certificate.

Proof. Write $\mathsf{B}_q(B) = \widehat{B}_q + E$. Then $|E| \le D_q$ and hence

$$|\widehat{B}_q^{-1}E| \le C_q. \tag{28}$$

For every positive integer s, $|(\widehat{B}_q^{-1}E)^s| \le C_q^s$. Since C_q is nonnegative and has spectral radius smaller than one, $\sum_{s\ge 0} C_q^s = (I - C_q)^{-1}$. The Neumann series therefore converges and gives

$$|(\widehat{B}_q + E)^{-1}| \le \sum_{s\ge 0} C_q^s|\widehat{B}_q^{-1}| = W_q.$$

This also proves invertibility. If $h = |z_q(B) - \widehat{z}_q|$, the resolvent identity gives

$$h \le |\widehat{B}_q^{-1}|D_q|z_q(B)| \le C_q(|\widehat{z}_q| + h).$$

Multiplication by the nonnegative matrix $(I - C_q)^{-1}$ proves the second claim in (27). □

Let e_x^{row} and e_y^{col} denote the first d coordinates of e_x and e_y. The resolvent core returns the following observable quantities, which are used by the certificate in Section F:

$$\underline{\rho} = \min_i \{\widehat{x}_i - (e_x^{\mathrm{row}})_i\}, \qquad \underline{\tau} = \min_j \{\widehat{y}_j - (e_y^{\mathrm{col}})_j\}, \tag{29}$$

$$\underline{\Delta} = \min_{j \in [d]} \{v_{[d]\setminus\{j\}}(L) - v(U)\}, \tag{30}$$

$$k_\infty = \frac{\|e_x^{\mathrm{row}}\|_\infty}{\varepsilon}, \qquad k_2 = \frac{\|e_x^{\mathrm{row}}\|_2}{\varepsilon}. \tag{31}$$

Assume from now on that the true game is regular and fully mixed. Define

$$\begin{aligned} \rho &= \min_i x_i^\star, & \tau &= \min_j y_j^\star, \\ \kappa &= \max_{q \in \{x,y\}} \|\mathsf{B}_q(A)^{-1}\|_\infty, & \overline{\kappa} &= \max\{1, \kappa\}. \end{aligned} \tag{32}$$

Theorem 4 (Fine-radius resolvent bounds)**.** *Let $d \geq 2$, and let $[L, U]$ be any rectangle containing A. Its interval resolvent tests pass whenever its maximum radius ε satisfies $0 < \varepsilon \leq \Gamma_{\mathrm{cert}}$, where*

$$\Gamma_{\mathrm{cert}} = \min\left\{ \frac{1}{8\overline{\kappa}d}, \frac{\rho}{64\overline{\kappa}\sqrt{d}}, \frac{\tau}{32\overline{\kappa}}, \frac{\Delta_{\mathrm{del}}}{64\overline{\kappa}d} \right\}. \tag{33}$$

At such a rectangle, the computed constants obey

$$k_\infty \leq 4\overline{\kappa}, \qquad k_2 \leq 4\overline{\kappa}\sqrt{d}, \qquad \underline{\rho} \geq \rho - 6\overline{\kappa}\varepsilon, \qquad \underline{\tau} \geq \tau - 6\overline{\kappa}\varepsilon. \tag{34}$$

$$\underline{\Delta} \geq \Delta_{\mathrm{del}} - 4\varepsilon. \tag{35}$$

Proof. Let $\varepsilon \leq \Gamma_{\mathrm{cert}}$. The difference between either augmented midpoint system and its true counterpart has infinity norm at most $d\varepsilon$. Since $\kappa d\varepsilon \leq 1/8$, the Neumann series from the true system to the midpoint system gives

$$\|\widehat{B}_q^{-1}\|_\infty \leq \frac{\kappa}{1 - \kappa d\varepsilon} \leq 2\overline{\kappa}. \tag{36}$$

The row sums of D_q are at most $d\varepsilon$, so $\|C_q\|_\infty \leq 2\overline{\kappa}d\varepsilon \leq 1/4$. Both spectral radius tests therefore pass and $\|(I - C_q)^{-1}\|_\infty \leq 4/3 < 2$.

The perturbation has more structure than its operator norm records. Put $E_q = \widehat{B}_q - B_q(A)$. Its last row and last column vanish, and the strategy component of $z_q(A)$ is a probability vector. Consequently, $\|E_q z_q(A)\|_\infty \leq \varepsilon$. The solution identity and (36) give

$$\|\widehat{z}_q - z_q(A)\|_\infty \leq 2\overline{\kappa}\varepsilon.$$

The ρ and τ components of (33) therefore make the row and column strategy components of $\widehat{z}_q$ strictly positive. The last augmented equation makes their coordinates sum to one, so both midpoint strategies are probability vectors before the interval envelope is bounded. Since D_q also has zero last row and last column, $\|D_q|\widehat{z}_q|\|_\infty \leq \varepsilon$. It follows from (26) that

$$\|e_q\|_\infty \leq 2(2\overline{\kappa})\varepsilon = 4\overline{\kappa}\varepsilon. \tag{37}$$

The true solution lies in the certified interval, and (37) therefore gives

$$\underline{\rho} \geq \rho - 6\overline{\kappa}\varepsilon, \qquad \underline{\tau} \geq \tau - 6\overline{\kappa}\varepsilon.$$

The robust deletion quantity satisfies

$$\underline{\Delta} \geq \Delta_{\mathrm{del}} - 4\varepsilon,$$

because both L and U are within 2ε of A in max norm and the zero-sum value is one-Lipschitz in that norm. These are the remaining inequalities in (34) and (35). □

Proposition 5 (Box-uniform resolution obstruction)**.** *For every $d \geq 2$, the entrywise box $\{B \in [0,1]^{d \times d} : \|B - I_d\|_{\max} \leq 1/d\}$ contains a game with zero deletion surplus and an optimal row strategy with a zero coordinate. Hence no sound box-uniform certificate of a positive deletion surplus and row floor can accept that box.*

Proof. Fix a column j and put $r = 1/d$. Define $M^{(j)} \in [0,1]^{d \times d}$ as follows. Every diagonal entry equals $1 - r$; every off-diagonal entry in column j equals r; and all remaining off-diagonal entries equal zero. Then $\|M^{(j)} - I_d\|_{\max} = r$.

Let $x_j = 0$ and $x_i = 1/(d-1)$ for $i \neq j$. For every $k \neq j$,

$$x^\top M^{(j)}_{:,k} = \frac{1 - r}{d - 1} = \frac{1}{d},$$

and $x^\top M^{(j)}_{:,j} = r = 1/d$. Hence the row player can guarantee $1/d$. Conversely, let y be uniform on all d columns. A row $i \neq j$ earns $((1-r)+r)/d = 1/d$ against y, while row j earns $(1-r)/d < 1/d$. Thus $v(M^{(j)}) = 1/d$ and x is an optimal row strategy with a zero coordinate.

After deleting column j, the same x guarantees $(1-r)/(d-1) = 1/d$. The column player can match this value by mixing uniformly over the remaining $d-1$ columns, so $v_{-j}(M^{(j)}) = 1/d = v(M^{(j)})$. The deletion surplus is therefore zero. Since $M^{(j)}$ belongs to every box $\mathcal{B}_\infty(I_d, r')$ with $r' \geq 1/d$, no box-uniform certificate of positive deletion surplus and interior optimality can soundly accept such a box. □

Proposition 6 (Sequential recertification cannot cycle indefinitely)**.** *Run one lifelong nested confidence rectangle. Before each refreshed local proxy is played, rerun the spectral-resolvent tests and the aggregate certificate in* (41)*. A local chain entered from global scale k_0 sets its root target $\varepsilon_0^{\mathrm{loc}}$ equal to the current maximum confidence radius. If $\varepsilon_e^{\mathrm{loc}}$ is the actual maximum radius frozen at its eth accepted epoch, the next refresh is attempted at the first observation for which*

$$\max_{i,j} r_{ij} \leq \varepsilon_e^{\mathrm{loc}}/2.$$

Thus each newly accepted actual radius is at most half its predecessor; an update may skip several geometric levels. If the refreshed certificate rejects, preserve all observations and return to the first global scale whose target ε_k is strictly smaller than the current maximum radius, provided that this index does not exceed K_T. If it does, IC-RF enters terminal mode and plays the uniform row distribution for the remaining horizon. After $K_T + 1$ accepted local epochs, the last accepted proxy is not refreshed. Then no rejected proxy is played, every nonterminal return strictly increases the global scale index, and each nonterminal global scale is entered at most once. On the lifelong confidence event, after the maximum radius first falls below Γ_{cert}, every subsequent recertification accepts.

Proof. The state-transition rule tests a proxy before choosing an action from it, so a rejected proxy produces no local round. At a rejected refresh, choose the least k with $\varepsilon_k < \max_{i,j} r_{ij}$. At least one column is then outside the active face, which prevents immediate repetition of the same full-face certificate. Every local chain, including one entered after a previous rejection, receives a fresh entry index k_0. Entry is possible only from a full active face at scale k_0, so $\max_{i,j} r_{ij} \leq \varepsilon_{k_0}$ at that entry and hence $\varepsilon_0^{\mathrm{loc}} \leq \varepsilon_{k_0}$. Therefore, at the eth attempted refresh,

$$\max_{i,j} r_{ij} \leq \frac{\varepsilon_0^{\mathrm{loc}}}{2^e} \leq \varepsilon_{k_0+e}.$$

Therefore the least scale whose target is strictly below the maximum radius has index greater than $k_0 + e$, hence greater than k_0. If that index is at most K_T, the global index strictly increases. Otherwise the terminal rule ends the core state machine, so it cannot create a zero-round loop. The accepted-epoch cap suppresses any further refresh.

On the confidence event the true matrix remains in every nested rectangle. Once the maximum radius is at most Γ_{cert}, it remains so. Theorem 6 then makes every later certificate accept, so no further local-to-global return is possible. □

The monotone-scale conclusion lets the proof of Theorem 3 combine all global blocks in one geometric ledger: rejected recertifications separate the blocks but never cause a scale to be charged twice.

F Observable certificate for local control

The certificate combines the interval resolvent test with the local quantities needed by the two-expert controller. The augmented systems, componentwise resolvent enclosures, and definitions of $\underline{\rho}$, $\underline{\tau}$, k_∞, and k_2 are those of Section E. For each column j, retain a lexicographically selected maximin row policy r_j of L_{-j} and define

$$\delta_j = v_{-j}(L) - v(U), \qquad \underline{\Delta} = \min_j \delta_j, \qquad d_j = \|r_j - \widehat{x}\|_2. \tag{38}$$

Set

$$\mathcal{R}_+ = k_2 + 2\sqrt{2}, \qquad \alpha_j = \begin{cases} 1, & d_j \leq k_2\varepsilon, \\ \min\left\{1, \dfrac{2\sqrt{2}\,\varepsilon}{d_j - k_2\varepsilon}\right\}, & d_j > k_2\varepsilon, \end{cases} \tag{39}$$

Let $P = I - \mathbf{1}\mathbf{1}^\top/d$ be the orthogonal projection onto the tangent space of the simplex. Define

$$\begin{aligned} S_{\min} &= \min_j \alpha_j \delta_j, & \overline{\mathcal{R}} &= \max\{1, \mathcal{R}_+\}, \\ H &= \max_j \|PM_{:j}\|_2, & D &= \max_j \frac{\|R_{:j}\|_2}{\varepsilon}, \\ C_+ &= 2^{\lceil \log_2 \overline{\mathcal{R}} \rceil}, & \lambda_+ &= (\sqrt{d}\, C_+)^{-1}, \\ \mathcal{G}_{\mathrm{F}} &= \frac{H^2}{2} + 2\mathcal{R}_+ D, & \mathcal{G}_{\mathrm{M}}^{(2)} &= \frac{\lambda_+ \mathcal{R}_+^2 H^2}{8} + \mathcal{R}_+ D, \\ \mathcal{G}^{(2)} &= \frac{\mathcal{G}_{\mathrm{F}} + \mathcal{G}_{\mathrm{M}}^{(2)}}{2}. \end{aligned} \tag{40}$$

If a larger cap is already active, the certificate uses that stored cap in place of C_+. This only decreases the master-curvature term. In addition to the two spectral-radius tests, the aggregate certificate accepts when

$$\begin{gathered} \underline{\rho} > 0, \qquad \underline{\tau} > 0, \qquad \underline{\Delta} > 0, \qquad \underline{\rho} \geq 2(\mathcal{R}_+ + k_\infty)\varepsilon, \\ \varepsilon \leq \frac{1}{2}, \qquad 4\varepsilon \leq \underline{\Delta}, \qquad \frac{16}{7}\mathcal{G}^{(2)}\varepsilon^2 \leq S_{\min}. \end{gathered} \tag{41}$$

Define the certified local set

$$\mathcal{X}^+ = \{x \in \Delta_d : \|x - \widehat{x}\|_2 \leq \mathcal{R}_+ \varepsilon\}. \tag{42}$$

Lemma 7 (Observable restricted-face interpolation)**.** *Suppose $A \in [L, U]$ and the certificate's spectral-radius and positivity tests hold, so the resolvent system supplies its unique fully mixed saddle $x^\star$. For every column j, define*

$$z_j = (1 - \alpha_j)x^\star + \alpha_j r_j.$$

Then $z_j \in \mathcal{X}^+$. Moreover, for every $\ell \neq j$,

$$(r_j - x^\star)^\top A_{:,\ell} \geq \delta_j, \qquad (r_j - x^\star)^\top A_{:,j} \geq -1, \qquad \alpha_j \geq 2\varepsilon.$$

Proof. The resolvent enclosure gives $\|x^\star - \widehat{x}\|_2 \leq k_2\varepsilon$. If $d_j \leq k_2\varepsilon$, convexity places the segment from $x^\star$ to r_j inside the ball of radius $k_2\varepsilon$ about $\widehat{x}$. Otherwise,

$$\|z_j - \widehat{x}\|_2 \leq k_2\varepsilon + \alpha_j(d_j - k_2\varepsilon) \leq (k_2 + 2\sqrt{2})\varepsilon.$$

Thus $z_j \in \mathcal{X}^+$. For $\ell \neq j$,

$$r_j^\top A_{:,\ell} \geq r_j^\top L_{:,\ell} \geq v_{-j}(L) = v(U) + \delta_j \geq v(A) + \delta_j.$$

Full support gives $(x^\star)^\top A_{:,\ell} = v(A)$ for every column. The payoff range gives the bound on column j. Finally $d_j \leq \sqrt{2}$. The second branch defining α_j is therefore at least 2ε before truncation, and truncation at one preserves the bound because $\varepsilon \leq 1/2$. The first branch has $\alpha_j = 1$. □

Theorem 5 (Soundness of the aggregate certificate)**.** *Suppose $A \in [L, U]$ and the tests in* (41) *accept. Then A has a unique fully mixed saddle $(x^\star, y^\star)$ and*

$$\min_i x_i^\star \geq \underline{\rho}, \qquad \min_j y_j^\star \geq \underline{\tau}, \qquad \Delta_{\mathrm{del}} \geq \underline{\Delta}. \tag{43}$$

Every policy in (42) *has every row probability at least $\underline{\rho}/2$. The policies and interpolation factors in* (38)– (39) *satisfy all conclusions of Lemma 7.*

Proof. Lemma 6 applied to both augmented systems places their normalized solutions inside the simplex with coordinate lower bounds $\underline{\rho}$ and $\underline{\tau}$. The two equal-payoff systems have the same value after multiplication by the opposite strategy, and therefore form a fully mixed saddle. Invertibility of both augmented systems makes this saddle unique. This proves the first two inequalities in (43). For every j, monotonicity of the zero-sum value gives

$$v_{-j}(A) - v(A) \geq v_{-j}(L) - v(U) = \delta_j,$$

which proves the deletion bound.

The enclosure gives $\|\widehat{x} - x^\star\|_\infty \leq k_\infty\varepsilon$. For every $x \in \mathcal{X}^+$,

$$x_i \geq x_i^\star - (\mathcal{R}_+ + k_\infty)\varepsilon \geq \underline{\rho}/2.$$

The remaining conclusions are proved directly in Lemma 7; that argument uses only the enclosure, the definitions above, and $\varepsilon \leq 1/2$. □

Theorem 6 (Fine-box acceptance and accepted-parameter comparison)**.** *Let A be a regular fully mixed instance and let Γ_{cert} be the sufficient radius in* (33)*. Every confidence rectangle containing A with maximum radius $0 < \varepsilon \le \Gamma_{\mathrm{cert}}$ passes the aggregate certificate. At every such fine rectangle,*

$$\mathcal{R}_+ \le 6\overline{\kappa}\sqrt{d}, \qquad \mathcal{G}^{(2)} \le 14\overline{\kappa}d. \tag{44}$$

At every accepted rectangle, regardless of its radius,

$$\underline{\rho} \ge \rho/2, \qquad \underline{\Delta} \ge \Delta_{\mathrm{del}}/2. \tag{45}$$

Proof. The resolvent calculations in the proof of Theorem 4 give

$$k_\infty \le 4\overline{\kappa}, \qquad k_2 \le 4\overline{\kappa}\sqrt{d}, \qquad \underline{\rho} \ge \rho - 6\overline{\kappa}\varepsilon, \qquad \underline{\Delta} \ge \Delta_{\mathrm{del}} - 4\varepsilon.$$

Because $d \ge 2$ and $\overline{\kappa} \ge 1$, $2\sqrt{2} \le 2\overline{\kappa}\sqrt{d}$. This proves the first inequality in (44). For the curvature constant, retain the sharper Neumann factors. Since $\kappa d\varepsilon \le 1/8$,

$$\|\widehat{B}_q^{-1}\|_\infty \le 8\kappa/7, \quad \|C_q\|_\infty \le 1/7, \quad \|(I - C_q)^{-1}\|_\infty \le 7/6.$$

The midpoint strategies are probability vectors by the preceding resolvent argument. Hence $\|D_q|\widehat{z}_q|\|_\infty \le \varepsilon$, $k_2 \le (4/3)\overline{\kappa}\sqrt{d}$, and $\mathcal{R}_+ \le (10/3)\overline{\kappa}\sqrt{d}$. Thus $\mathcal{G}_{\mathrm{F}} \le d/2 + (20/3)\overline{\kappa}d$. The two constituent curvature bounds are

$$\mathcal{G}_{\mathrm{F}} \le 11\overline{\kappa}d, \qquad \mathcal{G}_{\mathrm{M}}^{(2)} \le \frac{\mathcal{R}_+ H^2}{8\sqrt{d}} + \mathcal{R}_+\sqrt{d} \le 15\overline{\kappa}d.$$

Their average is at most $13\overline{\kappa}d$, and hence at most $14\overline{\kappa}d$, proving the second inequality in (44).

The ρ component of Γ_{cert} yields

$$2(\mathcal{R}_+ + k_\infty)\varepsilon \le 20\overline{\kappa}\sqrt{d}\,\varepsilon \le \frac{5}{16}\rho < \underline{\rho}.$$

The same radius bound makes $\underline{\rho}$ positive, and the τ component makes $\underline{\tau}$ positive. The deletion component gives

$$4\varepsilon \le \frac{\Delta_{\mathrm{del}}}{16\overline{\kappa}d} < \underline{\Delta}, \qquad 2\mathcal{G}^{(2)}\varepsilon \le \frac{7}{16}\Delta_{\mathrm{del}} < \underline{\Delta}.$$

Also $\varepsilon \le 1/2$. Since both r_j and $\widehat{x}$ are distributions, $d_j \le \sqrt{2}$. The definition of α_j therefore gives $\alpha_j \ge 2\varepsilon$, and hence

$$S_{\min} \ge 2\varepsilon\underline{\Delta} \ge \frac{16}{7}\mathcal{G}^{(2)}\varepsilon^2.$$

All aggregate tests accept.

For an arbitrary accepted radius, the same argument as in Theorem 4 gives $\rho \le \underline{\rho} + 2k_\infty\varepsilon$. The enlarged row-floor test implies $2k_\infty\varepsilon \le \underline{\rho}$, proving the first inequality in (45). Value Lipschitzness gives $\Delta_{\mathrm{del}} \le \underline{\Delta} + 4\varepsilon$, while the explicit gap-resolution test gives $4\varepsilon \le \underline{\Delta}$. This proves the second inequality. □

F.1 An observable radius choice

The fixed bonus $2\sqrt{2}$ in the local radius is sufficient for eventual acceptance, but it need not give the best finite-sample balance between the restricted-face surplus and the second-order online-learning term. This section gives a certified data-dependent choice. It changes neither the confidence event nor the target regret.

Let $b_0 = 2\sqrt{2}$ and let $\mathcal{B}$ be the finite geometric grid

$$\mathcal{B} = \{b_0 2^{q/2} : q \in \mathbb{N}_0,\ b_0 2^{q/2} \le 8\sqrt{d}\} \cup \{b_{\mathrm{edge}}\}, \tag{46}$$

where b_{edge} is included only when it lies in $[b_0, 8\sqrt{d}]$ and satisfies the strict row-floor inequality below. For $b \in \mathcal{B}$, define

$$\begin{aligned} \mathcal{R}(b) &= k_2 + b, \\ \alpha_j(b) &= \begin{cases} 1, & d_j \le k_2\varepsilon, \\ \min\left\{1, \dfrac{b\varepsilon}{d_j - k_2\varepsilon}\right\}, & d_j > k_2\varepsilon, \end{cases} \\ S(b) &= \min_j \alpha_j(b)\delta_j. \end{aligned} \tag{47}$$

Let $C_{\mathrm{prev}} \ge 1$ be the stored dyadic cap, with $C_{\mathrm{prev}} = 1$ before the first accepted epoch. Candidate b is scored with the already determined cap and learning rate

$$C(b) = \max\left\{C_{\mathrm{prev}}, 2^{\lceil \log_2 \max\{1, \mathcal{R}(b)\}\rceil}\right\}, \qquad \lambda(b) = \frac{1}{\sqrt{d}\,C(b)}.$$

Its two-expert curvature is

$$\begin{aligned}\mathcal{G}_{\mathrm{F}}(b) &= \frac{H^2}{2} + 2\mathcal{R}(b)D,\\ \mathcal{G}_{\mathrm{M}}^{(2)}(b) &= \frac{\lambda(b)\mathcal{R}(b)^2H^2}{8} + \mathcal{R}(b)D,\\ \mathcal{G}^{(2)}(b) &= \frac{\mathcal{G}_{\mathrm{F}}(b) + \mathcal{G}_{\mathrm{M}}^{(2)}(b)}{2}.\end{aligned} \tag{48}$$

All quantities in these definitions are observable at certification time. The algorithm chooses a grid point maximizing

$$\min\left\{\frac{\underline{\rho}}{2(\mathcal{R}(b)+k_\infty)\varepsilon}, \frac{7S(b)}{16\mathcal{G}^{(2)}(b)\varepsilon^2}\right\}, \tag{49}$$

with deterministic tie breaking. Fix any rational $\eta_{\mathrm{edge}} \in (0,1)$ before play and define

$$b_{\mathrm{edge}} = (1-\eta_{\mathrm{edge}})\min\left\{8\sqrt{d}, \frac{\underline{\rho}}{2\varepsilon} - k_2 - k_\infty\right\}.$$

Retain this endpoint only if it lies in $[b_0, 8\sqrt{d}]$; it then satisfies the strict row-floor test. The guarantees below hold for every such fixed slack and also if the endpoint is omitted, because the geometric grid always retains b_0.

Lemma 8 (Soundness of radius adaptation)**.** *Suppose the confidence rectangle contains A. For every $b \in [2\sqrt{2}, 8\sqrt{d}]$, if*

$$\underline{\rho} \geq 2(\mathcal{R}(b)+k_\infty)\varepsilon, \qquad \frac{16}{7}\mathcal{G}^{(2)}(b)\varepsilon^2 \leq S(b), \tag{50}$$

then the two-expert local lemmas remain valid with $\mathcal{R}(b)$, $\alpha_j(b)$, $S(b)$, and $\mathcal{G}^{(2)}(b)$ in place of their fixed-radius counterparts.

Proof. Fix a column j and set $z_j = (1-\alpha_j(b))x^\star + \alpha_j(b)r_j$. The interval resolvent gives $\|x^\star - \widehat{x}\|_2 \leq k_2\varepsilon$. If $d_j \leq k_2\varepsilon$, convexity places the whole segment between $x^\star$ and r_j inside the ball of radius $k_2\varepsilon$ about $\widehat{x}$. Otherwise, the triangle inequality and Equation (47) give

$$\|z_j - \widehat{x}\|_2 \leq k_2\varepsilon + \alpha_j(b)(d_j - k_2\varepsilon) \leq (k_2+b)\varepsilon.$$

Thus z_j belongs to the certified local set. Since two probability vectors are at Euclidean distance at most $\sqrt{2}$ and $b \geq 2\sqrt{2}$, the second branch satisfies

$$\alpha_j(b) \geq \min\{1, b\varepsilon/\sqrt{2}\} \geq 2\varepsilon,$$

where the aggregate certificate also requires $\varepsilon \leq 1/2$. The first branch has $\alpha_j(b) = 1$ and obeys the same inequality.

The first inequality in (50), together with the coordinatewise saddle enclosure, places every policy in the enlarged local set at row mass at least $\underline{\rho}/2$. The FTRL and exponential- weights potential calculations depend on the local set only through its radius. Repeating them with $\mathcal{R}(b)$ gives the curvature $\mathcal{G}^{(2)}(b)$. Set

$$\beta_b = \frac{\underline{\Delta}}{8(1+\underline{\Delta})}.$$

The second inequality in (50) bounds the pathwise second-order term by $7\alpha_j(b)\delta_j/8$ per round. When the least frequent column appears on at most $\beta_b s$ rounds, its adverse term is at most $\alpha_j(b)\delta_j/8$ per round. These two charges exactly exhaust the restricted-face surplus. When the column is more frequent, every column appears more than $\beta_b s$ times; Lemma 10 then bounds the epoch length by twice the quarter-share visit bound. This is an instance-dependent constant with the same dependence on the horizon. □

Proposition 7 (The adaptive test cannot delay certification)**.** *If the fixed-radius aggregate certificate accepts a confidence rectangle, then the adaptive-radius certificate also accepts it. Consequently the eventual-acceptance radius and the fixed-instance polylogarithmic dependence on the horizon remain valid.*

Proof. The grid contains $b_0 = 2\sqrt{2}$. At this point every quantity and every test equals its fixed-radius counterpart. Hence acceptance of the fixed certificate makes both ratios in (49) at least one. The selected grid point has score no smaller than the score at b_0, so it satisfies both inequalities in (50). All remaining certificate tests are independent of b. Lemma 8 therefore proves soundness. Inclusion of b_0 also gives the fine-box acceptance conclusion of Theorem 6. Finally, $\mathcal{R}(b) \leq k_2 + 8\sqrt{d}$ is independent of the horizon on every fixed instance, so the local epoch liabilities remain instance-dependent constants and only the number of refinements grows logarithmically. □

F.2 IDENTITY-GAME CALIBRATION FOR THE OBSERVABLE SADDLE CERTIFICATE

The general certificate is uniform over the regular fully mixed class. At the identity matrix, the deletion policies have additional geometry that gives a sharper sufficient resolution. The result below concerns the certificate used by the persistent two-expert controller; it is not an information-theoretic lower bound.

Proposition 8 (Identity-game certificate resolution)**.** *Let $A = I_d$ with $d \geq 2$. Every confidence rectangle $[L, U]$ containing I_d and having maximum entrywise half-width*

$$0 < \varepsilon \leq \Gamma_I := \frac{1}{512d^2}$$

passes the adaptive-radius observable saddle certificate for every retained cap rate $0 < \lambda \leq d^{-1/2}$, including the rate considered at first entry. Consequently, with probability at least $1 - \delta$,

$$\mathrm{Reg}_T = O\big(d^8\mathcal{L} + d^4\mathcal{L}\log(T \vee e)\big) \qquad (A = I_d).$$

Proof. For the identity game, $x^\star = y^\star = \mathbf{1}/d$, $v = 1/d$, and $\Delta_{\mathrm{del}} = 1/[d(d-1)]$. The two augmented saddle-system inverses are both, up to the transpose convention,

$$\begin{pmatrix} I_d - \mathbf{1}\mathbf{1}^\top/d & \mathbf{1}/d \\ -\mathbf{1}^\top/d & 1/d \end{pmatrix}.$$

Their maximum absolute row sum is $2 - 1/d \leq 2$. Applying the resolvent calculation from Lemma 6, and using $2d\varepsilon \leq 1/8$, gives

$$k_\infty \leq 8, \qquad k_2 \leq 8\sqrt{d}, \qquad \|\widehat{x} - x^\star\|_2 \leq 4\sqrt{d}\,\varepsilon,$$

and

$$\underline{\rho}, \underline{\tau} \geq \frac{1}{d} - 12\varepsilon, \qquad \underline{\Delta} \geq \frac{1}{d(d-1)} - 4\varepsilon.$$

The same inequalities make the midpoint saddle strictly positive, so the simplex-aware deletion-policy calculation applies.

Fix a column j. Let q_j put mass $1/(d-1)$ on every row except j, and let r_j be the lexicographically selected row maximin policy of the lower-confidence game with column j deleted. Since the rectangle is contained in $[0,1]^{d\times d}$ and contains I_d, the off-diagonal entries of L vanish and $1 - 2\varepsilon \leq L_{\ell\ell} \leq 1$. The policy q_j guarantees $(1-2\varepsilon)/(d-1)$ in L_{-j}. For every $\ell \neq j$,

$$(r_j)_\ell \geq (r_j)_\ell L_{\ell\ell} \geq v_{-j}(L) \geq \frac{1-2\varepsilon}{d-1} \geq \frac{1}{d-1} - 2\varepsilon.$$

The negative coordinate deviations of $r_j - q_j$ therefore sum to at most $2(d-1)\varepsilon$. Since all deviations sum to zero,

$$\|r_j - q_j\|_2 \leq \|r_j - q_j\|_1 \leq 4(d-1)\varepsilon.$$

Moreover, $\|q_j - x^\star\|_2 = 1/\sqrt{d(d-1)}$. Hence

$$d_j = \|r_j - \widehat{x}\|_2 \leq \frac{1}{\sqrt{d(d-1)}} + 4(d-1)\varepsilon + 4\sqrt{d}\,\varepsilon < \frac{2}{d}.$$

The adaptive grid always contains $b_0 = 2\sqrt{2}$. With $\mathcal{R}(b_0) = k_2 + b_0$, the interpolation rule in Equation (47) therefore gives $\alpha_j(b_0) \geq d\varepsilon$. Also $\underline{\Delta} \geq 1/(2d^2)$ at the stated radius, and thus

$$S(b_0) = \min_j \alpha_j(b_0)\delta_j \geq \frac{\varepsilon}{2d}.$$

At this grid point,

$$\mathcal{R}(b_0) \leq 10\sqrt{d}, \qquad D \leq \sqrt{d}, \qquad H^2 \leq 2.$$

For every retained $\lambda \leq d^{-1/2}$, the two-expert curvature in Equation (48) is at most $60d$. Consequently,

$$\frac{16}{7}\mathcal{G}^{(2)}(b_0)\varepsilon^2 \leq \frac{960}{7}d\varepsilon^2 < \frac{\varepsilon}{2d} \leq S(b_0).$$

The row-floor left side is bounded by

$$2\{\mathcal{R}(b_0) + k_\infty\}\varepsilon \leq 32\sqrt{d}\,\varepsilon \leq \frac{1}{16d} < \underline{\rho}.$$

The preceding resolvent bounds also verify the spectral tests, while $\varepsilon \leq 1/2$ and $4\varepsilon \leq \underline{\Delta}$ verify the remaining scalar tests. Thus b_0 is feasible. The grid maximizer has no smaller score, so the adaptive certificate accepts as well.

For the regret bound, the global ledger evaluated at Γ_I contributes $O(d^8\mathcal{L})$. It remains to bound local epochs that may have been accepted earlier. The gap-resolution test gives $\varepsilon_e \le \Delta_{\mathrm{del}}/4$. For $d \ge 3$, this implies $2d\varepsilon_e \le 1/4$; repeating the preceding resolvent and simplex-aware estimates yields

$$k_{2,e} \le 4\sqrt{d}, \quad \mathcal{R}_e \le 12\sqrt{d}, \quad H_e^2 \le 2, \quad D_e^{\mathrm{norm}} \le \sqrt{d}, \quad \mathcal{G}_e^{(2)} \le 66d.$$

The accepted-parameter comparison also gives

$$\rho_e \ge \frac{1}{4d}, \qquad \beta_e \ge \frac{1}{32d(d-1)}, \qquad \lambda_e^{-1} < 24d.$$

The last inequality includes dyadic rounding: every retained cap is less than twice the largest preceding value of $\max\{1, \mathcal{R}_e\}$. Substitution in Theorem 7, followed by the geometric cap-restart bound and the self-financing epoch liabilities, gives $O(d\log(T \vee e) + d^4\mathcal{L}\log(T \vee e))$. When $d = 2$, all identity instance parameters are fixed numerical constants and Theorem 3 gives the same displayed order. Adding the global and local contributions completes the proof. □

Proposition 5 gives a box obstruction at order d^{-1}, whereas Proposition 8 proves sufficiency at order d^{-2}. The remaining factor of d is a certificate-resolution gap; neither proposition determines the optimal regret dependence on the dimension.

G CERTIFIED-CENTER FTRL AND CONFIDENCE-WIDTH REFRESH

We now establish the pathwise FTRL inequality and the sampling statement used in every accepted local epoch. The certified restricted comparator is constructed later in Lemma 7.

Fix an accepted local epoch with midpoint proxy M, radius matrix R, maximum radius ε, and certificate outputs $\underline{\rho}$ and $\mathcal{R}$. Define

$$\rho_{\mathrm{floor}} = \frac{\underline{\rho}}{2}. \tag{51}$$

With $P = I - \mathbf{1}\mathbf{1}^\top/d$, set

$$H = \max_j \|PM_{:j}\|_2, \qquad D = \max_j \frac{\|R_{:j}\|_2}{\varepsilon}, \qquad \mathcal{G}_{\mathrm{F}} = \frac{H^2}{2} + 2\mathcal{R}D.$$

Let $\widehat{x}$ be the row center returned by the augmented midpoint system and let

$$\mathcal{X} = \{x \in \Delta_d : \|x - \widehat{x}\|_2 \le \mathcal{R}\varepsilon\}$$

be the certified local set.

G.1 THE ADAPTIVE FTRL SEGMENT

Set $\eta = \varepsilon^2$. Starting with zero cumulative gradient, the adaptive segment plays

$$x_t = \arg\max_{x \in \mathcal{X}} \left\{ \left\langle x, \sum_{s<t} M_{:j_s} \right\rangle - \frac{\|x - \widehat{x}\|_2^2}{2\eta} \right\}. \tag{52}$$

This optimization is well-defined because $\mathcal{X}$ is compact and convex. The opponent may choose every j_t after seeing x_t.

Lemma 9 (Uniform pathwise FTRL inequality)**.** *For every comparator $z \in \mathcal{X}$ and every prefix of length s of the adaptive segment,*

$$\sum_{t=1}^{s} \langle z - x_t, A_{:j_t} \rangle \le \frac{\mathcal{R}^2}{2} + \mathcal{G}_{\mathrm{F}} s\varepsilon^2. \tag{53}$$

The inequality is deterministic for the realized column sequence.

Proof. Put $\psi(x) = \|x - \widehat{x}\|_2^2/2 + I_{\mathcal{X}}(x)$, where $I_{\mathcal{X}}$ is zero on $\mathcal{X}$ and infinite elsewhere, and define

$$F(u) = \sup_x \{\langle u, x\rangle - \psi(x)/\eta\}.$$

The function ψ/η is $1/\eta$-strongly convex, so its conjugate F is η-smooth. If $g_t = PM_{:j_t}$ and $S_t = \sum_{s \le t} g_s$, then $x_t = \nabla F(S_{t-1})$ and smoothness gives

$$F(S_t) \le F(S_{t-1}) + \langle x_t, g_t\rangle + \frac{\eta}{2}\|g_t\|_2^2.$$

Summing, using $F(0) = 0$, and lower-bounding $F(S_s) \geq \langle S_s, z\rangle - \psi(z)/\eta$ gives

$$\sum_{t=1}^{s} \langle z - x_t, M_{:j_t}\rangle \leq \frac{\|z - \widehat{x}\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{s} \|PM_{:j_t}\|_2^2, \tag{54}$$

To see why the projected gradients are valid, replace each $M_{:j_t}$ by $PM_{:j_t}$, where $P = I - \mathbf{1}\mathbf{1}^\top/d$. This subtracts a scalar multiple of $\mathbf{1}$, which changes neither the maximizer over the simplex nor any inner product with $z - x_t$.

Both z and x_t lie in the ball of radius $\mathcal{R}\varepsilon$ around $\widehat{x}$. Since $|A - M| \leq R$ componentwise,

$$|\langle z - x_t, A_{:j_t} - M_{:j_t}\rangle| \leq 2\mathcal{R}\varepsilon\|R_{:j_t}\|_2.$$

Use $\eta = \varepsilon^2$ and the definition of $\mathcal{G}_{\mathrm{F}}$ to obtain (53). □

G.2 The observable refresh time

Retain the logarithmic factor $\mathcal{L}$ from (12). For an epoch with target ε, define the deterministic proof envelopes

$$N^{\mathrm{ref}}(\varepsilon) = \left\lceil \frac{36\mathcal{L}}{\varepsilon^2} \right\rceil, \qquad a(\varepsilon, \rho_{\mathrm{floor}}) = \left\lceil \frac{2\{N^{\mathrm{ref}}(\varepsilon) + 4\mathcal{L}\}}{\rho_{\mathrm{floor}}} \right\rceil. \tag{55}$$

These quantities are not stopping rules. The epoch refreshes as soon as the observable maximum confidence radius is at most $\varepsilon/2$.

Lemma 10 (Appearance-to-refresh conversion)**.** *There is an event $\mathcal{E}_{\mathrm{local}}$ of probability at least $1 - \delta/4$ on which the following statement holds simultaneously for every accepted adaptive segment and every column. Once a column appears $a(\varepsilon, \rho_{\mathrm{floor}})$ times in that segment, each entry in the column has received at least $N^{\mathrm{ref}}(\varepsilon)$ fresh observations. If every column reaches this appearance count, the confidence-width refresh has occurred.*

Proof. By Theorem 5, every action in $\mathcal{X}$ assigns each row conditional probability at least ρ_{floor}. Condition on the random history at which an adaptive segment begins and index the subsequent appearances of a fixed column. Complete the row-sampling indicators with ones if the segment or horizon ends. Their conditional means remain at least ρ_{floor}. Let $C_i(a)$ be the completed sum through $a = a(\varepsilon, \rho_{\mathrm{floor}})$ appearances; it equals the actual fresh sample count whenever these appearances occur within the segment. A multiplicative martingale Chernoff bound gives, for every row,

$$\mathbb{P}\left(C_i(a) < N^{\mathrm{ref}}(\varepsilon) \,\middle|\, \text{start history}\right) \leq e^{-\mathcal{L}}.$$

Indeed, the conditional mean count is at least $\rho_{\mathrm{floor}} a \geq 2\{N^{\mathrm{ref}}(\varepsilon) + 4\mathcal{L}\}$; its lower half is above $N^{\mathrm{ref}}(\varepsilon)$, and the usual lower-tail exponent is at least $\rho_{\mathrm{floor}} a/8 \geq \mathcal{L}$.

The state machine admits at most $K_T + 1$ accepted local epochs, and hence at most $K_T + 1$ adaptive segments. A union bound over their d^2 row–column pairs costs at most $d^2(K_T + 1)e^{-\mathcal{L}} \leq \delta/4$. Conditioning is removed by the tower property.

Finally, $4a_1^2 < 9$ and $4a_2 < 9$. Corollary 2, applied at target $\varepsilon/2$, shows that $N^{\mathrm{ref}}(\varepsilon)$ total observations of every entry suffice for every radius to be at most $\varepsilon/2$. Fresh observations alone meet this sufficient condition, so lifelong observations can only make the refresh earlier. □

G.3 A self-financing local controller

This section gives the pathwise guarantee used by the finite-sample local controller. Fix an accepted local epoch e. Let M_e be its frozen confidence midpoint, let R_e be the matrix of entrywise confidence half-widths, and let

$$\mathcal{X}_e = \{x \in \Delta_d : \|x - \widehat{x}_e\|_2 \leq a_e\}$$

be the certified local set. Here $a_e = \mathcal{R}_e\varepsilon_e$. The dyadic-cap two-expert controller of Section H produces an anchor $h_t \in \mathcal{X}_e$. A second predictable controller may produce any proposal $q_t \in \mathcal{X}_e$. The implementation obtains q_t by combining the persistent capped controller and a scale-free controller over the same center–FTRL pair. The argument below only uses predictability and membership in $\mathcal{X}_e$.

For completeness, the implemented proposal is specified as follows. Let $c_{t,1} = \widehat{x}_e$ and $c_{t,2} = f_t$ be the center and centered-FTRL experts. Before round t, an inner weight $\omega_t \in \Delta_2$ forms $s_t = \omega_{t,1}c_{t,1} + \omega_{t,2}c_{t,2}$. After j_t is revealed, define $u_{t,r} = c_{t,r}^\top M_{e,:,j_t}$ and

$$\widetilde{u}_t = u_t - \frac{u_{t,1} + u_{t,2}}{2}\mathbf{1}.$$

With $V_t = \sum_{r \le t} \|\widetilde{u}_r\|_2^2$, update

$$\omega_{t+1} = \Pi_{\Delta_2}(\omega_t + \eta_t \widetilde{u}_t), \qquad \eta_t = \frac{\sqrt{2}}{\sqrt{V_t}},$$

using $\omega_{t+1} = \omega_t$ when $V_t = 0$. An outer weight $\nu_t \in \Delta_2$ applies the same update to the proxy-payoff pair $(h_t^\top M_{e,:,j_t}, s_t^\top M_{e,:,j_t})$ and forms

$$q_t = \nu_{t,1} h_t + \nu_{t,2} s_t.$$

The inner and outer weights and their variation sums persist across local refinements and reset only when the dyadic cap grows. Projection onto Δ_2 keeps s_t and q_t in the convex set $\mathcal{X}_e$.

After column j_t is revealed, define the observable lower confidence gain

$$\underline{g}_t = (q_t - h_t)^\top M_{e,:,j_t} - |q_t - h_t|^\top R_{e,:,j_t}. \tag{56}$$

Let $P = I - \mathbf{1}\mathbf{1}^\top/d$ be the projection onto the tangent space of the simplex, and put

$$H_e = \max_j \|P M_{e,:,j}\|_2, \qquad J_e = \max_j \|R_{e,:,j}\|_2, \qquad B_e = 2a_e(H_e + J_e).$$

The controller initializes $Z_e = B_e$. On each round it sets

$$\alpha_t = \min\{1, Z_t/B_e\}, \qquad x_t = (1-\alpha_t)h_t + \alpha_t q_t, \qquad Z_{t+1} = Z_t + \alpha_t \underline{g}_t. \tag{57}$$

When $B_e = 0$, the convention is $\alpha_t = 1$.

Lemma 11 (Self-financing domination)**.** *On the simultaneous confidence event, $Z_t \ge 0$ throughout epoch e, and every prefix I of that epoch satisfies*

$$\sum_{t \in I} \left(v(A) - x_t^\top A_{:,j_t} \right) \le \sum_{t \in I} \left(v(A) - h_t^\top A_{:,j_t} \right) + B_e.$$

Proof. Because $q_t - h_t$ sums to zero and both policies belong to $\mathcal{X}_e$,

$$\|q_t - h_t\|_2 \le 2a_e, \qquad |(q_t - h_t)^\top M_{e,:,j_t}| \le 2a_e H_e.$$

Cauchy–Schwarz also gives

$$|q_t - h_t|^\top R_{e,:,j_t} \le 2a_e J_e.$$

Hence $\underline{g}_t \ge -B_e$. If $B_e > 0$, the update in (57) therefore yields

$$Z_{t+1} \ge Z_t - \alpha_t B_e = Z_t - \min\{B_e, Z_t\} \ge 0.$$

The confidence event and (56) imply

$$(q_t - h_t)^\top A_{:,j_t} \ge \underline{g}_t.$$

Consequently, for every epoch prefix I,

$$\begin{aligned} &\sum_{t \in I} \left(v(A) - x_t^\top A_{:,j_t} \right) - \sum_{t \in I} \left(v(A) - h_t^\top A_{:,j_t} \right) \\ &\quad = -\sum_{t \in I} \alpha_t (q_t - h_t)^\top A_{:,j_t} \\ &\quad \le -\sum_{t \in I} \alpha_t \underline{g}_t = Z_e - Z_{\max I + 1} \le B_e. \end{aligned}$$

If $B_e = 0$, the two displayed norm bounds make the true payoff difference zero, and the conclusion follows directly. □

Corollary 1 (Preservation of the local regret theorem)**.** *Replace the published anchor in every accepted local epoch by the policy in* (57)*. The bound in Theorem 7 increases by at most* $\sum_e B_e$*. In particular, geometric confidence refinement preserves its polylogarithmic horizon dependence.*

Proof. Apply Lemma 11 separately to every local epoch and add the resulting inequalities to the global-stage bound. Index all accepted epochs consecutively, including epochs separated by a global-mode return. Nestedness and the refresh rule give $\varepsilon_{e+1} \le \varepsilon_e/2$. On the confidence event the accepted parameter bounds give a finite instance-dependent $\mathcal{R}_{\max} \ge \sup_e \mathcal{R}_e$. Since $H_e \le \sqrt{d}$, $J_e \le \sqrt{d}\varepsilon_e$, and $\varepsilon_e \le 1/2$, we have

$$B_e = 2\mathcal{R}_e\varepsilon_e(H_e + J_e) \le 3\sqrt{d}\mathcal{R}_{\max}\varepsilon_e, \qquad \sum_e B_e \le 6\sqrt{d}\mathcal{R}_{\max}\varepsilon_1.$$

This bounds the total publication liability without requiring $a_e = \mathcal{R}_e\varepsilon_e$ to be monotone. The main theorem uses a looser per-epoch bound, which is also valid. □

H DYADIC-CAP TWO-EXPERT ANALYSIS

This section analyzes the persistent two-expert anchor in the IC-RF state machine. Its inputs are the confidence event, the certified local set, the restricted comparator, and the appearance-to-refresh bound established in the preceding sections.

Throughout this section, set $\Lambda_T = \log\{8(T \vee 2)\}$.

For an accepted local epoch e, let I_e be its local rounds, $s_e = |I_e|$, and let M_e, R_e, ε_e, $\widehat{x}_e$, and $\mathcal{R}_e$ be the frozen objects in Section B. By the exact-zero branch of that state machine, $\varepsilon_e > 0$. Put

$$P = I - \frac{\mathbf{1}\mathbf{1}^\top}{d}, \qquad H_e = \max_j \|PM_{e,:,j}\|_2, \qquad D_e^{\mathrm{norm}} = \max_j \frac{\|R_{e,:,j}\|_2}{\varepsilon_e}.$$

Let $r(e)$ index the cap segment containing epoch e and write

$$\lambda_{r(e)} = \frac{1}{\sqrt{d}C_{r(e)}}.$$

The center and FTRL proposals lie on a segment of length at most $\mathcal{R}_e\varepsilon_e$. Define

$$\mathcal{G}_e^{\mathrm{M},2} = \frac{\lambda_{r(e)}\mathcal{R}_e^2 H_e^2}{8} + \mathcal{R}_e D_e^{\mathrm{norm}} \tag{58}$$

and let $\mathcal{G}_e^{(2)}$ be the combined master–FTRL curvature appearing in the certificate.

Lemma 12 (Cap-segment comparison)**.** *On the lifelong confidence event, for either expert label* $a \in \{1,2\}$ *and every prefix of cap segment* r*,*

$$\sum_{e:r(e)=r}\sum_{t\in I_e}(c_{t,a} - h_t)^\top A_{:,j_t} \le \frac{\log(1/\pi_{r,a})}{\lambda_r} + \sum_{e:r(e)=r}\mathcal{G}_e^{\mathrm{M},2}s_e\varepsilon_e^2,$$

where the final epoch may be truncated by the prefix.

Proof. For $t \in I_e$, set $u_{t,a} = c_{t,a}^\top M_{e,:,j_t}$ and $\pi_{t,a} = w_{t,a}/(w_{t,1} + w_{t,2})$. Exponential weights over the concatenated local rounds of one cap segment gives

$$\sum_t(u_{t,a} - \pi_t^\top u_t) \le \frac{\log(1/\pi_{r,a})}{\lambda_r} + \frac{\lambda_r}{8}\sum_t \mathrm{range}(u_t)^2.$$

The advice may change predictably at epoch boundaries. Since the center is one endpoint of the center–FTRL segment,

$$\mathrm{range}(u_t) \le \mathcal{R}_e\varepsilon_e H_e.$$

Moreover every expert and the mixture are within $\mathcal{R}_e\varepsilon_e$ of one another. Confidence coverage and Cauchy–Schwarz therefore give

$$|(c_{t,a} - h_t)^\top(A - M_e)_{:,j_t}| \le \mathcal{R}_e D_e^{\mathrm{norm}}\varepsilon_e^2.$$

Adding the two bounds over the segment proves the claim. Global-mode gaps contain no master update and do not alter the potential. □

Lemma 13 (Total cap-restart cost)**.** *If every accepted multiplier is at most* $\mathcal{R}_{\max}$*, then the caps satisfy*

$$C_{r+1} \ge 2C_r, \qquad \sum_r C_r < 4\max\{1, \mathcal{R}_{\max}\}.$$

For the horizon-clipped prior scale $\widetilde{\varepsilon}_{\sigma_r}$ in the state machine,

$$\sum_r \frac{1}{\lambda_r} \log \frac{1+\widetilde{\varepsilon}_{\sigma_r}}{\widetilde{\varepsilon}_{\sigma_r}} < 4\sqrt{d}\max\{1, \mathcal{R}_{\max}\}\Lambda_T.$$

Proof. Every restart moves to a strictly larger power of two, hence at least doubles the cap. Dyadic rounding gives $C_{\max} < 2\max\{1, \mathcal{R}_{\max}\}$, and the geometric series is less than $2C_{\max}$. Moreover,

$$\widetilde{\varepsilon}_{\sigma_r} \geq \frac{1}{4(T \vee 2)} \quad \Longrightarrow \quad \log \frac{1+\widetilde{\varepsilon}_{\sigma_r}}{\widetilde{\varepsilon}_{\sigma_r}} \leq \log\{1 + 4(T \vee 2)\} \leq \Lambda_T.$$

Substitute $\lambda_r^{-1} = \sqrt{d}C_r$ and sum over the caps. □

For epoch e, let j_e be a least frequent realized column, let N_e be its count, and let the observable interpolation lemma supply α_e, deletion surplus δ_e, row floor ρ_e, and threshold

$$\beta_e = \frac{\underline{\Delta}_e}{8(1+\underline{\Delta}_e)}.$$

The two-expert certificate accepts only if its reserved surplus dominates the combined curvature. This is the same certificate condition used by the rare-column step below, evaluated with the candidate cap rate.

Theorem 7 (Persistent two-expert local regret)**.** *On the confidence and local row-sampling events, every prefix spanning accepted local epochs satisfies the following bound. If the state machine reaches its terminal certified epoch, denote that epoch by $e_\dagger$; otherwise omit every term containing $e_\dagger$:*

$$\begin{aligned}\mathrm{Reg}_{\mathrm{anchor}} \leq \sum_r \frac{1}{\lambda_r} \log \frac{1+\widetilde{\varepsilon}_{\sigma_r}}{\widetilde{\varepsilon}_{\sigma_r}} \\ + \sum_e \frac{\mathcal{R}_e^2}{2} + \sum_{e \neq e_\dagger} \frac{166\mathcal{G}_e^{(2)}\mathcal{L}}{\rho_e \beta_e} \\ + \mathbf{1}\{e_\dagger \textit{ exists}\} 2\mathcal{G}_{e_\dagger}^{(2)} T \varepsilon_{e_\dagger}^2. \end{aligned} \tag{59}$$

Moreover, a terminal certified epoch satisfies

$$\varepsilon_{e_\dagger} \leq 2^{-(K_T+1)} \leq \frac{1}{4(T \vee 2)}. \tag{60}$$

Proof. Apply Lemma 12 to the FTRL label in each cap segment. The master prior is then paid once per segment, while Lemma 9 may use a different comparator in each epoch because it holds uniformly over the certified local set.

Fix a nonterminal epoch e, choose a least frequent realized column j_e, and let N_e be its count. If $N_e \leq \beta_e s_e$, use the restricted comparator from Lemma 7. Combining the master and FTRL inequalities gives

$$\frac{\mathcal{R}_e^2}{2} + 2\mathcal{G}_e^{(2)} s_e \varepsilon_e^2 - \alpha_e \delta_e s_e + \alpha_e(1+\delta_e)N_e,$$

apart from the already booked segment prior. The certificate's surplus-reserve test makes the last three terms nonpositive.

If instead $N_e > \beta_e s_e$, every column is frequent. Use $x^\star$ as the FTRL comparator. Certificate soundness places $x^\star$ in the local set, and full mixing gives $(x^\star)^\top A_{:j} = v(A)$ for every column. The comparator contribution is therefore zero, and the epoch regret is at most

$$\frac{\mathcal{R}_e^2}{2} + 2\mathcal{G}_e^{(2)} s_e \varepsilon_e^2.$$

Because this epoch is nonterminal, it refreshes as soon as its confidence radius halves. Lemma 10 and the least-frequency condition give

$$s_e \leq \frac{83\mathcal{L}}{\rho_e \beta_e \varepsilon_e^2}.$$

Substitution gives the corresponding nonterminal term in (59). The same implication holds when the horizon truncates a nonterminal epoch: if its length exceeded the display, every column would already have forced a refresh.

It remains to handle $e_\dagger$. Use $x^\star$ again and bound its length by T, which gives the last term in (59) without invoking the refresh lemma. At the first accepted epoch every confidence half-width is at most $1/2$. Every later accepted epoch begins only after the previous accepted radius has halved; a rejected recertification can only delay the next accepted epoch, while nested confidence intervals preserve the achieved radius. The (K_T+1)st accepted epoch consequently satisfies $\varepsilon_{e_\dagger} \leq 2^{-(K_T+1)}$. Since $K_T = 1 + \lceil \log_2(T \vee 2) \rceil$, this is at most $1/[4(T \vee 2)]$, proving (60). Summing the epoch and segment bounds completes the proof. □

I A LIFELONG VARIANCE-ADAPTIVE CONFIDENCE RECTANGLE

This section constructs the single confidence event used by every stage of IC-RF. It does not assume independence across rounds or across payoff entries.

Observation model. Let $A \in [0,1]^{d\times d}$ be fixed. On round t, all variables determining the selected entry (i_t, j_t) are revealed before the payoff Y_t. Write $\mathcal{P}_t$ for the resulting pre-payoff sigma-field. We assume

$$0 \le Y_t \le 1, \qquad \mathbb{E}[Y_t \mid \mathcal{P}_t] = A_{i_t j_t}. \tag{61}$$

The selection rule may otherwise depend arbitrarily on the complete past.

Fix an entry (i,j) and index its observed rewards by intrinsic visit time: X_s^{ij} is the reward on the sth round for which $(i_t, j_t) = (i,j)$. Define

$$\widetilde{\mu}_s^{ij} = \frac{\frac{1}{2} + \sum_{q=1}^{s-1} X_q^{ij}}{s}, \qquad V_n^{ij} = \sum_{s=1}^{n} \big(X_s^{ij} - \widetilde{\mu}_s^{ij}\big)^2. \tag{62}$$

The predictor $\widetilde{\mu}_s^{ij}$ is measurable before X_s^{ij} is observed and belongs to $[0,1]$.

An explicit uniform boundary. Fix $\eta = 2$, $m = 1$, and $h(x) = \zeta(2)(x+1)^2$, where ζ is the Riemann zeta function. Let

$$a_1 = \frac{2^{1/4} + 2^{-1/4}}{\sqrt{2}}, \qquad a_2 = \frac{\sqrt{2}+1}{2}, \tag{63}$$

and, for $v \ge 1$ and $\alpha \in (0,1)$, define

$$q_\alpha(v) = \log\frac{\zeta(2)}{\alpha} + 2\log\big(1 + \log_2 v\big), \tag{64}$$

$$u_\alpha(v) = \sqrt{a_1^2 v q_\alpha(v) + a_2^2 q_\alpha(v)^2} + a_2 q_\alpha(v). \tag{65}$$

Theorem 4 of Howard et al. (2021) applies to every $[0,1]$-valued adapted sequence, every $[0,1]$-valued predictable predictor, and any sub-exponential uniform boundary with scale $b - a = 1$. We instantiate its boundary with the stitched construction in their Theorem 1, taking $\eta = 2$, $m = 1$, and $h(k) = \zeta(2)(k+1)^2$. Substitution into that theorem gives exactly the constants a_1, a_2 in (63) and shows that $v \mapsto u_\alpha(v \vee 1)$ is a valid boundary for the self-normalized process in their Theorem 4. The looser inequality

$$u_\alpha(v) \le a_1\sqrt{v q_\alpha(v)} + 2a_2 q_\alpha(v) \tag{66}$$

will be useful for deterministic sample envelopes.

We allocate probability $\delta/2$ to the confidence rectangle and set $\alpha_\star = \delta/(4d^2)$. After $n \ge 1$ observations of entry (i,j), let

$$\widehat{A}_n^{ij} = \frac{1}{n}\sum_{s=1}^{n} X_s^{ij}, \qquad b_n^{ij} = \frac{u_{\alpha_\star}(V_n^{ij} \vee 1)}{n}. \tag{67}$$

The raw clipped interval is

$$I_n^{ij} = \big[(\widehat{A}_n^{ij} - b_n^{ij}) \vee 0, (\widehat{A}_n^{ij} + b_n^{ij}) \wedge 1\big]. \tag{68}$$

For $n = 0$, set $I_0^{ij} = [0,1]$. The algorithm stores the running intersection

$$[L_n^{ij}, U_n^{ij}] = \bigcap_{s=0}^{n} I_s^{ij}, \qquad M_n^{ij} = \frac{L_n^{ij} + U_n^{ij}}{2}, \qquad r_n^{ij} = \frac{U_n^{ij} - L_n^{ij}}{2}. \tag{69}$$

On a path where this intersection is empty, IC-RF plays the uniform row distribution for the remaining horizon. Theorem 8 shows that this convention is never invoked on the joint confidence event.

Lemma 14 (Optional skipping preserves the entry process)**.** *For each entry (i,j), its intrinsic-time observations, up to every realized visit, satisfy*

$$\mathbb{E}[X_s^{ij} - A_{ij} \mid \mathcal{H}_s^{ij}] = 0, \tag{70}$$

where $\mathcal{H}_s^{ij} = \mathcal{P}_{\tau_s^{ij}}$ is the pre-payoff sigma-field at the sth visit. Moreover, every nonnegative exponential supermartingale used by the confidence-sequence theorem remains a supermartingale when inspected only at these visits. Consequently its uniform crossing bound holds over all realized intrinsic visit counts, including paths on which the entry is visited only finitely many times.

Proof. Let τ_s^{ij} be the calendar time of the sth visit to entry (i,j). The selection indicator is $\mathcal{P}_t$-measurable because the entry is chosen before the current payoff. Hence, on $\{\tau_s^{ij} < \infty\}$, equation (61) evaluated at the predictable visit time gives

$$\mathbb{E}[Y_{\tau_s^{ij}} - A_{ij} \mid \mathcal{P}_{\tau_s^{ij}}] = 0.$$

This is precisely (70).

For completeness, consider a nonnegative exponential supermartingale used to prove the empirical Bernstein boundary cited above. Index it by intrinsic observations. Between two visits leave its value unchanged, and at a visit apply its ordinary one-observation update. The conditional supermartingale inequality at a visit follows from the preceding display; off a visit it holds with equality. Thus the calendar-time process is a nonnegative supermartingale. Ville's inequality controls its crossing probability over all calendar times, and therefore over the possibly finite subsequence of realized visits. The stitched mixture has the same property by nonnegative integration. No fictitious sample after a non-stopping "last visit" is needed. □

Theorem 8 (Lifelong entrywise confidence rectangle)**.** *Under* (61)*, with probability at least* $1-\delta/2$*, simultaneously for every entry* (i,j) *and every intrinsic sample size* $n \geq 0$*,*

$$A_{ij} \in [L_n^{ij}, U_n^{ij}]. \tag{71}$$

Consequently,

$$|M_n^{ij} - A_{ij}| \leq r_n^{ij} \leq b_n^{ij}, \tag{72}$$

and r_n^{ij} *is nonincreasing in* n*.*

Proof. Fix (i,j). Lemma 14 gives a $[0,1]$-valued adapted sequence with constant conditional mean A_{ij}, and $\widetilde{\mu}_s^{ij}$ in (62) is a $[0,1]$-valued predictable sequence. Theorem 4 of Howard et al. (2021), instantiated with the stitched boundary from their Theorem 1 exactly as described above, gives

$$\mathbb{P}\Big(\forall n \geq 1 : |\widehat{A}_n^{ij} - A_{ij}| < b_n^{ij}\Big) \geq 1 - 2\alpha_\star = 1 - \frac{\delta}{2d^2}.$$

Clipping to $[0,1]$ preserves coverage. A union bound over the d^2 entries gives a common event of probability at least $1-\delta/2$. On that event every raw interval contains A_{ij}, hence so does every running intersection. The midpoint inequality follows from $A_{ij} \in [L_n^{ij}, U_n^{ij}]$. Since the intersection includes the current raw interval, its half-width is no larger than b_n^{ij}; nestedness makes the half-width nonincreasing. □

Worst-variance sample envelope. The confidence sequence is horizon-free. For analysis over a horizon T, define

$$\mathcal{L}_T(\alpha) = \log\frac{\zeta(2)}{\alpha} + 2\log\big(1 + \log_2(T \vee 1)\big) \tag{73}$$

and

$$\overline{N}_T(\varepsilon,\alpha) = \left\lceil \max\left\{ \frac{4a_1^2\mathcal{L}_T(\alpha)}{\varepsilon^2}, \frac{4a_2\mathcal{L}_T(\alpha)}{\varepsilon} \right\} \right\rceil. \tag{74}$$

Corollary 2 (Worst-variance crossing time)**.** *For every observation path, every* $n \leq T$*, and every entry,*

$$n \geq \overline{N}_T(\varepsilon,\alpha_\star) \quad \Longrightarrow \quad r_n^{ij} \leq \varepsilon. \tag{75}$$

For $\varepsilon \leq 1$*,* $\overline{N}_T(\varepsilon,\alpha_\star) = O(\mathcal{L}_T(\alpha_\star)/\varepsilon^2)$*.*

Proof. Both X_s^{ij} and $\widetilde{\mu}_s^{ij}$ lie in $[0,1]$, so $V_n^{ij} \leq n$. Equations (66) and (73) yield

$$b_n^{ij} \leq a_1\sqrt{\frac{\mathcal{L}_T(\alpha_\star)}{n}} + \frac{2a_2\mathcal{L}_T(\alpha_\star)}{n}.$$

The two requirements in (74) make the first and second terms at most $\varepsilon/2$. The conclusion follows from $r_n^{ij} \leq b_n^{ij}$. □

Low-variance acceleration. For constants $\nu \in [0,1]$ and $c_0 \geq 0$, define

$$\overline{N}_T^{\mathrm{var}}(\varepsilon,\alpha;\nu,c_0) = \left\lceil \max\left\{ \frac{9a_1^2\nu\mathcal{L}_T(\alpha)}{\varepsilon^2}, \frac{3a_1\sqrt{(c_0 \vee 1)\mathcal{L}_T(\alpha)}}{\varepsilon}, \frac{6a_2\mathcal{L}_T(\alpha)}{\varepsilon} \right\} \right\rceil. \tag{76}$$

Corollary 3 (Pathwise variance-adaptive crossing time)**.** *Suppose an entry satisfies* $V_n^{ij} \leq \nu n + c_0$ *at some* $n \leq T$*. If* $n \geq \overline{N}_T^{\mathrm{var}}(\varepsilon,\alpha_\star;\nu,c_0)$*, then* $r_n^{ij} \leq \varepsilon$*. For a deterministic entry,* $X_s^{ij} = A_{ij}$ *for every visit and*

$$V_n^{ij} = (A_{ij} - \tfrac{1}{2})^2 \sum_{s=1}^{n} \frac{1}{s^2} \leq \frac{\pi^2}{24} < 1. \tag{77}$$

Hence its radius crosses ε *after* $O(\log(d/\delta)/\varepsilon)$ *intrinsic observations, with no dependence on* T *required for this deterministic specialization.*

Proof. Using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ in (66) gives

$$b_n^{ij} \le a_1 \sqrt{\frac{\nu \mathcal{L}_T(\alpha_\star)}{n}} + \frac{a_1 \sqrt{(c_0 \vee 1)\mathcal{L}_T(\alpha_\star)}}{n} + \frac{2a_2 \mathcal{L}_T(\alpha_\star)}{n}.$$

The three requirements in (76) make these terms at most $\varepsilon/3$. For deterministic observations, $A_{ij} - \widetilde{\mu}_s^{ij} = (A_{ij} - 1/2)/s$, which proves (77). Substitution into (67) gives the stated $1/\varepsilon$ sample order. □

J Proof of Proposition 1 and a regularity neighborhood

Proof of Proposition 1. Fix a column j. If deleting j did not change the value, an optimal column distribution for the reduced game, extended by zero mass on j, would still hold every row's payoff to at most $v(A)$. It would therefore be an optimal column distribution for the original game, different from $y^\star$ because $y_j^\star > 0$. This contradicts uniqueness. There are finitely many columns, so $\Delta_{\mathrm{del}} > 0$.

Suppose $\mathsf{B}_x(A)$ were singular. Its null space would contain a nonzero pair (h, c) satisfying $A^\top h = c\mathbf{1}$ and $\mathbf{1}^\top h = 0$. The pair must have $h \neq 0$. Since $x^\star$ has full support, both $x^\star + \eta h$ and $x^\star - \eta h$ belong to Δ_d for small enough $\eta > 0$. Every column gives these two strategies payoffs $v(A) + \eta c$ and $v(A) - \eta c$, respectively. Optimality of $v(A)$ forces $c = 0$; otherwise one strategy would guarantee more than $v(A)$. With $c = 0$, both strategies are optimal, contradicting uniqueness. Hence $\mathsf{B}_x(A)$ is nonsingular. The argument for $\mathsf{B}_y(A)$ exchanges rows and columns. □

Proposition 9 (A quantitative regularity neighborhood)**.** *Let A satisfy Assumption 1, and define*

$$r_A = \min\left\{ \frac{1}{2\kappa d}, \frac{\rho}{4\kappa}, \frac{\tau}{4\kappa}, \frac{\Delta_{\mathrm{del}}}{4} \right\}.$$

If $A' = A + E \in [0,1]^{d\times d}$ and $\|E\|_{\max} < r_A$, where $\|E\|_{\max} = \max_{i,j} |E_{ij}|$, then A' also has a unique fully mixed equilibrium. Its difficulty parameters satisfy

$$\rho(A') \ge \frac{\rho}{2}, \qquad \tau(A') \ge \frac{\tau}{2}, \qquad \Delta_{\mathrm{del}}(A') \ge \frac{\Delta_{\mathrm{del}}}{2}, \qquad \kappa(A') \le 2\kappa.$$

Proof. Put $\epsilon = \|E\|_{\max}$ and, for $q \in \{x, y\}$, write $K_q = \mathsf{B}_q(A)$ and $K_q' = \mathsf{B}_q(A')$. The perturbation $D_q = K_q' - K_q$ has a zero last row and column and satisfies $\|D_q\|_\infty \le d\epsilon$. Since $\|K_q^{-1} D_q\|_\infty \le \kappa d\epsilon < 1/2$, the Neumann-series bound gives

$$\|(K_q')^{-1}\|_\infty \le \frac{\|K_q^{-1}\|_\infty}{1 - \|K_q^{-1} D_q\|_\infty} \le 2\kappa. \tag{78}$$

Thus both perturbed augmented systems are invertible and $\kappa(A') \le 2\kappa$.

Let $b = (\mathbf{0}^\top, 1)^\top$, and write $z_x = (x^{\star\top}, v(A))^\top = K_x^{-1} b$ and $z_x' = (x'^\top, v_x')^\top = (K_x')^{-1} b$. The resolvent identity yields

$$z_x' - z_x = -(K_x')^{-1} D_x z_x.$$

The first d coordinates of $D_x z_x$ equal $E^\top x^\star$, while its last coordinate is zero. Because $x^\star$ is a probability vector, $\|D_x z_x\|_\infty \le \epsilon$. Equation (78) therefore implies

$$\|z_x' - z_x\|_\infty \le 2\kappa\epsilon.$$

The same argument applied to K_y gives $\|z_y' - z_y\|_\infty \le 2\kappa\epsilon$, where the first d coordinates of z_y and z_y' are $y^\star$ and y', respectively. The assumptions on ϵ now give

$$\min_i x_i' \ge \rho - 2\kappa\epsilon > \frac{\rho}{2}, \qquad \min_j y_j' \ge \tau - 2\kappa\epsilon > \frac{\tau}{2}.$$

The augmented equations state that every column has payoff v_x' against x' and every row has payoff v_y' against y'. Hence $x'^\top A' y' = v_x' = v_y'$, so (x', y') is a fully mixed saddle point of A'. It is unique. Indeed, any other optimal row strategy must make every column constraint tight because y' has strictly positive coordinates; it would therefore solve the invertible system $K_x' z = b$. The column side is identical.

Finally, the value of any fixed restricted matrix game is one-Lipschitz in the entrywise maximum norm. For each column j,

$$\begin{aligned} v_{[d]\setminus\{j\}}(A') - v(A') &\ge v_{[d]\setminus\{j\}}(A) - \epsilon - v(A) - \epsilon \\ &\ge \Delta_{\mathrm{del}} - 2\epsilon > \frac{\Delta_{\mathrm{del}}}{2}. \end{aligned}$$

Taking the minimum over j proves the asserted deletion-gap bound and completes the proof. □

K Proof of Theorem 2

Proof. Fix $d \geq 2$ and $T \geq 1$. Write $\mathbf{1} \in \mathbb{R}^d$ for the all-ones vector, $u = \mathbf{1}/d$ for the uniform distribution, and $e_i \in \mathbb{R}^d$ for the ith coordinate vector. Let

$$\alpha = \frac{1}{16}, \qquad \varepsilon = \frac{1}{16} \min\left\{1, \sqrt{\frac{d}{T}}\right\}.$$

For each $h \in [d]$, set

$$x_h^{(h)} = 1 - \alpha, \qquad x_i^{(h)} = \frac{\alpha}{d-1} \quad (i \neq h).$$

Choose any permutation matrix Π_h satisfying $\Pi_h e_1 = e_h$, and define

$$v_h = \frac{1}{2} + (1-\alpha)\varepsilon, \qquad A_h = v_h \mathbf{1}\mathbf{1}^\top + \varepsilon\big(\Pi_h - \mathbf{1}(x^{(h)})^\top \Pi_h\big). \tag{79}$$

If π_h is the permutation associated with Π_h, then

$$(A_h)_{ij} = v_h + \varepsilon\big(\mathbf{1}\{i = \pi_h(j)\} - x^{(h)}_{\pi_h(j)}\big).$$

The term in parentheses belongs to $[-1, 1]$. Thus every entry lies between $1/2 - \alpha\varepsilon > 0$ and $1/2 + (2-\alpha)\varepsilon < 1$, so $A_h \in [0,1]^{d \times d}$.

We first verify the game-theoretic properties of the construction. Since $\Pi_h u = u$ and $(x^{(h)})^\top \Pi_h u = 1/d$,

$$A_h u = v_h \mathbf{1}, \qquad (x^{(h)})^\top A_h = v_h \mathbf{1}^\top.$$

Thus $(x^{(h)}, u)$ is a fully mixed saddle point with value v_h. It is unique. Indeed, if x is any optimal row strategy, then $x^\top A_h u = v_h$ and $x^\top A_h \geq v_h \mathbf{1}^\top$. Because u has full support, all these column inequalities must be equalities. Equation (79) then implies $(x - x^{(h)})^\top \Pi_h = 0$, so $x = x^{(h)}$. Similarly, if y is an optimal column strategy, the full support of $x^{(h)}$ forces $A_h y = v_h \mathbf{1}$. Therefore $\Pi_h y = \mathbf{1}(x^{(h)})^\top \Pi_h y$, so $\Pi_h y$ is constant. As y sums to one, $y = u$. This proves uniqueness.

The opponent now selects column 1 on every round. Conditional on the sampled row, draw independent Bernoulli payoffs with the corresponding mean in A_h. Since $\Pi_h e_1 = e_h$ and $x_h^{(h)} = 1 - \alpha$, the revealed column is

$$A_h e_1 = \frac{1}{2}\mathbf{1} + \varepsilon e_h. \tag{80}$$

Thus the learner observes exactly a d-armed Bernoulli bandit with one unknown superior row h, whose mean exceeds the others by ε. Its regret relative to the mixed game value retains a signed correction: if N_h is the number of times row h is sampled,

$$\mathbb{E}_h[\mathrm{Reg}_T] = \varepsilon\big((1-\alpha)T - \mathbb{E}_h N_h\big). \tag{81}$$

Here $\mathbb{E}_h$ refers to the interaction law under A_h.

For completeness, we bound how often a learner can select the unknown row. Let P_0 be the auxiliary interaction law in which every row in the revealed column has Bernoulli mean $1/2$, and let P_h be the law under A_h. The learner uses the same causal decision kernel in both laws. The chain rule for relative entropy therefore yields

$$\mathrm{KL}(P_0 \| P_h) = \mathbb{E}_0 N_h \ \mathrm{kl}\left(\frac{1}{2}, \frac{1}{2} + \varepsilon\right) \leq 4\varepsilon^2 \mathbb{E}_0 N_h.$$

The last inequality holds because $\varepsilon \leq 1/16$ and $\mathrm{kl}(1/2, 1/2 + \varepsilon) = -\frac{1}{2}\log(1 - 4\varepsilon^2) \leq 4\varepsilon^2$. Pinsker's inequality and $0 \leq N_h/T \leq 1$ imply

$$\mathbb{E}_h N_h \leq \mathbb{E}_0 N_h + T\sqrt{\tfrac{1}{2}\,\mathrm{KL}(P_0 \| P_h)} \leq \mathbb{E}_0 N_h + \sqrt{2}\,\varepsilon T \sqrt{\mathbb{E}_0 N_h}.$$

Average over $h \in [d]$. Since $\sum_h \mathbb{E}_0 N_h = T$, Jensen's inequality gives

$$\frac{1}{d}\sum_{h=1}^{d} \mathbb{E}_h N_h \leq \frac{T}{d} + \sqrt{2}\,\varepsilon T \sqrt{\frac{T}{d}} \leq \frac{T}{d} + \frac{\sqrt{2}}{16} T. \tag{82}$$

Combining (81) and (82), and using $d \geq 2$, yields

$$\frac{1}{d}\sum_{h=1}^{d} \mathbb{E}_h[\mathrm{Reg}_T] \geq \varepsilon T \left(1 - \frac{1}{16} - \frac{1}{d} - \frac{\sqrt{2}}{16}\right) \geq \frac{\varepsilon T}{4} = \frac{1}{64}\min\{T, \sqrt{dT}\}.$$

At least one h attains this average. Taking $A_{d,T} = A_h$ proves the theorem. □

L BIT COMPLEXITY OF THE GLOBAL RESTRICTED-GAME ORACLE

The statistical theorem uses exact real arithmetic. For a rational proxy matrix, the global controller's restricted maximin policy and value have polynomial-bit exact implementations; the true matrix is never an input.

For a rational number, let its encoding length be the total binary length of its signed numerator and positive denominator in lowest terms. Let $H \in \mathbb{Q}^{d\times m}$ have entry encoding length at most B, and let $S \subseteq [m]$ be nonempty.

Theorem 9 (Exact restricted-game oracle)**.** *There is a deterministic Turing algorithm that returns*

1. *the exact value $v_S(H)$;*
2. *an exact rational maximin policy $x_S(H) \in \Delta_d$; and*
3. *the lexicographically least maximin policy, if a unique output is required,*

using a number of bit operations polynomial in d, m, and B. Every returned numerator and denominator has polynomial encoding length. Computing all m one-column deletion values remains polynomial in the same parameters.

Proof. The restricted row problem is the rational linear program

$$\begin{aligned} \text{maximize} \quad & z \\ \text{subject to} \quad & \sum_{i=1}^{d} x_i H_{ij} \geq z \quad (j \in S), \\ & \sum_{i=1}^{d} x_i = 1, \\ & x_i \geq 0 \qquad\qquad (i \in [d]). \end{aligned} \tag{83}$$

It has $d+1$ variables and at most $m+d+1$ constraints. Rational linear programming is solvable in time polynomial in the number of variables, constraints, and coefficient encoding length (Khachiyan, 1979). This supplies exact feasibility, optimal-value, and optimal-solution routines in the Turing model.

For completeness, the output-size requirement follows directly from the geometry of (83). An optimum exists because the simplex is compact, and an optimal basic feasible solution can be chosen. After clearing denominators, a basis is a square system of order at most $d+1$ whose integer coefficients have bit length polynomial in d, m, and B. Cramer's rule expresses every basic coordinate as a ratio of two basis determinants. Hadamard's inequality bounds the binary length of each determinant by

$$O\big(d^2 B + d\log(d+1)\big)$$

when each basis row clears at most $d+2$ input denominators. Thus an exact optimizer and its value have polynomial encoding length; the exact rational LP algorithm never needs a super-polynomial output.

To impose deterministic lexicographic tie breaking, first solve (83) and add the equality $z = v_S(H)$. Minimize x_1, fix it at its optimum, then minimize x_2, and continue. This performs at most $d+1$ rational linear programs. Every newly fixed coordinate has the polynomial bit bound just proved, so the later LP inputs also retain polynomial length.

For a deletion value, replace S by $S \setminus \{j\}$. At most m additional calls preserve polynomial complexity. Active-face membership uses rational comparisons, and each other global game computation is an LP of this form. This proves all claims, including polynomial total cost over T rounds when the proxy inputs have polynomial encoding length. □

The controller therefore never maximizes game value over a confidence set. It solves an ordinary rational zero-sum game at an already constructed midpoint or optimistic proxy. The next two sections provide outward rational evaluation of the confidence boundary, a rational certificate, controlled local projections, and the remaining local updates.

M A POLYNOMIAL-BIT IMPLEMENTATION

For the finite-input model, T, d, and rational δ are encoded in binary, and each observed payoff is a dyadic rational with at most b fractional bits. The unknown mean matrix is not supplied. Complexity is polynomial in T, d, b, and the encoding length of δ.

M.1 RATIONAL CONFIDENCE RECTANGLES

For $\alpha \in (0,1)$, define the integer

$$L_T^{\mathbb{Q}} = 3 + \left\lceil \log_2 \frac{1}{\alpha} \right\rceil + 2\left\lceil \log_2\{1 + \lceil \log_2(T \vee 2)\rceil\}\right\rceil. \tag{84}$$

For a nonnegative rational q, let $\mathrm{usqrt}(q)$ be zero when $q = 0$ and otherwise be a positive dyadic rational satisfying

$$\sqrt{q} \le \mathrm{usqrt}(q) \le 2\sqrt{q}. \tag{85}$$

Such a number is obtained by locating the binary exponent of q and then performing integer bisection; both its encoding length and computation time are polynomial in the encoding length of q.

Replace the stitched boundary in (65) by

$$u_{\mathbb{Q}}(v) = \mathrm{usqrt}\left(\frac{9}{4}\{vL_T^{\mathbb{Q}} + (L_T^{\mathbb{Q}})^2\}\right) + \frac{3}{2}L_T^{\mathbb{Q}}. \tag{86}$$

Lemma 15 (Rational boundary domination)**.** *For every $1 \le v \le T$, $u_{\mathbb{Q}}(v)$ is a rational upper bound on the stitched boundary used in Theorem 8. Hence replacing every raw interval by the interval formed with $u_{\mathbb{Q}}(V_n \vee 1)/n$ preserves the same lifelong coverage event. Moreover,*

$$\frac{u_{\mathbb{Q}}(V_n \vee 1)}{n} \le 3\sqrt{\frac{L_T^{\mathbb{Q}}}{n}} + \frac{9L_T^{\mathbb{Q}}}{2n}. \tag{87}$$

Proof. The inequalities $\zeta(2) < 2$ and $\log x \le \log_2 x$ for $x \ge 1$ show that $L_T^{\mathbb{Q}}$ upper-bounds $q_\alpha(v)$ in (65). The constants in (63) satisfy $a_1^2 < 9/4$ and $a_2 < 3/2$. Monotonicity of the square root and (85) therefore prove boundary domination. The probability statement follows because enlarging a confidence sequence cannot destroy coverage.

For the deterministic bound, use the upper half of (85), $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$, and $V_n \vee 1 \le n$ for $n \ge 1$:

$$u_{\mathbb{Q}}(V_n \vee 1) \le 3\sqrt{(V_n \vee 1)L_T^{\mathbb{Q}}} + \frac{9}{2}L_T^{\mathbb{Q}} \le 3\sqrt{nL_T^{\mathbb{Q}}} + \frac{9}{2}L_T^{\mathbb{Q}}.$$

Divide by n. □

Thus

$$N_{\mathbb{Q}}(\varepsilon) = \left\lceil \max\left\{\frac{36L_T^{\mathbb{Q}}}{\varepsilon^2}, \frac{9L_T^{\mathbb{Q}}}{\varepsilon}\right\}\right\rceil \tag{88}$$

is a valid rational sample-count envelope. Replacing N_k by this quantity, W_k by $640d^2L_T^{\mathbb{Q}}/\varepsilon_k^3$, and the local refresh count by $N_{\mathbb{Q}}(\varepsilon/2)$ makes every count inequality in Sections D and G hold with slack. Only universal numerical constants change.

M.2 EXACT RATIONAL CERTIFICATE TESTS

The inverse and solution calculations in (25)–(26) use fraction-free Gaussian elimination. Their outputs have polynomial encoding length by Cramer's rule and Hadamard's determinant bound. The apparent spectral-radius oracle can also be removed.

Lemma 16 (Rational spectral-radius test)**.** *For a nonnegative rational square matrix C,*

$$\rho_{\mathrm{sp}}(C) < 1 \quad \Longleftrightarrow \quad \textit{there exists } w \ge 0 \textit{ with } (I - C)w \ge \mathbf{1}. \tag{89}$$

Consequently the test $\rho_{\mathrm{sp}}(C) < 1$ is decidable in polynomial bit complexity by rational linear programming.

Proof. If $\rho_{\mathrm{sp}}(C) < 1$, then $w = (I - C)^{-1}\mathbf{1} = \sum_{s \ge 0} C^s\mathbf{1}$ is nonnegative and satisfies the displayed inequalities with equality. Conversely, feasibility gives $Cw \le w - \mathbf{1} < w$. In particular $w > 0$, and

$$\max_i \frac{(Cw)_i}{w_i} < 1.$$

The Collatz–Wielandt upper bound for a nonnegative matrix yields $\rho_{\mathrm{sp}}(C) < 1$. The feasibility problem in (89) is a rational LP with polynomial input size, so Theorem 9's rational-LP argument applies. □

It remains to rationalize both the local radius and the interpolation coefficients. Work only at a positive radius ε and after the row-positivity test; then $\widehat{x}$ and every restricted-game policy r_j are probability vectors. Apply (85) to squared rational norms to obtain

$$k_2 \le \widetilde{k}_2 \le 2k_2, \qquad D \le \widetilde{D} \le 2D, \qquad d_j \le \overline{d}_j := \min\left\{\frac{3}{2}, \operatorname{usqrt}(\|r_j - \widehat{x}\|_2^2)\right\} \le \frac{3}{2}, \tag{90}$$

where $\widetilde{D}$ is the maximum upper-approximated column norm divided by ε. The cap at $3/2$ remains an upper bound on d_j because two probability vectors are at distance at most $\sqrt{2} < 3/2$. Set $\underline{k}_2 = \widetilde{k}_2/2$, so $\underline{k}_2 \le k_2$. For any rational bonus $b \ge 3$, define

$$\widetilde{\mathcal{R}}(b) = \widetilde{k}_2 + b, \qquad \widetilde{\alpha}_j(b) = \begin{cases} 1, & \overline{d}_j \le \underline{k}_2\varepsilon, \\ \min\left\{1, \dfrac{b\varepsilon}{\overline{d}_j - \underline{k}_2\varepsilon}\right\}, & \overline{d}_j > \underline{k}_2\varepsilon. \end{cases} \tag{91}$$

Every operation in (91) is rational. In the second branch, $d_j - k_2\varepsilon \le \overline{d}_j - \underline{k}_2\varepsilon$; in the first branch, $d_j \le k_2\varepsilon$. Thus, with $\widetilde{z}_j = (1-\widetilde{\alpha}_j)x^\star + \widetilde{\alpha}_j r_j$, the triangle inequality gives

$$\|\widetilde{z}_j - \widehat{x}\|_2 \le (k_2 + b)\varepsilon \le \widetilde{\mathcal{R}}(b)\varepsilon. \tag{92}$$

Moreover $\overline{d}_j - \underline{k}_2\varepsilon \le 3/2$, so $b \ge 3$ and $\varepsilon \le 1/2$ imply $\widetilde{\alpha}_j(b) \ge 2\varepsilon$ in both branches. For the fixed rational candidate $b = 3$, $\widetilde{\mathcal{R}}(3) \ge k_2 + 2\sqrt{2}$ and $\widetilde{\mathcal{R}}(3) \le 2(k_2 + 2\sqrt{2})$.

The rational FTRL curvature is

$$\widetilde{\mathcal{G}}_{\mathrm{F}}(b) = \frac{1}{2}\max_j \|PM_{:j}\|_2^2 + 2\widetilde{\mathcal{R}}(b)\widetilde{D} + \frac{1}{4}. \tag{93}$$

The squared projected norms are rational. The last term pays the projection error in (96). The aggregate rational test, including its two-expert curvature, is specified and bounded in Section N. Its row-floor and surplus tests use (91) and (93); hence the comparator geometry and all numerical acceptance tests are conservative. The rational bonus 3 dominates $2\sqrt{2}$, while the one-sided norm bounds make interpolation sound without evaluating d_j exactly.

M.3 Rational local FTRL

At an adaptive local round, exact FTRL is the Euclidean projection of the rational vector

$$z_t = \widehat{x} + \eta \sum_{s<t} M_{:j_s} \tag{94}$$

onto the intersection of the simplex and the rational Euclidean ball $\mathcal{X}$.

Lemma 17 (Polynomial-time feasible projection)**.** *Let $z, c \in \mathbb{Q}^d$, $r \in \mathbb{Q}_+$, and $\mathcal{C} = \{x \in \Delta_d : \|x - c\|_2 \le r\}$ with $c \in \mathcal{C}$. For every rational $\xi > 0$, a rational $\widetilde{x} \in \mathcal{C}$ satisfying*

$$\|\widetilde{x} - \Pi_{\mathcal{C}}(z)\|_2 \le \xi \tag{95}$$

can be computed in time polynomial in the input length, d, and $\log(1/\xi)$.

Proof. If $r = 0$, return c, the only feasible point. Otherwise projection onto the simplex is rational and exact: after sorting the coordinates, it has the form $[\Pi_{\Delta_d}(w)]_i = \max\{w_i - \theta, 0\}$ for a rational threshold θ. Let $q(a) = \Pi_{\Delta_d}\{c + a(z - c)\}$ for $a \in [0, 1]$. If $q(1)$ lies in the ball, it is the desired projection. Otherwise the Karush–Kuhn–Tucker conditions give

$$\Pi_{\mathcal{C}}(z) = q(a_\star), \qquad \|q(a_\star) - c\|_2 = r$$

for some $a_\star \in (0, 1)$. The squared distance $\|q(a) - c\|_2^2$ is continuous and nondecreasing in a: equivalently, increase the multiplier on $\|x - c\|_2^2$ in the projection objective and compare the two optimality inequalities.

Rational bisection therefore finds $a_- \le a_\star$ with

$$a_\star - a_- \le \frac{\xi}{1 + \|z - c\|_2}.$$

Every bisection comparison squares rational quantities, and $q(a_-) \in \mathcal{C}$. Nonexpansiveness of simplex projection gives

$$\|q(a_-) - q(a_\star)\|_2 \le (a_\star - a_-)\|z - c\|_2 \le \xi.$$

Use the rational bound $1 + \|z - c\|_1$ in the bisection stopping test. The number of iterations is $O(1 + \log^+((1 + \|z - c\|_1)/\xi))$, where $\log^+ u = \max\{0, \log_2 u\}$, which is polynomial in the general input encoding length and $\log^+(1/\xi)$. For (94), the more specific bound $\|z - c\|_2 \le T\sqrt{d}$ also holds. □

If $\varepsilon = 0$, the certified local set is the singleton $\{\widehat{x}\}$ and its projection is exact. Otherwise choose the largest dyadic tolerance not exceeding the target,

$$\xi_t = 2^{-\lceil \log_2\{4\lceil\sqrt{d}\rceil/\varepsilon^2\}\rceil} \le \frac{\varepsilon^2}{4\lceil\sqrt{d}\rceil}. \tag{96}$$

and play the feasible approximation from Lemma 17. For the realized adaptive column,

$$|\langle \widetilde{x}_t - x_t, A_{:j_t}\rangle| \le \sqrt{d}\,\xi_t \le \frac{\varepsilon^2}{4}.$$

The column may depend on $\widetilde{x}_t$; Lemma 9 is uniform over realized column sequences, so this adaptivity causes no additional term. The added $\frac{1}{4}$ in (93) pays the approximation error. Feasibility preserves the certified row floor.

Theorem 10 (Polynomial-bit implementation of the IC-RF primitives)**.** *Under the finite-input model above, the confidence rectangle, global restricted-face controller, rational saddle-system and interpolation calculations, and centered-FTRL expert have implementations whose expected running time and memory are polynomial in T, d, b, and the encoding length of δ. Their corresponding coverage, ledger, comparator-geometry, and FTRL bounds change only by universal numerical constants.*

Proof. The predictable means and squared residual sums are exact rationals. Lemma 15 constructs containing rational confidence rectangles; the inflated budgets following (88) preserve the promotion, ledger, and refresh proofs. Theorem 9 implements every global game computation. Lemma 16, fraction-free elimination, and the upper norm approximations implement the saddle-system tests, while (92) preserves restricted-comparator geometry. The full aggregate test and its fine-radius acceptance check are in Section N; a permanently hidden column is handled by the proper-face ledger. Lemma 17 and (96) implement local FTRL and charge its numerical error to $\widetilde{\mathcal{G}}_{\mathrm{F}}$.

We bound stored quantities in dependency order. Payoff sums, predictable means, and residual sums have polynomial encoding length after at most T updates; confidence endpoints depend only on these statistics and the input confidence level. Intersections select existing endpoints. Thus every positive stored radius ε has polynomial encoding length L and is at least 2^{-L}. The tolerance in (96) therefore has polynomial encoding length, and the projection bisection takes polynomially many iterations. At zero radius no bisection is needed. Exact LP outputs and fraction-free linear algebra have polynomial output length by Theorem 9 and Hadamard's bound. Local weights, proposal states, and wealth are bounded separately in Lemma 18. Finally, a rational distribution can be sampled exactly from fair random bits in expected polynomial time. Draw a uniform binary integer over the next power-of-two interval above a common denominator, reject its unused tail, and use cumulative rational masses. The acceptance probability is at least one half. Multiplying polynomial expected per-round work by T proves the claim. □

N End-to-end polynomial-bit implementation

Section M implements the confidence rectangle, saddle system, interpolation, maximin programs, projection, and centered FTRL. We complete the aggregate certificate, persistent update, proposal, and gate.

For the adaptive-radius version, use the rational grid

$$\mathcal{B}_{\mathbb{Q}} = \{3\cdot 2^q : q \in \mathbb{N}_0, (3\cdot 2^q)^2 \le 64d\} \cup \{b^{\mathbb{Q}}_{\mathrm{edge}}\}.$$

For a fixed rational slack $\eta_{\mathrm{edge}} \in (0,1)$ and fixed integer precision $p_{\mathrm{edge}} \ge 1$, let $b^{\mathbb{Q}}_{\mathrm{edge}}$ be the largest multiple of $2^{-p_{\mathrm{edge}}}$ satisfying

$$b \ge 3, \qquad b^2 \le 64d, \qquad b \le (1-\eta_{\mathrm{edge}})\left\{\frac{\rho}{2\varepsilon} - \widetilde{k}_2 - k_\infty\right\}.$$

Omit it if this finite set is empty. These are rational comparisons, and the endpoint strictly satisfies the implemented row-floor test. Here $\widetilde{k}_2$ is the rational upper enclosure used in (91). Every bonus is at least $3 > 2\sqrt{2}$, and the first grid point supplies the fixed soundness candidate. The edge point is optional for the guarantee and may be omitted. For each candidate b, use $\widetilde{\mathcal{R}}(b)$ and $\widetilde{\alpha}_j(b)$ from (91). Let $C^{\mathbb{Q}}(b)$ be the larger of the stored dyadic cap and the least power of two not below $\max\{1, \widetilde{\mathcal{R}}(b)\}$. Choose and retain a positive rational $\widetilde{\lambda}(b)$ satisfying

$$\frac{1}{2\sqrt{d}\,C^{\mathbb{Q}}(b)} \le \widetilde{\lambda}(b) \le \frac{1}{\sqrt{d}\,C^{\mathbb{Q}}(b)}.$$

Rational square-root bisection supplies such a value. With the exact rational quantity $H^2 = \max_j \|PM_{:j}\|_2^2$, define

$$\widetilde{\mathcal{G}}_{\mathrm{M}}(b) = \frac{\widetilde{\lambda}(b)\widetilde{\mathcal{R}}(b)^2 H^2}{8} + \widetilde{\mathcal{R}}(b)\widetilde{D} + \frac{1}{4},$$
$$\widetilde{\mathcal{G}}^{(2)}(b) = \frac{\widetilde{\mathcal{G}}_{\mathrm{F}}(b) + \widetilde{\mathcal{G}}_{\mathrm{M}}(b)}{2}, \qquad \widetilde{S}(b) = \min_j \widetilde{\alpha}_j(b)\delta_j. \tag{97}$$

Use these quantities in the row-floor and surplus tests and in the score of (49). The $O(\log d)$ grid candidates, certificate tests, and score comparisons are rational; ties are broken lexicographically. The added quarters pay the projection and exponential approximation errors below.

Here is a quantitative acceptance check. If a confidence rectangle containing a regular fully mixed A has $0 < \varepsilon \le \Gamma_{\mathrm{cert}}/16$, the resolvent estimates in (34) give $k_2 \le 4\overline{\kappa}\sqrt{d}$ and $k_\infty \le 4\overline{\kappa}$. At the fixed candidate $b = 3$,

$$\widetilde{\mathcal{R}}(3) \le 12\overline{\kappa}\sqrt{d}, \qquad \widetilde{D} \le 2\sqrt{d}, \qquad H^2 \le d.$$

Since $C^{\mathbb{Q}}(3) \ge \widetilde{\mathcal{R}}(3)$, (93) and (97) imply

$$\widetilde{\mathcal{G}}_{\mathrm{F}}(3) \le 49\overline{\kappa}d, \qquad \widetilde{\mathcal{G}}_{\mathrm{M}}(3) \le 26\overline{\kappa}d, \qquad \widetilde{\mathcal{G}}^{(2)}(3) \le 40\overline{\kappa}d.$$

The ρ and deletion components of (33) then yield

$$2\{\widetilde{\mathcal{R}}(3) + k_\infty\}\varepsilon \le 32\overline{\kappa}\sqrt{d}\,\varepsilon \le \rho/32 < \underline{\rho},$$

and, using $\widetilde{\alpha}_j(3) \ge 2\varepsilon$ and $\underline{\Delta} \ge \Delta_{\mathrm{del}}/2$,

$$\frac{16}{7}\widetilde{\mathcal{G}}^{(2)}(3)\varepsilon^2 \le \varepsilon\Delta_{\mathrm{del}} \le 2\varepsilon\underline{\Delta} \le \widetilde{S}(3).$$

The remaining positivity and gap tests hold at $\Gamma_{\mathrm{cert}}/16$ by Theorem 6. Thus the rational candidate $b = 3$ is accepted at that radius; the score-maximizing candidate also passes both tests. At such a fine rectangle, every grid candidate has $\widetilde{\mathcal{R}}(b) \le 8\overline{\kappa}\sqrt{d} + 8\sqrt{d}$. At a coarser accepted rectangle, $\varepsilon > \Gamma_{\mathrm{cert}}/16$, the row-floor test and $\underline{\rho} \le 1$ instead give $\widetilde{\mathcal{R}}(b) < 1/(2\varepsilon) < 8/\Gamma_{\mathrm{cert}}$. These two bounds, together with $\widetilde{D} \le 2\sqrt{d}$, bound the accepted curvature and dyadic cap by universal multiples of the quantities used in Theorem 3. The accepted row-floor and gap tests still give $\underline{\rho} \ge \rho/2$ and $\underline{\Delta} \ge \Delta_{\mathrm{del}}/2$ by the same comparison in (45). Hence the rational controller retains the main theorem's horizon and instance-parameter orders, with changed universal constants.

Lemma 18 (Rational final local epoch)**.** *Under the finite-input model of Section M, one accepted local epoch and its persistent two-expert state have a rational implementation with polynomial bit complexity. The bound in Theorem 7 remains valid after increasing each curvature coefficient by a universal constant. Relative to the exact learning rate for the same rational cap, each cap-segment prior term increases by at most a factor of two. The self-financing prefix inequality remains exact for the implemented rational policies, with a universal-factor larger booked liability.*

Proof. Use the fraction-free certificate solver for the center and Lemma 17 for centered FTRL. The exact-zero branch plays the identified maximin strategy without creating local state. For $\varepsilon > 0$, halve the dyadic tolerance in (96); every feasible rational projection then has Euclidean error at most

$$\xi_t \le \frac{\varepsilon^2}{8\lceil\sqrt{d}\rceil}.$$

For every subsequently revealed column, replacing an exact projection by this feasible approximation changes its payoff by at most $\sqrt{d}\,\xi_t \le \varepsilon^2/8$. This contribution is added to the FTRL curvature coefficient.

For each cap segment, use a positive rational learning rate $\widetilde{\lambda}$ satisfying

$$\frac{\lambda}{2} \le \widetilde{\lambda} \le \lambda, \qquad \lambda = (\sqrt{d}\,C_r)^{-1},$$

Binary bisection applied to the rational upper square-root operation in Section M computes such a number with polynomial encoding length. Replacing λ by $\widetilde{\lambda}$ increases $\log(1/\pi_a)/\widetilde{\lambda}$ by at most a factor two relative to the exact prior term and can only decrease the quadratic Hoeffding term.

It remains to implement the exponential update. At round t, compute a positive rational multiplier $m_{t,a}$ for each expert such that

$$e^{\widetilde{\lambda}u_{t,a}}(1-\zeta_t) \le m_{t,a} \le e^{\widetilde{\lambda}u_{t,a}}(1+\zeta_t), \qquad \zeta_t = \min\left\{\frac{1}{16}, \frac{\widetilde{\lambda}\varepsilon_{e(t)}^2}{16}\right\}.$$

Here $e(t)$ is the current epoch. Approximate the exponential to absolute error $\zeta_t/4$ and round to a multiple of 2^{-p_e}, where $p_e = \lceil \log_2(8/\zeta_t) \rceil$ within epoch e. Since the argument lies in $[0,1]$, this gives a positive $O(p_e)$-bit multiplier with the stated relative error. Normalize weights only when forming the expert mixture.

Writing $\theta_{t,a} = \log m_{t,a} - \widetilde{\lambda} u_{t,a}$ gives $|\theta_{t,a}| \le 2\zeta_t$. To account for this error directly, let $W_t = \sum_a w_{t,a}$, let $q_{t,a} = w_{t,a}/W_t$, and initialize $w_{1,a} = \pi_a$ with the cap-segment prior. Hoeffding's log-sum-exp inequality gives

$$\log \sum_a q_{t,a} e^{\widetilde{\lambda} u_{t,a}} \le \widetilde{\lambda} \sum_a q_{t,a} u_{t,a} + \frac{\widetilde{\lambda}^2}{8} \operatorname{range}(u_t)^2.$$

Indeed, for $f(s) = \log \sum_a q_{t,a} e^{s u_{t,a}}$, $f'(0) = \sum_a q_{t,a} u_{t,a}$ and $f''(s) = \operatorname{Var}_{q^{(s)}}(u_t) \le \operatorname{range}(u_t)^2/4$. Integrating twice from $f(0) = 0$ proves the bound. The multiplier bounds and $\log(1+\zeta_t) \le \zeta_t$ therefore give

$$\log \frac{W_{t+1}}{W_t} \le \widetilde{\lambda} \sum_a q_{t,a} u_{t,a} + \frac{\widetilde{\lambda}^2}{8} \operatorname{range}(u_t)^2 + \zeta_t.$$

For either expert a, $W_{s+1} \ge w_{s+1,a}$ also gives

$$\log \frac{W_{s+1}}{W_1} \ge \log \pi_a + \widetilde{\lambda} \sum_{t=1}^{s} u_{t,a} - 2 \sum_{t=1}^{s} \zeta_t.$$

Summing the first display and comparing with the second adds at most

$$\frac{3}{\widetilde{\lambda}} \sum_{t=1}^{s} \zeta_t \le \frac{3}{16} \sum_{t=1}^{s} \varepsilon_{e(t)}^2 \le \frac{1}{4} \sum_e s_e \varepsilon_e^2$$

to every expert comparison on a prefix of length s, while preserving the range-squared term required by Lemma 12. If the positive rational radius ε has encoding length L, then $\varepsilon \ge 2^{-L}$ and $\log(1/\zeta_t) = O(L + \log T + \log d)$, so the exponential approximation uses polynomial resources. Confidence radii and certificate outputs share a polynomial bit bound P_0 depending only on observation statistics; thus p_e is independent of prior weight denominators. Each retained weight is its rational prior times at most T dyadic multipliers and has length at most $P_0 + O(T \max_e p_e)$. The mixture and anchor inherit this bound.

The scale-free controllers supply only the proposal in Lemma 11, not comparators in Theorem 7. Set $p_{\mathrm{prop}} = \lceil 4 \log_2(T + d + 2) \rceil$. After each inner and outer projected update, round the first mixture coordinate to the nearest $2^{-p_{\mathrm{prop}}}$ multiple in $[0, 1]$ and set the second to its complement. Obtain nonzero square-root learning rates within a factor two by rational bisection; skip zero-variation updates. Variation sums are exact: their summands depend on the center, FTRL policy, anchor, and rounded weights, but not preceding variation sums. Each summand has a common polynomial bit bound, preserved through at most T additions. The predictable proposal remains in $\mathcal{X}_e$; the theorem requires no trajectory identity with an exact-real implementation.

Compute rational upper bounds $\overline{H}_e \ge H_e$ and $\overline{J}_e \ge J_e$, each within a factor two, using the upper square-root routine in Section M. Define the booked amount $\overline{B}_e = 2a_e(\overline{H}_e + \overline{J}_e)$. Then $\underline{g}_t$ is rational and the updates $\alpha_t = \min\{1, Z_t/\overline{B}_e\}$ and $Z_{t+1} = Z_t + \alpha_t \underline{g}_t$ are exact rational operations. The proof of Lemma 11 applies with $\overline{B}_e$: wealth remains nonnegative, and the additional prefix regret is at most $\overline{B}_e$. The upper-norm approximation changes the booked liability only by a universal factor. Moreover, $\underline{g}_t$ depends on the bounded anchor and proposal states, not on the wealth. It and $\overline{B}_e$ have a common polynomial encoding bound P_1. When $\overline{B}_e > 0$, the exact wealth update has the form

$$Z_{t+1} = \begin{cases} Z_t + \underline{g}_t, & Z_t \ge \overline{B}_e, \\ Z_t(1 + \underline{g}_t/\overline{B}_e), & 0 \le Z_t < \overline{B}_e. \end{cases}$$

Each branch contains one occurrence of the retained wealth and an operand of $O(P_1)$ bits. Its numerator and denominator lengths therefore grow by at most $O(P_1)$ per round, giving $O(TP_1)$ bits over the horizon. At $\overline{B}_e = 0$ the gain and wealth remain zero. Forming α_t and the published distribution now also has polynomial cost. Observed payoffs enter the confidence statistics with their supplied b-bit encoding, so the policy denominator does not enter those statistics.

All expert policies remain feasible, so their convex mixtures preserve the certified row floor. Every operation uses rational inputs of polynomial encoding length, and the projection bisections and exponential series require polynomially many iterations. □

Corollary 4 (Turing implementation of IC-RF)**.** *Under the finite-input model, the rationalized IC-RF controller has expected running time and memory polynomial in T, d, the payoff and δ encoding lengths. On the regular fully mixed instance class of Theorem 3, its high-probability regret has the same horizon and instance-parameter orders, with different universal constants.*

Proof. Theorem 10 implements the core primitives, and Lemma 18 implements the persistent local state. The $b = 3$ candidate is accepted at $\Gamma_{\mathrm{cert}}/16$, and $L_T^{\mathbb{Q}} = O(\mathcal{L})$ makes the count-envelope cost a universal factor. Equation (92), the range-squared comparison, and the two radius bounds supply the local and gate inequalities of Theorem 3. Its proof applies with $\Gamma_{\mathrm{cert}}/16$ and changed universal constants. At most T rounds and $O(\log T)$ scales, including exact fair-bit sampling, have polynomial expected cost by Theorem 10. □