ADAPTIVE MAXIMIN LEARNING WITH INTERMITTENT OPPONENT OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study an unknown zero-sum matrix game against an adaptive opponent, with m learner actions, n opponent actions, and T rounds. Rewards are always observed; opponent actions are independently revealed with unknown probability $q \in [0, 1]$. The benchmark is the best mean payoff a fixed action can guarantee against every opponent action. Regret is the expected sum of this value minus the mean payoff of each chosen action pair. A base policy given only m, n, T achieves worst-case regret $O(\sqrt{mT \log(mnT)})$ and two gap-dependent logarithmic bounds. One uses revealed actions and scales as $1/q$ for $q > 0$. The other holds with the same constant for every q at a strict pure saddle, a pair of actions that are unique best responses to each other. On a family with tied opponent best responses, the revealed-action bound has a matching fixed-instance lower bound for policies with logarithmic worst-case regret on every fixed matrix at every fixed $q > 0$. An enhanced policy combines row-reward estimates with revealed-action frequencies. It uses the same inputs, preserves the base guarantees, and attains $q^{-1/2}$ dependence on a different two-row family for $q > 0$.

1 INTRODUCTION

Repeated games model decisions whose outcomes depend on the choices of other players. When payoffs are unknown, a player must learn from the rewards received during play, while the opponent may keep changing its behavior. How much payoff can the player guarantee throughout this process? Exploration is part of play, so even the learning rounds count. Early work studies this cost within known game families (Shimkin & Shwartz, 1995; 1996; Conitzer & Sandholm, 2003); later work considers adaptation and protection against exploitation when payoffs are known (DiGiovanni & Tewari, 2022).

Consider a zero-sum matrix game. Rows and columns represent the learner's and opponent's actions; each entry gives the learner's mean payoff, and the opponent's mean payoff is its negative. With *bandit feedback*, the learner observes only the reward of the chosen action pair. To assess learning, regret compares accumulated payoffs with a benchmark. Adversarial bandits compare with the best fixed action in hindsight (Auer et al., 2002). For unknown matrices with noisy rewards and observed actions, O'Donoghue et al. (2021) use the Nash value, the payoff an optimal randomized strategy guarantees against every opponent strategy. Related work studies the observations needed to identify approximate equilibria in noisy two-column games (Maiti et al., 2023), or pure equilibria under noisy queries to arbitrary matrix entries (Maiti et al., 2024). At an equilibrium, neither player gains by changing its strategy alone; at a pure equilibrium, each chooses a fixed action.

A useful starting point for learning during play is the payoff that a single fixed action can guarantee. Even an opponent that knows this action cannot drive its mean payoff below the smallest entry in its row. Taking the largest of these row minima gives the *pure maximin value*. Ito et al. (2026) introduce pure-strategy maximin regret (PSMR), the expected sum of this value minus the mean payoff of each chosen action pair. Favorable responses offset losses on other rounds: an opponent that makes exploration look attractive may also pay for part of it. Their work obtains logarithmic regret under two conditions. Rewards alone suffice at a *strict pure saddle*, where each player's action is its unique best response to the other's action. When opponent actions are always observed, another logarithmic bound depends on entries below the pure maximin value. Either bound can be

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.

smaller, and choosing between the policies requires information about the unknown matrix. Can a single learner obtain both guarantees?

We answer this question when opponent actions are revealed intermittently. Rewards are observed every round; the opponent's action is independently revealed with an unknown probability $q \in [0, 1]$. Our base policy π^0 receives only the numbers m, n of learner and opponent actions and the number of rounds T. It simultaneously attains worst-case regret $O(\sqrt{mT \log(mnT)})$ and both logarithmic guarantees. The revealed-action bound includes a factor $1/q$ for $q > 0$. The strict-saddle bound holds with the same constant for every q, including zero. The learner benefits from whichever bound is smaller, using the observations available during play.

A revealed action locates its reward in the matrix. When the column is hidden, a low reward may reflect the row, the opponent's choice, or noise. The cost of limited feedback also motivates label-efficient prediction (Cesa-Bianchi et al., 2005), feedback graphs (Mannor & Shamir, 2011; Alon et al., 2015; Esposito et al., 2022), and partial monitoring (Bartók et al., 2014; Lattimore & Szepesvári, 2019). In our model, rewards are always available, and column observations help interpret them. Proposition 1 shows that the $1/q$ dependence is necessary on a family with tied opponent best responses: it gives a matching fixed-instance lower bound for policies with logarithmic worst-case regret on every fixed matrix at every fixed $q > 0$.

The base policy combines entry estimates from revealed rounds with comparisons of row rewards. These comparisons account for gains that offset exploration losses. Combining guarantees within one policy also motivates best-of-both-worlds bandits (Bubeck & Slivkins, 2012; Auer & Chiang, 2016; Zimmert & Seldin, 2021), data-dependent improvements (Wei & Luo, 2018), and extensions to broader learning problems (Dann et al., 2023). Our reward-comparison rule adjusts its learning rate during play using the methods of Ito et al. (2024); Nguyen et al. (2025). Its comparison cost depends on how often the learner tries each alternative row, which lets gains on favorable rounds offset the cost of these choices.

The payoff structure can make each revealed action more useful. Revealed actions estimate column frequencies; these frequencies help separate the columns' contributions to rewards even on rounds with hidden actions. Our enhanced policy π^+ uses these frequencies to separate the two columns' payoffs. It weights row rewards so that both rows are compared against the same opponent choices. Given the same inputs m, n, T, it preserves the base guarantees and improves the dependence on q to $q^{-1/2}$ on a different two-row family for $q > 0$. We first construct the base policy and explain how reward comparisons and entry estimates give its three guarantees. We then show how column-frequency estimates extract further information from rewards and yield the enhanced bound.

2 Model and simultaneous guarantee

Write $[k] = \{1, \dots, k\}$ and $\Delta_k = \{x \in \mathbb{R}^k : x_i \geq 0, \sum_i x_i = 1\}$. The vector e_i is the ith standard basis vector, $\mathbf{1}_k$ is the all-ones vector, $\mathbf{1}\{E\}$ indicates an event E, and $(x)_+ = \max\{x, 0\}$. We use natural logarithms in the following parts.

Let $A \in [-1, 1]^{m \times n}$ be a fixed, unknown payoff matrix, with integer dimensions $m, n \geq 2$ and known integer horizon $T \geq 1$. Rows are learner actions, columns are opponent actions, and entries are mean payoffs. The benchmark is the best payoff a fixed row can guarantee against every column: $v = \max_i \min_j A_{ij}$. We call v the pure maximin value and a row attaining it a maximin row or a safe row. The learner uses observations of play to approach this safety level.

Let $\mathcal{F}_{t-1}$ denote the complete history through round $t - 1$, including past opponent actions. The learner uses the part it observes. At round t, it chooses a distribution $p_t \in \Delta_m$. The opponent then chooses a column $J_t \in [n]$, using the history and p_t. Next the learner makes its fresh draw $I_t \sim p_t$ and observes a reward $r_t \in [-1, 1]$. This timing permits an adaptive opponent while keeping the current actions independent conditional on the history and p_t. Rewards may depend on the past; their conditional means satisfy

$$\mathbb{E}[r_t \mid \mathcal{F}_{t-1}, J_t, I_t] = A_{I_t J_t}. \tag{1}$$

Fix a revelation probability $q \in [0, 1]$, unknown to the learner. Independent revelation indicators $B_t \sim \text{Bernoulli}(q)$ reveal J_t exactly when $B_t = 1$ and are independent of the past, current actions,

and reward. The learner observes r_t, B_t, and, on a revealed round, J_t. Thus every round supplies a reward, and a fraction q of rounds supply its column label in expectation.

We compare each round's mean payoff with v, allowing gains above this level to offset losses below it. The expected sum of these differences is signed pure-strategy maximin regret (PSMR):

$$R_T = \mathbb{E}\sum_{t=1}^{T}(v - A_{I_t J_t}). \tag{2}$$

For example, one mean payoff $v + \epsilon$ offsets one mean payoff $v - \epsilon$ for any $\epsilon > 0$ within the reward range. These gains can therefore reduce the cost of exploration.

Two features of the matrix determine the logarithmic guarantees. First, an entry below v incurs a loss each time it is played. For an entry with positive gap Δ, a confidence interval reaches accuracy proportional to Δ after a number of revealed samples proportional to $1/\Delta^2$, up to logarithmic factors. Multiplying by its per-play loss Δ yields an inverse-gap cost. We therefore use

$$\Delta_{ij} = v - A_{ij} \quad (A_{ij} < v), \qquad C_c = \sum_{(i,j):A_{ij}<v} \frac{1}{\Delta_{ij}}. \tag{3}$$

The empty sum is zero. The second feature concerns both losses and gains at a pure saddle.

A pair $(i_*, j_*) \in [m] \times [n]$ is a *pure saddle* if $A_{ij_*} \le A_{i_*j_*} \le A_{i_*j}$ for all i, j, with $A_{i_*j_*} = v$. It is *strict* when both inequalities are strict for $i \neq i_*$ and $j \neq j_*$. Against the saddle column, other rows incur losses; along the saddle row, other columns increase payoff. At a strict saddle, we summarize these gaps by

$$\alpha_i = v - A_{ij_*} > 0 \ (i \neq i_*), \quad \beta = \min_{j \neq j_*}(A_{i_*j} - v) > 0, \quad C_s = \frac{1}{\beta}\sum_{i \neq i_*}\frac{1}{\alpha_i}. \tag{4}$$

Here α_i measures the loss from row i against the saddle column, and β is the gain guaranteed to the safe row when the opponent uses another column. A smaller β leaves less gain to offset costly row choices; Section 4 explains their product in C_s.

Either of C_s and C_c can be smaller. For $0 < \varepsilon \le 1/4$, the matrices

$$\begin{pmatrix} 0 & \varepsilon \\ -1/4 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 1/4 \\ -1/4 & -\varepsilon \end{pmatrix}$$

have $(C_s, C_c) = (4/\varepsilon, 4)$ and $(16, 4 + 1/\varepsilon)$, respectively. In the first matrix, decreasing ε increases C_s while C_c stays fixed; in the second, it increases C_c while C_s stays fixed. The policy below satisfies all applicable bounds simultaneously, so its regret is controlled by whichever is smallest.

Theorem 1 (Simultaneous adaptation)**.** *There is a policy π^0, given only m, n, T, and a universal constant C such that, with $L_T = \log(192m^2(n+1)(T+1)^5)$,*

$$\begin{aligned} &R_T \le C\sqrt{mTL_T} + 1 && \text{for every } A \text{ and } q, \\ &R_T \le CL_TC_c/q && \text{for every } A \text{ if } q > 0, \\ &R_T \le CL_TC_s + 1 && \text{at a strict pure saddle, for every } q. \end{aligned} \tag{5}$$

These bounds hold simultaneously. If $C_c = 0$, every policy has $R_T \le 0$.

The same policy satisfies all three bounds without knowing the gaps or q. Reward comparisons control exploration costs, while revealed actions estimate matrix entries. Appendix A.2 states the result with explicit constants; Section 4 explains the three guarantees.

To understand when the factor $1/q$ is necessary, consider a safe row that pays zero in every column, while each other row is negative in column 1 and zero elsewhere. The opponent can make column 1 rare, weakening its effect on rewards when columns are hidden. Revealed labels allow its observations to be counted separately. To learn on every matrix, a policy must also distinguish this matrix from ones where another row is best. This requires trying rows that are costly on the original matrix.

Proposition 1 (Matching lower bound for revealed actions)**.** *For every $m, n \ge 2$, there is a family of matrices with a pure saddle and tied opponent best responses. Consider any policy that achieves*

logarithmic worst-case regret on every fixed matrix at every fixed positive revelation probability. For each fixed matrix A in this family and each fixed $q > 0$, the policy incurs worst-case regret $\Omega(C_c(A) \log T/q)$ as $T \to \infty$. The policy uses the observations of Section 2; the worst case ranges over the opponents and reward laws allowed there.

Appendix A.3 gives the precise statement and Appendix B.1 its proof.

Policy and feedback	Worst case	Strict saddle	Entries below v
Ito et al. (2026), reward-based	$\sqrt{mT}$	LC_s	—
Ito et al. (2026), informed	$\sqrt{mnTL}$	—	LC_c
This paper, unknown q	$\sqrt{mTL}$	LC_s	LC_c/q

Table 1: Orders of the guarantees compared here, with logarithmic terms collected in $L = O(\log(mnT))$. Each row refers to one policy; the two prior rows are different policies. The bounds in the last row hold simultaneously, with the entry bound omitted at $q = 0$. The prior informed policy assumes every column is observed.

3 THE BASE POLICY

Each round uses past observations to choose a *leader*: the row with the largest upper bound on its minimum mean payoff. The policy gives it at least half the play probability. An internal learner allocates the rest among eligible rows. Every reward updates row bounds and the learner; revealed columns also update entry and column bounds.

Write $L = L_T$. Statistics at the start of round t use observations through $t - 1$; we omit the current-round subscript. Ties follow a fixed rule.

Bounds available before play. Let U_{ij} be an upper confidence bound on A_{ij}. It starts at one and takes the minimum of the bounds computed from revealed observations of this entry. When these bounds hold, $\min_j U_{ij}$ upper-bounds row i's minimum payoff.

To stop exploring a low-payoff entry, we need evidence of higher payoffs in the same column. Pool the rewards from all rows played when column j is revealed. Their conditional means are the successive row distributions' payoffs against j, each at most $\max_i A_{ij}$. For every length $1, 2, 4, \ldots$, partition this sequence into consecutive blocks. If $\widehat{r}_j(D)$ is a completed block's average reward, its lower confidence bound also lower-bounds the column maximum. Set

$$H_j = \max\left\{-1, \max_D \left(\widehat{r}_j(D) - \sqrt{2L/|D|}\right)\right\}. \tag{6}$$

Initially $H_j = -1$. Short blocks capture brief periods; long blocks narrow the confidence interval. The running maximum preserves earlier evidence.

Choosing the action. Store an upper bound M_i on each row's minimum mean payoff, initially one. Set $w = \max_i M_i$ and $b \in \arg\max_i M_i$. A row i is *eligible* if at most one column satisfies $U_{ij} \le w$ and that column, if present, satisfies $U_{ij} \ge H_j$. Denote the eligible set by E. Allowing one column with $U_{ij} \le w$ accommodates the strict saddle row's single minimizing column whenever w lies below its other payoffs. The second test excludes a row when this entry's upper bound falls below H_j.

The internal learner returns positive recommendation probabilities $z \in \Delta_m$. It uses follow-the-regularized-leader (FTRL), minimizing cumulative estimated loss plus a convex regularizer. Map eligible recommendations to their own rows and the others to the leader: $f(i) = i$ for $i \in E$, and $f(i) = b$ otherwise. The mapped recommendation selects action a with probability $Q_a = \sum_{i:f(i)=a} z_i$. The play distribution and active alternatives are

$$p = \tfrac{1}{2}e_b + \tfrac{1}{2}Q, \qquad \mathcal{A} = E \setminus \{b\}, \qquad s = \sum_{i \in \mathcal{A}} z_i. \tag{7}$$

Thus $p_i = z_i/2$ on $\mathcal{A}$ and $p_b = 1 - s/2 \ge 1/2$. If E is empty, as initially when $n \ge 2$, then $p = e_b$. The leader's rewards still update its bound, even without revelations.

Updating the payoff bounds. A revealed column updates U, H. Every reward also gives an estimate of the leader's payoff:

$$X_t = \mathbf{1}\{I_t = b_t\} r_t / p_t(b_t). \tag{8}$$

It has conditional mean $A_{b_t J_t}$ and magnitude at most two; playing another row records zero.

To compare with an unknown maximin row, we save records for every candidate a. In the regret decomposition, the leader's deficit has coefficient $1/2$ when a is eligible and one when a is mapped to the leader. Keep these cases separate: for each a, append the same X_t to group $(b_t, a, \mathbf{1}\{a \in E_t\})$, using the saved eligibility set. Each group's true average payoff is at least its leader row's minimum. Its empirical mean after k records plus $\sqrt{8L/k}$ therefore upper-bounds that minimum. Let $P_{i,a,e}$ be the running minimum of these bounds, initially one. Store

$$M_i = \max\{-1, \min(1, \min_j U_{ij}, \min_{a,e} P_{i,a,e})\} \tag{9}$$

for the next round.

Updating the internal learner. The learner estimates each active alternative's loss relative to the leader. Define the observed loss $\ell = (1 - r)/2$ and mean entry loss $\bar{\ell}_{ij} = (1 - A_{ij})/2$. Initialize $u \in \mathbb{R}^m$ to zero. If $I = i \in \mathcal{A}$, set $u_i = \ell / p_i$. If $I = b$ and $s > 0$, privately draw $K \in \mathcal{A}$ with probability $\nu_i = z_i / s$, independently of reward and revelation indicator, and set $u_K = -\ell/(p_b \nu_K)$. Use $u = 0$ when $s = 0$. For each active i, the positive branch has expected contribution $\bar{\ell}_{iJ_t}$ and the negative branch $-\bar{\ell}_{bJ_t}$, giving

$$\mathbb{E}[u_i \mid \mathcal{F}_{t-1}, J_t] = \bar{\ell}_{iJ_t} - \bar{\ell}_{bJ_t} \qquad (i \in \mathcal{A}). \tag{10}$$

Each feedback vector has at most one nonzero coordinate. The FTRL regularizer and coordinate-wise learning-rate update follow Nguyen et al. (2025, Eq. (17)). Mixing its output with the uniform distribution at rate $1/(T+1)$ ensures $z_i \geq 1/[m(T+1)]$, making the divisions valid. Appendices B.2–B.3 give the full updates and implementation.

Algorithm 1 The base policy π^0

Require: Dimensions m, n and horizon T.

1: Set $M_i = U_{ij} = P_{i,a,e} = 1$ and $H_j = -1$; initialize empty reward records and the internal FTRL learner.
2: **for** $t = 1, \ldots, T$ **do**
3: Use the stored row bounds M_i to choose the leader b, the largest bound w, and the eligible set E.
4: Obtain z from the FTRL state. Form f, Q, p and the active set $\mathcal{A}$ as in (7).
5: Announce p. After the opponent chooses J_t, draw $I \sim p$ and observe the reward and revelation feedback.
6: **if** the column is revealed **then**
7: Update $U_{I_t J_t}$ and H_{J_t} with the revealed reward.
8: **end if**
9: Form X by (8). For each a, append it to $(b, a, \mathbf{1}\{a \in E\})$ and update that P; update M by (9).
10: Form the loss-difference estimate u using the saved probabilities and update the internal learner.
11: **end for**

4 Regret analysis

Reward comparisons prove the worst-case and strict-saddle bounds using every reward. We then count revealed entry observations and show how column bounds stop further exploration. Independence of the current revelation indicator converts this second calculation into total regret.

Reward comparisons. Write $L = L_T$ and fix a maximin row a. For the realized history and chosen column J_t, let $\mu_t = \sum_i p_t(i) A_{iJ_t}$ be the actual mean payoff. Compare $p_t = (e_{b_t} + Q_t)/2$

with the reference distribution $(e_{b_t} + e_{f_t(a)})/2$, using the same saved leader, eligibility set, and chosen column. Its first half selects the leader; its second selects a when eligible and the leader otherwise. The learner's comparison term has conditional mean equal to the reference mean payoff minus μ_t. Thus

$$R_T = \mathbb{E}\sum_t d_t^a + \mathbb{E}\sum_t(\langle z_t, u_t\rangle - u_{ta}), \qquad d_t^a = \tfrac{1}{2}(v - A_{b_t J_t}) + \tfrac{1}{2}(v - A_{f_t(a) J_t}), \tag{11}$$

where d_t^a is the benchmark minus the reference payoff.

Set $\delta = 1/[6(T+1)^2]$. With probability at least $1-\delta$, all entry, reward-group, and block confidence bounds hold at every update. Then $M_a \geq v$ and $w \geq v$. Consider a group's first $K \geq 2$ records. Before the last is added, its row is still the leader, so its reward upper bound is at least v. With radius $\varepsilon = \sqrt{8L/(K-1)}$, the recorded average must be at least $v - \varepsilon$; concentration places the true average at least $v - 2\varepsilon$. The last deficit is at most two. For the corresponding rounds $\mathcal{D}$, including $K = 0, 1$,

$$\sum_{t\in\mathcal{D}}(v - A_{b_t J_t}) \leq 4\sqrt{2LK} + 2. \tag{12}$$

The learner's comparison cost can be expressed through actual choices. Let $N_i^{\text{play}} = \sum_t \mathbf{1}\{I_t = i\}$. The FTRL guarantee implies, for each fixed recommendation a,

$$\mathbb{E}\sum_t(\langle z_t, u_t\rangle - u_{ta}) \leq C\sqrt{L}\sum_{i\neq a}\sqrt{\mathbb{E}N_i^{\text{play}}} + CmL, \tag{13}$$

where C denotes a universal constant. An active alternative is played with probability $z_{ti}/2$. Whenever a itself is active, another row is the leader and receives probability at least one half. Thus the feedback costs for all coordinates can be bounded using choices of rows other than a.

For the worst-case bound, an eligible maximin row pays at least v. Apply (12) to its eligible and ineligible groups, and sum over leaders. Since $\sum_i N_i^{\text{play}} = T$, Cauchy–Schwarz in (11) gives $O(\sqrt{mTL} + mL)$. For $T < mL$, use $R_T \leq 2T$.

The bound at a strict saddle. Take $a = i_*$. On the confidence event, $U_{i_* j} \geq v + \beta$ for $j \neq j_*$ and $H_{j_*} \leq v \leq U_{i_* j_*}$. Hence $i_* \notin E_t$ requires $w_t \geq v + \beta$. The record argument then uses the higher benchmark: $\sum(v - A_{b_t J_t}) = \sum(v + \beta - A_{b_t J_t}) - \beta K$. Maximizing the square-root uncertainty against half this negative term leaves $O(L/\beta) - \beta K/2$ per group.

When i_* is eligible, consider a nonsaddle leader i. Let K_i^* count this group's rounds with column j_* and O_i its other rounds. Its cumulative deficit is at least $\alpha_i K_i^* - 2O_i$ and at most $O(\sqrt{L(K_i^* + O_i)} + 1)$ by (12). These inequalities bound it by $CL/(\alpha_i\beta) + \beta O_i/8$. On the same O_i rounds, the comparator pays at least $v + \beta$, offsetting the linear term.

Let $\mathcal{J} = \{t : i_* \in E_t, J_t = j_*\}$ and $U = \mathbb{E}(T - |\mathcal{J}|)$. Define the total gap-weighted choices of nonsaddle rows, including their plays as leader, by

$$S = \sum_{i\neq i_*}\alpha_i \mathbb{E}N_i^{\text{play}}, \qquad A_\alpha = \sum_{i\neq i_*}\alpha_i^{-1}.$$

Weighted Cauchy–Schwarz gives $\sum_{i\neq i_*}\sqrt{\mathbb{E}N_i^{\text{play}}} \leq \sqrt{A_\alpha S}$. Summing the leader bounds and comparator gains therefore yields

$$R_T \leq CLC_s + CmL + C\sqrt{LA_\alpha S} - \frac{7\beta}{16}U + 4T\delta. \tag{14}$$

On $\mathcal{J}$, $\sum_{i\neq i_*} p_t(i)\alpha_i = v - \mu_t$. Elsewhere this sum is at most two, while $v - \mu_t \geq -2$. Thus

$$S \leq R_T + 4U. \tag{15}$$

Substituting $U \geq (S - R_T)/4$ leaves a square root minus a linear term in S, with maximum $O(LA_\alpha/\beta) = O(LC_s)$. The coefficient of R_T is $1 - 7\beta/64 \geq 25/32$, and $m \leq 8C_s$ absorbs CmL. This proves $O(LC_s) + 1$ for every $q \in [0, 1]$.

The bound from revealed entries. Return to an arbitrary matrix. For a entry with mean payoff below v, set $\Delta_{ij} = v - A_{ij} > 0$ and $k_{ij} = \lceil 512L/\Delta_{ij}^2 \rceil$. Let c_t sum the gaps times their play probabilities over such entries in the current column J_t whose revealed counts before this round are below their thresholds. Each such entry has at most k_{ij} revealed visits before reaching its threshold, so, unconditionally, $\mathbb{E}\sum_t B_t c_t \le \sum_{\Delta_{ij}>0} \Delta_{ij} k_{ij} = O(LC_c)$.

After reaching its threshold, a entry satisfies $M_i \le U_{ij} \le v - 7\Delta_{ij}/8 < v \le w$ on the confidence event, excluding its row from leadership. It can still be explored if this is its only column with $U_{ij} \le w$ and $H_j \le U_{ij}$. Let $\rho_j = (v - H_j)_+$ denote how far H_j lies below v. Such entries have gaps at most $8\rho_j/7$ and total play probability at most one half. On a round with $J_t = j$, therefore, $v - \mu_t \le c_t + 4\rho_j/7$; when $H_j < v$,

$$\mu_t - H_j \ge 3\rho_j/7 - c_t.$$

This explains how H_j rises. Its blocks pool all rewards on revealed visits to column j. When their average c_t is small, their mean payoffs exceed the old lower bound; enough samples make a completed block's confidence interval narrow enough to raise it. Large c_t can delay this increase by lowering the column average. Divide the possible distances $\rho_j \in (0, 2]$ into successive intervals $(x, 3x/2]$. For one interval, let C_x sum c_t over its revealed visits. The block argument bounds their number by $O(C_x/x + L/x^2)$ and their cumulative deficit by $O(C_x + L/x)$. Thus the same c_t sum covers both its direct loss and the extra loss during this delay, up to a constant factor.

For each column containing such entries, let Δ_j be its smallest positive gap. Once $\rho_j < 7\Delta_j/8$, every such entry past its threshold fails the H_j test. The terms L/x sum geometrically to $O(L/\Delta_j)$; over these columns, $\sum_j 1/\Delta_j \le C_c$. The intervals use disjoint visits, so the C_x terms sum to at most $\sum_t B_t c_t$. With $d_t = v - \mu_t$, accounting for confidence failures and using independence of the current revelation indicator gives

$$qR_T = \mathbb{E}\sum_t B_t d_t \le 33000LC_c + 2T\delta. \tag{16}$$

For $C_c > 0$, $C_c \ge 1/2$, proving the entry bound. Appendix C gives the complete proofs and constants.

5 The enhanced policy

Column-frequency estimates can help separate the payoffs of two columns even when some column labels are hidden.

The enhanced policy adds entry inference when $m = n = 2$ and uses the base policy otherwise. In column 1, the safe row has mean payoff zero and the other row has mean payoff $-\alpha$, with $0 < \alpha \le 1/16$. In column 2, their mean payoffs are positive: β and $1/2 - \alpha$, respectively, with $0 < \beta \le 1/96$. Frequent choices of column 2 can make the other row's average look favorable; these choices also reduce regret. The two row orderings give

$$A^+ = \begin{pmatrix} 0 & \beta \\ -\alpha & 1/2 - \alpha \end{pmatrix}, \qquad A^- = \begin{pmatrix} -\alpha & 1/2 - \alpha \\ 0 & \beta \end{pmatrix}. \tag{17}$$

Here $v = 0$, $C_c = 1/\alpha$, and $C_s = 1/(\alpha\beta)$, so the two base logarithmic bounds scale as $1/(\alpha q)$ and $1/(\alpha\beta)$. Estimating how often column 2 appears helps distinguish its favorable payoffs from the first-column loss.

Theorem 2 (Enhanced adaptation)**.** *A policy π^+ knowing only m, n, T satisfies all three bounds of Theorem 1 up to universal constants, with logarithm $L_+ = \log(2^{30}m^2(n+1)(T+1)^{10})$. For either matrix in* (17)*, it also satisfies*

$$R_T \le CL_+ \min\left\{\alpha T, \sqrt{T}, \frac{1}{\alpha} + \min\left\{\frac{1}{\alpha\beta}, \frac{1}{\sqrt{\alpha q}}\right\}\right\}$$

for a universal C. Interpret the revelation term as $+\infty$ at $q = 0$.

The terms αT and $\sqrt{T}$ follow from per-round losses and the worst-case bound; $1/(\alpha\beta)$ is the strict-saddle term. The new $1/\sqrt{\alpha q}$ term uses column frequencies, while $1/\alpha$ accounts for row-mean estimation.

There are two ways to incur regret: try the worse row while the safe row leads, or play the worse row while it leads. We compare row rewards to handle the first case and sharpen the leader's payoff bound to handle the second.

Comparing row rewards with the same weights. An opponent may use column 1 early and column 2 later. If the learner changes its row probabilities, unweighted row averages give different weights to these periods. To compare the rows against the same opponent choices, fix an ordered pair (b, a) and use completed rounds with leader b and eligible alternative a. Let W sum the probabilities of choosing a, and form the reward sums

$$W = \sum_t p_t(a), \qquad S_a = \sum_t \mathbf{1}\{I_t = a\} r_t, \qquad S_b = \sum_t \frac{p_t(a)}{p_t(b)} \mathbf{1}\{I_t = b\} r_t. \tag{18}$$

All sums use the same rounds. For row b, the selection probability cancels the denominator: $p_t(b)[p_t(a)/p_t(b)]A_{bJ_t} = p_t(a)A_{bJ_t}$. Row a has the same weight $p_t(a)$. Thus both sums compare payoffs under the same column frequencies.

The algorithm knows W. To describe its column-2 part, define $O = \sum_t p_t(a)\mathbf{1}\{J_t = 2\}$ and, for $W > 0$, $\theta = O/W$. Hidden columns make O unavailable to the algorithm. The normalized sums S_i/W estimate

$$\bar{\mu}_i = (1 - \theta)A_{i1} + \theta A_{i2}, \qquad i \in \{a, b\}. \tag{19}$$

The observable weighted counts $\sum_t p_t(a)B_t$ and $\sum_t p_t(a)B_t\mathbf{1}\{J_t = 2\}$ have conditional-mean sums qW and qO. Their ratio therefore suggests an estimate of θ. The algorithm uses confidence intervals for both counts to bound this ratio, accounting for the unknown q.

To recover the first-column payoff, solve the mixture identity: for $\theta < 1$, $A_{i1} = (\bar{\mu}_i - \theta A_{i2})/(1 - \theta)$. Use the intervals for the row mean, θ, and the entry A_{i2} estimated from revealed samples to bound the right-hand side. When the entire interval for θ is below one, intersect this inferred interval with the revealed-sample interval. Let L_{ij} and U_{ij} be the best lower and upper endpoints so far. Recompute exploration eligibility with these bounds each round. Also exclude row i from exploration if $U_{ij} \le w$ and $U_{ij} < \max_k L_{kj}$ for some column j: this column limits its minimum payoff to at most the current best upper bound w, and another row is already known to pay more in that column.

Limiting play of the leading row. A row's average reward can be high because column 2 is favorable, even when its minimum payoff is low. Among rounds when the row leads and is played, let u bound its average mean payoff from above, $l > 0$ bound its column-2 payoff from below, and $g \le 1/2$ bound the fraction of column-2 plays from below. If its minimum payoff is v_i, at least fraction g has mean payoff at least l, and all other plays have mean payoff at least v_i. Hence $u \ge gl + (1 - g)v_i$, giving the upper bound $(u - gl)/(1 - g)$. The cap on g keeps the denominator at least $1/2$. Take the minimum with the original row bound.

Each leader also has its own internal learner: update it only while that row leads, and resume its saved state when the row leads again. Its cumulative comparison then covers precisely those rounds. The base guarantees continue to hold. On the confidence event, the new exploration test can exclude a strict-saddle row only through a nonsaddle column, since that row maximizes payoff at the saddle column. Such exclusion still requires $w \ge v + \beta$, the condition used in Section 4.

Why the revelation probability enters through a square root. Write h for the safe row and a for the other row. First consider rounds led by h, and let W_s, O_s be the pair's totals through round s. Row a incurs loss α in column 1 and gains $1/2 - \alpha$ in column 2; safe-row plays have nonpositive regret. The cumulative conditional regret is therefore at most

$$\Gamma_s = \alpha(W_s - O_s) - (1/2 - \alpha)O_s = \alpha W_s - O_s/2.$$

The weighted payoff difference between h and a is $\Gamma_s + \beta O_s$. We need a comparison bound at every time, because the last eligible round depends on observations. The internal learner's pathwise comparison inequality, together with concentration of its random loss estimates, gives, with probability at least $1 - 1/[10(T + 1)^3]$, for all $s \le T$,

$$\Gamma_s + \beta O_s \le 1000\big(\sqrt{L_+ W_s} + L_+\big). \tag{20}$$

On the joint confidence event, for $q > 0$, three sufficient conditions make the observable exploration test exclude a: $W \geq CL_+/\alpha^2$, $qO \geq CL_+$, and $\theta \leq 8\alpha$, for a sufficiently large universal C. These are conditions for the proof; the algorithm checks its computed intervals. The first gives row-mean accuracy proportional to α; the second supplies enough column-2 observations. The last makes the column-2 estimation error small after multiplication by θ in the inference formula.

At the last eligible round, at least one condition had yet to hold before the new observation. Write W, O, Γ for the final totals, taking them to be zero if the pair was never eligible, and put $L = L_+$. The last round increases W, O by at most $1/2$ each. Thus at least one of $W \leq CL/\alpha^2 + 1$, $O \leq CL/q + 1$, or $O \geq 8\alpha(W - 1)$ holds. The first gives $\Gamma \leq CL/\alpha$ after increasing C; the last gives $\Gamma \leq 4\alpha$. In the remaining case, $O \leq CL/q + 1$. Substitute $W = (\Gamma + O/2)/\alpha$ into (20). For $\Gamma > 0$, the inequality, valid for every $c > 0$, $c\sqrt{L\Gamma/\alpha} \leq \Gamma/2 + c^2 L/(2\alpha)$ moves half of Γ to the left. The remaining square-root term is $O(\sqrt{LO/\alpha})$, which yields

$$\Gamma \leq CL\left(\frac{1}{\alpha} + \frac{1}{\sqrt{\alpha q}}\right). \tag{21}$$

The bound also holds for $\Gamma \leq 0$.

When a leads, its upper bound must stay nonnegative, since the safe row's bound is at least zero. Repeated first-column losses lower its reward average and eventually contradict this requirement. Second-column gains can delay this drop, but they also reduce regret. Once their payoff and frequency are estimated, the corrected upper bound subtracts a fixed fraction of these gains. This bounds the delay; the resulting inequality has a square-root term in the number of plays and a negative term in the number of column-2 plays, giving (21) again. Add the two cases and the expected regret on confidence failures, then take the minimum with the base bounds and $R_T \leq \alpha T$ to prove Theorem 2; Appendix D.2 gives the full updates and inequalities.

The analysis also covers $\beta = 0$, with a matching fixed-instance lower bound for sufficiently small q (Appendix D.3). Finally, equal C_s, C_c can accompany different learning difficulty: Appendix D.4 gives two known candidate families with these same parameters and different regret orders.

6 Conclusion

Theorem 1 makes two forms of adaptation available within one policy. The learner can use revealed actions to estimate individual entries and rewards to compare rows, without knowing in advance which source will give the smaller regret bound. This matters when observations are scarce: every fixed positive revelation probability gives logarithmic regret on a fixed matrix, while a strict pure saddle allows a logarithmic guarantee even with rewards alone. The worst-case bound controls the cost when small gaps make these logarithmic bounds large. Thus the learner's guarantees follow the available information and the payoff structure throughout play.

The two matrix families explain why the price of missing action observations depends on the game. In Proposition 1, an opponent can make the column with negative entries rare, while every other column pays zero on average. The resulting information cost requires the factor $1/q$. In the enhanced family, choices that make the worse row look attractive also produce positive payoffs. Estimating their frequency helps interpret rewards from hidden-column rounds, and those payoffs offset exploration losses. Extending this combination of payoff comparison and column-frequency inference to larger matrices is a natural next step. A fuller account of learning difficulty would need to describe how informative opponent choices affect the learner's payoff as well as what they reveal.

AI USAGE STATEMENT

We used large language models (LLMs) solely to assist with literature searches, translation, and language editing to improve the clarity and readability of the manuscript. We did not use LLMs for any other tasks listed in the ICLR 2027 AI policy, including developing theoretical models or conceptual frameworks; formulating mathematical claims or contributing to the development or writing of proofs; proposing or refining hypotheses; designing or providing feedback on research methodology or experiments; implementing methods; generating, cleaning, or reformatting datasets; conducting qualitative or thematic analyses; or interpreting results. The authors have reviewed all LLM-assisted content and take full responsibility for the final content of this work, including its accuracy, originality, and integrity.

REPRODUCIBILITY STATEMENT

The supplement provides complete algorithms and proofs, explicit lower-bound matrices, reward laws and opponent schedules, and the computational accounting for statistical updates.

REFERENCES

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 23–35. PMLR, 2015. URL https://proceedings.mlr.press/v40/Alon15.html.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 116–120. PMLR, 2016. URL https://proceedings.mlr.press/v49/auer16.html.

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. doi: 10.1137/S0097539701398375. URL https://epubs.siam.org/doi/10.1137/S0097539701398375.

Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014. doi: 10.1287/moor.2014.0663. URL https://pubsonline.informs.org/doi/10.1287/moor.2014.0663.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 42.1–42.23. PMLR, 2012. URL https://proceedings.mlr.press/v23/bubeck12b.html.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005. doi: 10.1109/TIT.2005.847729. URL https://stoltz.perso.math.cnrs.fr/Publications/CBLS-LabelEff.pdf.

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/e19347e1c3ca0c0b97de5fb3b690855a-Abstract.html.

Vincent Conitzer and Tuomas Sandholm. BL-WoLF: A framework for loss-bounded learnability in zero-sum games. In *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 91–98, 2003. URL https://arxiv.org/abs/cs/0307006.

Chris Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5503–5570. PMLR, 2023. URL https://proceedings.mlr.press/v195/dann23a.html.

Anthony DiGiovanni and Ambuj Tewari. Balancing adaptability and non-exploitability in repeated games. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 559–568. PMLR, 2022. URL https://proceedings.mlr.press/v180/digiovanni22a.html.

Emmanuel Esposito, Federico Fusco, Dirk van der Hoeven, and Nicolò Cesa-Bianchi. Learning on the edge: Online learning with stochastic feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34776–34788. Curran Associates, Inc., 2022. doi: 10.52202/068431-2520. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/e0e956681b04ac126679e8c7dd706b2e-Abstract-Conference.html.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adaptive learning rate for follow-the-regularized-leader: Competitive analysis and best-of-both-worlds. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 2522–2563. PMLR, 2024. URL https://proceedings.mlr.press/v247/ito24a.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 Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 3666–3692. PMLR, 2026. URL https://proceedings.mlr.press/v336/ito26a.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://jmlr.org/papers/v17/kaufman16a.html.

T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. doi: 10.1016/0196-8858(85)90002-8. URL https://www.sciencedirect.com/science/article/pii/0196885885900028.

Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2111–2139. PMLR, 2019. URL https://proceedings.mlr.press/v99/lattimore19a.html.

Arnab Maiti, Kevin Jamieson, and Lillian J. 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 J. 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.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/hash/e1e32e235eee1f970470a3a6658dfdd5-Abstract.html.

Quan Nguyen, Shinji Ito, Junpei Komiyama, and Nishant A. Mehta. Data-dependent bounds with T-optimal best-of-both-worlds guarantees in multi-armed bandits using stability-penalty matching. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4386–4451. PMLR, 2025. URL https://proceedings.mlr.press/v291/nguyen25a.html.

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.

Nahum Shimkin and Adam Shwartz. Asymptotically efficient adaptive strategies in repeated games part I: Certainty equivalence strategies. *Mathematics of Operations Research*, 20(3):743–767, 1995. doi: 10.1287/moor.20.3.743. URL https://pubsonline.informs.org/doi/10.1287/moor.20.3.743.

Nahum Shimkin and Adam Shwartz. Asymptotically efficient adaptive strategies in repeated games part II. asymptotic optimality. *Mathematics of Operations Research*, 21(2):487–512, 1996. doi: 10.1287/moor.21.2.487. URL https://pubsonline.informs.org/doi/10.1287/moor.21.2.487.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1263–1291. PMLR, 2018. URL https://proceedings.mlr.press/v75/wei18a.html.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021. URL https://jmlr.org/papers/v22/19-753.html.

A CONTEXT AND MAIN RESULTS

The table records the information and policy assumptions used by each result.

Result	Policy inputs or assumptions	Scope
Theorem 1	Only m, n, T	All matrices; each applicable bound holds simultaneously.
Proposition 1	Logarithmic regret for each fixed A and $q > 0$	Matching fixed-instance lower bound for the revealed-action guarantee.
Theorem 2	Only m, n, T	General guarantees and the specified two-row subfamily.
Appendix D.3	π^+: only m, n, T; coefficient comparison: the policy guarantees in Appendix A.3	Matching fixed-instance lower and upper coefficients for the enhanced policy.
Appendix D.4	Known candidate family, parameters, q, T	Identical gap parameters with different minimax regret.

Table A.1: Information assumptions. All upper bounds allow the adaptive opponents and bounded conditionally unbiased noise of Section 2.

A.1 RELATED METHODS AND INFORMATION MODELS

Ito et al. (2026) establish logarithmic PSMR and pose the simultaneous-adaptation question. Their Theorems 1 and 4 give the first two rows of Table 1. In the latter, write $\Delta = v - A_{ij} > 0$ for a gap between the benchmark and a matrix entry below it. Then $0 < \Delta \leq 2$ and $\Delta \leq 4/\Delta$ allow the additive sum of gaps to be absorbed into C_c.

The hybrid regularizer specializes Nguyen et al. (2025, Section 5, Eq. (17)) to parameter $1/2$, building on stability–penalty matching (Ito et al., 2024) and Tsallis-INF (Zimmert & Seldin, 2021). Section C.4 proves the signed sparse-feedback bound used here. Shimkin & Shwartz (1995; 1996) study fully observed games from known finite families of reward laws. Part II gives exact logarithmic coefficients under mutual absolute continuity, unique optimal strategies, and distinct game values. Appendices A.3 and D.3 compare fixed-instance coefficients on two families with nonunique minimizing strategies, using the same policy class over all matrices and allowing masked actions and bounded conditionally unbiased noise. BL-WoLF also studies bounded learning loss for known game families (Conitzer & Sandholm, 2003). Here rewards remain visible when column labels are missing; the fixed-instance lower bounds account for both kinds of feedback and for payoffs above the safety value.

A.2 EXPLICIT CONSTANTS AND CONFIDENCE PARAMETERS

We give expected regret bounds with explicit constants, followed by the confidence-parameter form used in their proof. We use the model and notation of Section 2: regret is signed and the opponent chooses its column before the learner's fresh random draws. The auxiliary update draw is independent of reward and revelation indicator conditional on the precomputed distribution and actual row. At a strict saddle write $A_\alpha = \sum_{i \neq i_*} \alpha_i^{-1} = \beta C_s$.

Theorem A.1. The policy of Section 3, given only m, n, T, satisfies

$$\begin{aligned} R_T &\leq 200\sqrt{mTL_T} + 1 && \text{for every matrix and } q \in [0,1], \\ R_T &\leq 33001 L_T C_c/q && \text{for every matrix if } q > 0, \\ R_T &\leq 160000 L_T C_s + 1 && \text{at a strict saddle, for all } q \in [0,1], \end{aligned} \tag{A.1}$$

where $L_T = \log(192m^2(n+1)(T+1)^5)$. These bounds hold simultaneously with $R_T \leq 2T$. If $C_c = 0$, every mean-payoff deficit is nonpositive, so $R_T \leq 0$.

For the proof, choose any $0 < \delta < 1$ as the failure-probability parameter of the confidence intervals, and run the same policy with

$$L = \log \frac{32m^2(n+1)(T+1)^3}{\delta}. \tag{A.2}$$

The resulting expected regret satisfies

$$R_T \le 200\sqrt{mTL} + 3T\delta \qquad \text{for every matrix,} \tag{A.3}$$
$$R_T \le 160000LC_s + 6T\delta \qquad \text{at a strict saddle,} \tag{A.4}$$
$$qR_T \le 33000LC_c + 2T\delta \qquad \text{for every matrix and } q > 0. \tag{A.5}$$

Setting $\delta = 1/[6(T+1)^2]$ recovers the policy and logarithm in Theorem A.1. Appendix C proves these inequalities, using the learner analysis in Appendix C.4.

A.3 Fixed-instance lower bound for revealed actions

Here we give the precise form of Proposition 1. The hard matrices have one all-zero row. Every other row has a negative entry in column 1 and zero entries elsewhere. The lower bound applies to any policy with logarithmic worst-case regret on every fixed matrix at every fixed positive revelation probability. For comparisons of logarithmic coefficients, we also specify the general-matrix guarantees required of our common policy class. The definitions below serve both this result and Appendix D.3.

Fix $m, n \ge 2$ and $W_T = \sqrt{mT\log(emn(T+1))}$. A sequence $\pi = (\pi_T)_{T\ge1}$ receives only m, n, T and the observed history. Define

$$\Phi_\pi(A, q, T) = \left[\sup_{\text{allowed opponents and noises}} \mathbb{E}_\pi \sum_{t=1}^{T} (v(A) - A_{I_t J_t})\right]_+ . \tag{A.6}$$

The supremum covers all opponents and reward-noise processes allowed in Section 2. Let Π be the class of sequences satisfying (i) $\Phi_\pi(A, q, T) \le C_\pi W_T$ for all A, q, T with the same finite C_π; (ii) $\Phi_\pi(A, q, T) = O_{\pi,A,q}(\log(eT))$ for every fixed A and $q > 0$; and (iii) the same property holds at $q = 0$ for every strict pure saddle. The matrix and its safety value are learned from feedback.

Property (ii) is the requirement used by the lower bound: it applies both to the displayed instances and to the alternative matrices used in the proof. Conditions (i) and (iii) complete the common class in which we compare logarithmic coefficients. For fixed A and $q > 0$, define

$$c_\pi(A, q) = \limsup_{T\to\infty} \frac{\Phi_\pi(A, q, T)}{\log T}, \qquad c_*(A, q) = \inf_{\pi\in\Pi} c_\pi(A, q). \tag{A.7}$$

Take each sequence's limit superior before the infimum. This order compares the asymptotic coefficient of each complete policy sequence. Conditions (i)–(iii) allow changes at any single horizon, which leave that coefficient unchanged.

For probability laws P, Q, write $\mathrm{KL}(P\|Q)$ for their Kullback–Leibler divergence. For binary rewards with means $a, b \in (-1, 1)$, write $\mathrm{d}(a, b) = \mathrm{kl}((1+a)/2, (1+b)/2)$, where $\mathrm{kl}(p, s) = p\log(p/s) + (1-p)\log((1-p)/(1-s))$ is the binary relative entropy.

Proposition A.2 (matching lower bound for revealed actions). Fix $m, n \ge 2$, $q \in (0, 1]$, and $\alpha_i \in (0, 1/4]$ for $i = 2, \dots, m$. Let

$$A_{1j} = 0 \quad (j \in [n]), \qquad A_{i1} = -\alpha_i, \quad A_{ij} = 0 \quad (i \ge 2,\ j \ge 2). \tag{A.8}$$

This matrix has safety value zero, a pure saddle at $(1, 1)$ with tied column responses, and $C_c(A) = \sum_{i=2}^{m} \alpha_i^{-1}$. Every $\pi \in \Pi$ satisfies

$$\liminf_{T\to\infty} \frac{\Phi_\pi(A, q, T)}{\log T} \ge \frac{1}{q} \sum_{i=2}^{m} \frac{\alpha_i}{\mathrm{d}(-\alpha_i, 0)} \ge \frac{C_c(A)}{q}. \tag{A.9}$$

The policy π^0 of Theorem 1 belongs to Π, and

$$\frac{C_c(A)}{q} \le c_*(A, q) \le c_{\pi^0}(A, q) \le 5C\frac{C_c(A)}{q}, \tag{A.10}$$

where C is the universal constant in that theorem. All matrix parameters and q are fixed before taking the horizon limit. The lower bound uses property (ii) of Π and also holds for policies receiving q and satisfying property (ii).

Thus, for each fixed matrix in (A.8) and fixed $q > 0$, $c_*(A, q) = \Theta(C_c(A)/q)$ with universal comparison constants. Since $L_T = 5 \log T + O(1)$ at fixed dimensions, this family matches the time logarithm, revelation rate, and row-gap dependence of the second branch of Theorem 1.

Appendix B.1 gives the proof.

B THE BASE POLICY AND LOWER BOUNDS

B.1 PROOF OF THE FIXED-INSTANCE LOWER BOUND

We prove Proposition A.2 using the policy class and coefficient definitions in Appendix A.3. The opponent makes column 1 rare. Without its label, the negative entry is averaged with zero entries, so reward observations become less informative about that entry. We compare this matrix with one in which the same row is the unique maximin row. A policy with logarithmic regret on both matrices must distinguish them.

Proof. Fix $\pi \in \Pi$ and $\rho \in (0, 1)$. Let the opponent choose columns independently, with probabilities ρ and $1 - \rho$ on columns 1 and 2. At a cell with mean a, draw an independent binary reward with probabilities $(1+a)/2$ and $(1-a)/2$ on $+1$ and -1, respectively, and use the prescribed independent revelation revelation indicators. Write $N_i(T) = \sum_{t \le T} \mathbf{1}\{I_t = i\}$ and $R_T(A; \rho)$ for the regret in this environment. Under this column law, the rows have mean rewards zero and $-\rho\alpha_i$, so

$$R_T(A; \rho) = \sum_{i=2}^{m} \rho\alpha_i \mathbb{E}_A N_i(T) \ge 0, \qquad \mathbb{E}_A N_i(T) = O(\log T) \quad (i \ge 2). \tag{B.1}$$

The count bounds follow from property (ii); their constants may depend on the fixed parameters, including ρ.

The original environment requires the policy to play row i only $O(\log T)$ times. We now make low regret require frequent play of that row. For each $i \ge 2$ and fixed $\varepsilon \in (0, 1/4]$, form $A^{i,\varepsilon}$ by replacing every entry of row i with ε. Keep the same column law, binary reward construction, and revelation rate. The safety value becomes ε, and every other row has entries at most zero. Consequently,

$$R_T(A^{i,\varepsilon}; \rho) \ge \varepsilon \mathbb{E}_{A^{i,\varepsilon}}[T - N_i(T)], \qquad \mathbb{E}_{A^{i,\varepsilon}}[T - N_i(T)] = O(\log T). \tag{B.2}$$

Property (ii) applies to this fixed alternative matrix. Both environments have nonnegative regret increments computed from the mean payoffs. Since the safety value is unknown, the learner receives the same external inputs in the two experiments.

After a row-i play with the column label hidden, the reward has mean $-\rho\alpha_i$ in the original environment and ε in the alternative. When the column is revealed, the learner observes both the reward and its label. The observed revelation indicator distinguishes these two branches, whose divergences are therefore weighted by $1 - q$ and q. Thus the divergence per row-i play is

$$\begin{aligned} D_i^{\varepsilon}(\rho, q) = (1 - q)\mathrm{d}(-\rho\alpha_i, \varepsilon) \\ + q\big[\rho \mathrm{d}(-\alpha_i, \varepsilon) + (1 - \rho)\mathrm{d}(0, \varepsilon)\big]. \end{aligned} \tag{B.3}$$

Let P_T and Q_T be the laws of the learner's record under A and $A^{i,\varepsilon}$, respectively. The record comprises its private randomization and $(I_t, r_t, B_t, \widetilde{J}_t)_{t \le T}$, where $\widetilde{J}_t = J_t$ when $B_t = 1$ and $\widetilde{J}_t = \dagger$ when $B_t = 0$; $\dagger$ denotes a missing label. Action kernels coincide at the same observed history, and all other rows have the same feedback kernels. The KL chain rule gives

$$\mathrm{KL}(P_T \| Q_T) = D_i^{\varepsilon}(\rho, q) \mathbb{E}_A N_i(T). \tag{B.4}$$

Positive probabilities for both rewards, common column and revelation indicator supports, and identical private randomness laws ensure mutual absolute continuity.

For $E_i = \{N_i(T) \le T/2\}$, write $p_T = P_T(E_i)$ and $s_T = Q_T(E_i)$. Equations (B.1)–(B.2) and Markov's inequality give $p_T \ge 1 - O(\log T/T)$ and $s_T = O(\log T/T)$. Data processing and the binary entropy bound imply

$$\begin{aligned} D_i^{\varepsilon}(\rho, q) \mathbb{E}_A N_i(T) &\ge \mathrm{kl}(p_T, s_T) \\ &\ge p_T \log(1/s_T) - \log 2 \ge (1 - o(1)) \log T. \end{aligned} \tag{B.5}$$

First take the lower limit in T with ρ, ε, q, A fixed, then let $\varepsilon \downarrow 0$. Continuity of d and the lower-limit inequality for the finite sum of nonnegative terms in (B.1) yield

$$\liminf_{T\to\infty} \frac{R_T(A;\rho)}{\log T} \geq \sum_{i=2}^{m} \frac{\rho\alpha_i}{(1-q)\mathrm{d}(-\rho\alpha_i, 0) + q\rho\mathrm{d}(-\alpha_i, 0)}. \tag{B.6}$$

Since $\Phi_\pi(A, q, T) \geq R_T(A;\rho)$, the right-hand side bounds $\liminf_T \Phi_\pi(A, q, T)/\log T$ for every fixed $\rho > 0$. Letting $\rho \downarrow 0$ in these bounds and using $\mathrm{d}(-\rho\alpha_i, 0) = O(\rho^2\alpha_i^2)$ proves the first inequality in (A.9). This gives the lower bound on the worst-opponent coefficient using fixed positive choices of ρ.

For $f(u) = \mathrm{d}(u, 0)$, we have $f(0) = f'(0) = 0$ and $f''(u) = 1/(1-u^2) \leq 4/3$ on $|u| \leq 1/2$. Hence $\mathrm{d}(-\alpha_i, 0) \leq \alpha_i^2$, proving the second inequality in (A.9). The three guarantees of Theorem 1 imply $\pi^0 \in \Pi$. Its revelation bound and $L_T/\log T \to 5$ give the upper bound in (A.10); the remaining inequalities follow from (A.9) and the definition of c_*. □

Equation (B.6) explains the dependence on q. After $\varepsilon \downarrow 0$, the two feedback branches contribute $(1-q)\mathrm{d}(-\rho\alpha_i, 0)$ and $q\rho\mathrm{d}(-\alpha_i, 0)$ per row-i play, where $\mathrm{d}(-\rho\alpha_i, 0) = \Theta(\rho^2\alpha_i^2)$ and $\mathrm{d}(-\alpha_i, 0) = \Theta(\alpha_i^2)$. The expected regret cost is $\rho\alpha_i$. Taking small fixed ρ at fixed $q > 0$ gives a cost-to-information ratio of order $1/(q\alpha_i)$.

B.2 A SINGLE ALGORITHM FOR EVERY REVELATION PROBABILITY

Use the policy and notation of Section 3, with L from (A.2). Statistics use observations from completed rounds; ties are resolved by a fixed rule.

B.2.1 CELL AND COLUMN CONFIDENCE BOUNDS

Let N_{ij} count the completed revealed observations of cell (i, j), and let $\widehat{A}_{ij,k}$ be the mean of its first k rewards. Its upper bound is

$$U_{ij} = \min\left\{1, \min_{1\leq k\leq N_{ij}} \left(\widehat{A}_{ij,k} + \sqrt{2L/k}\right)\right\}, \tag{B.6a}$$

with value one before the first observation. The column lower bound (6) starts at $H_j = -1$. Write $\tau_{j,\ell}$ for column j's ℓth revealed round. Its canonical blocks are

$$D_{r,k} = \{r2^k + 1, \ldots, (r+1)2^k\}, \qquad r, k \geq 0 \text{ integers}. \tag{B.7}$$

Here $\widehat{r}_j(D) = |D|^{-1}\sum_{\ell\in D} r_{\tau_{j,\ell}}$, and $t \in D$ in calendar sums means $t \in \{\tau_{j,\ell} : \ell \in D\}$. Include every completed block's lower bound in the running maximum H_j, even after merging its records; U, H change only on revelations. Section B.3 gives the streaming implementation.

B.2.2 ROW UPPER BOUNDS AND ELIGIBILITY

Use M_i, w_t, b_t, E_t from Section 3 and (9), and the distribution z_t defined below, which assigns positive probability to every recommendation. The recommendation map is

$$f_t(i) = i \quad (i \in E_t), \qquad f_t(i) = b_t \quad (i \notin E_t). \tag{B.8}$$

Equation (7) defines $\mathcal{A}_t, s_t, Q_t, p_t$; thus $p_t(i) = z_{ti}/2$ on $\mathcal{A}_t$ and $p_t(b_t) \geq 1/2$. Append the estimate X_t from (8) to every triple (i, a, e) satisfying

$$b_t = i, \qquad \mathbf{1}\{a \in E_t\} = e. \tag{B.9}$$

For the K records $X_1, \ldots, X_K$ in each subsequence, including zero values and unrevealed rounds, define

$$P_{i,a,e} = \min\left\{1, \min_{1\leq k\leq K} \left(\frac{1}{k}\sum_{h=1}^{k} X_h + \sqrt{8L/k}\right)\right\}. \tag{B.10}$$

An empty subsequence has bound one. The selection conditions depend on completed records and are determined before the current reward and revelation indicator are observed.

B.2.3 A SPARSE, CENTERED LEARNER UPDATED ON EVERY REWARD

Small sampling probabilities can produce large importance-weighted loss estimates. The regularizer controls how one such estimate changes the minimizing distribution. We use a square-root term and an entropy term, together with the logarithmic barrier below. Let $D_0 = m(T+1)$, $\eta = 1/D_0$, and

$$F(x) = -2\sqrt{x} + (1-x)\log(1-x) + x, \qquad F(0) = 0, \quad F(1) = -1. \tag{B.11}$$

Let $G_t = \sum_{s \le t} u_s$, $G_0 = 0$, $\lambda_{1i} = 16$, and $\lambda_t = (\lambda_{ti})_{i \in [m]}$. With $G = G_{t-1}$ and $\lambda_i = \lambda_{ti}$, compute

$$x_t = \arg\min_{x \in \Delta_m} \left\{ \langle G, x \rangle + \sum_i \lambda_i F(x_i) - 64 \sum_i \log x_i \right\}, \qquad z_{ti} = \left(1 - \frac{1}{T+1}\right) x_{ti} + \eta. \tag{B.12}$$

Strict convexity and the barrier ensure a unique interior minimizer.

Set $\ell_t = (1 - r_t)/2$. If $I_t = i \neq b_t$, use a vector with its only possibly nonzero coordinate

$$u_{ti} = \ell_t / p_t(i). \tag{B.13}$$

If $I_t = b_t$ and $s_t > 0$, draw a fresh private recommendation $K \in \mathcal{A}_t$, independently of reward and revelation indicator, with probability $\nu_i = z_{ti}/s_t$, and use

$$u_{tK} = -\ell_t / [p_t(b_t)\nu_K]. \tag{B.14}$$

All other coordinates are zero; use $u_t = 0$ if $s_t = 0$.

Add u_t to G. We also increase the regularizer coefficient for the updated coordinate. The rule below makes the contribution from changing this coefficient comparable to the contribution from the loss update in the regret bound. For the possibly nonzero coordinate j, put

$$\begin{aligned} a_{tj} &= 8u_{tj}^2 \min\{z_{tj}^{3/2},\, 1 - z_{tj},\, \lambda_j z_{tj}^2/64\}, \\ h_{tj} &= 2\sqrt{z_{tj}}, \qquad \lambda_j \leftarrow \lambda_j + \frac{a_{tj}}{\lambda_j h_{tj}}. \end{aligned} \tag{B.15}$$

The right-hand side uses pre-update values. The three expressions in the minimum correspond to the square-root, entropy, and logarithmic-barrier parts of the regularizer. In the proof of Lemma C.2, each controls the loss-update term in the cases where it can be the minimum, giving the bound a_{tj}/λ_{tj}; the coefficient change contributes at most three times this quantity. For every other coordinate set $a_{ti} = 0$ and $h_{ti} = 2\sqrt{z_{ti}}$. Every reward updates the learner; a zero update leaves its coefficients unchanged.

B.3 EXACT STREAMING IMPLEMENTATION

Arithmetic operations, comparisons, square roots, and logarithms each have unit cost; the T exact convex optimization problems in (B.12) are counted separately. Each cell or observation group stores only a count, sum of recorded values, and historical bound. When a row first becomes the leader, allocate its $2m$ groups indexed by the comparison row a and state $e \in \{0, 1\}$, and cache the minimum of their bounds. Each round updates one group for each a, according to its current state, for m group updates. At most T rows require allocation, so these operations cost $O(mT)$.

For each revealed column, maintain the binary decomposition of the observations collected so far. Insert a singleton and merge equal-length aligned terminal blocks, updating H_j with each new lower bound before discarding children. Each canonical block appears at its right endpoint: the first child survives until the second completes. An insertion creates at most $1 + \lfloor \log_2 T \rfloor$ blocks; r visits create

$$\sum_{k \ge 0} \lfloor r/2^k \rfloor < 2r. \tag{B.16}$$

Thus block updates cost $O(T)$. Eligibility scans cost $O(mn)$ per round; leader selection and sampling cost $O(m)$. The cost excluding convex optimization is $O(T(mn + \log(T+1)))$ under a per-round worst-case bound, and

$$O(mnT) \quad \text{amortized arithmetic operations in total.} \tag{B.17}$$

The subsequences, cells, blocks, and learner use $O(m^2)$, $O(mn)$, $O(n \log(T+1))$, and $O(m)$ words, respectively, totaling

$$O\big(m^2 + mn + n\log(T+1)\big) \quad \text{arithmetic words.} \tag{B.18}$$

C Regret analysis

C.1 Confidence on the two observation streams

Write

$$\mu_t = \sum_i p_t(i) A_{iJ_t}, \qquad d_t = v - \mu_t. \tag{C.1}$$

Write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}, J_t]$ for the conditional expectation before the fresh row and auxiliary draws. Then p_t is fixed and $\mu_t = \mathbb{E}_t r_t$.

Lemma C.1. With probability at least $1 - \delta$, simultaneously through T, the following hold. Each bound holds for every available sample count $k \geq 1$, or for every completed block of size k.

1. The empirical mean of the first k revealed observations of each cell differs from its cell mean by at most $\sqrt{2L/k}$.
2. The empirical mean of every completed canonical block in a revealed-column sequence differs from the average of its corresponding μ_t's by at most $\sqrt{2L/k}$.
3. The empirical mean of the first k records of each selected leader-observation subsequence differs from the average of its corresponding $A_{b_t J_t}$'s by at most $\sqrt{8L/k}$.

Proof. Independence allows the revelation indicator to be drawn privately before the row and reward. Include all intervening history in each selected-sample filtration. Conditioning on the column, row, and revelation indicator gives cell mean A_{ij}; conditioning on the column and revelation indicator before the row draw gives column mean μ_t. The predictable selection condition for X_t preserves its mean $A_{b_t J_t}$. The lengths of the three ranges are $2, 2, 4$, respectively.

For a centered increment Y with range length c, Hoeffding's lemma gives $\mathbb{E}[e^{\theta Y} \mid \text{past}] \leq e^{\theta^2 c^2/8}$. The resulting exponential supermartingale, started after a fixed sample index and stopped at the next k samples or at T, bounds the failure probability by $2e^{-L}$ for each stated radius whenever those samples exist. There are at most mnT pairs of cells and sample counts, $2m^2T$ pairs of selected subsequences and record counts, and $2nT$ canonical blocks. The union bound and (A.2) give probability at least $1 - \delta$, also covering sample counts and blocks selected later. □

Call this event $\mathcal{G}$. The case $C_c = 0$ was settled in Appendix A.2; henceforth $C_c > 0$ and therefore $v > -1$.

On $\mathcal{G}$,

$$U_{ij} \geq A_{ij}, \qquad U_{ij} \leq A_{ij} + 2\sqrt{2L/N_{ij}} \quad (N_{ij} > 0). \tag{C.2}$$

For a maximin row a, every cell mean A_{aj} is at least v, so all cell bounds U_{aj} are at least v. Each reward record collected while a is the leader also has conditional mean at least v. Lemma C.1 therefore gives $P_{a,h,e} \geq v$ for every comparison row h and state e. Taking the minimum of these cell and group bounds gives $M_a \geq v$. Since $v > -1$, lower clipping is inactive for the chosen leader. Every bound entering $M_{b_t} = w_t$ is therefore at least w_t, giving

$$w_t \geq v, \qquad U_{b_t j} \geq w_t \quad \text{for all } j. \tag{C.3}$$

Also $H_j \leq \max_i A_{ij}$, because its block lower bounds are at most the corresponding average conditional means. If a strict pure saddle exists, this yields $H_{j_*} \leq v$, and the eligibility rule then implies

$$w_t < v + \beta \quad \Longrightarrow \quad i_* \in E_t. \tag{C.4}$$

Indeed the off-saddle cell bounds of row i_* are strictly above w_t; its only possible low bound is at j_*, where it is at least $v \geq H_{j_*}$. These assertions also hold at $q = 0$, when the bounds based on revealed observations stay at their initial values.

Finally, for any completed canonical block D of column j, (6) and Lemma C.1 give, at every subsequent time,

$$H_j \geq \frac{1}{|D|} \sum_{t \in D} \mu_t - 2\sqrt{2L/|D|}. \tag{C.5}$$

The running maximum H_j preserves (C.5) after the block records are merged.

C.2 LEADER CONTROL USING ALL REWARDS

For a set of rounds I, define the cumulative leader deficit

$$R_b(I) = \sum_{t \in I} (v - A_{b_t J_t}). \tag{C.6}$$

The following bound holds on every matrix.

C.2.1 A BOUND FOR A COMPLETED OBSERVATION SUBSEQUENCE

On $\mathcal{G}$, consider a selected subsequence with leader i and $K \geq 2$ records. Just before its last assignment, its upper bound computed from the preceding $K-1$ records is at least $w_t \geq v$. Lemma C.1 bounds the deficit on the first $K-1$ records by $4\sqrt{2L(K-1)}$. The last deficit is at most two, giving, also for $K = 0, 1$,

$$r_i := \sum_{\text{subsequence entries}} (v - A_{iJ_t}) \leq 4\sqrt{2LK} + 2. \tag{C.7}$$

Each assignment in the selected subsequence contributes one record, including zero values and unrevealed rounds.

C.2.2 THE TRUE SADDLE ROW IS INELIGIBLE

For the rest of this section, assume a strict saddle. Put $\mathcal{I}_0 = \{t : i_* \notin E_t\}$, $\mathcal{I}_1 = [T] \setminus \mathcal{I}_0$, and $T_{\text{off}}(I) = \sum_{t \in I} \mathbf{1}\{J_t \neq j_*\}$.

By (C.4), every record indexed by $(i, i_*, 0)$, on which the saddle row fails the exploration test, has $w_t \geq v + \beta$. Applying the same argument to the earlier records with this larger threshold gives

$$r_i \leq 4\sqrt{2LK} - \beta K + 4 \leq 16L/\beta - \beta K/2 + 4. \tag{C.8}$$

This includes $K = 0$; the last step maximizes $4\sqrt{2LK} - \beta K/2$.

Since $\alpha_i, \beta \leq 2$,

$$m/\beta \leq 4C_s, \qquad m \leq 8C_s. \tag{C.9}$$

Summing (C.8) gives $R_b(\mathcal{I}_0) \leq 96LC_s - \beta|\mathcal{I}_0|/2$ on $\mathcal{G}$. On its complement, $R_b \leq 2T$ and restoring the negative count costs at most $\beta T\delta/2 \leq T\delta$. Hence

$$\mathbb{E}R_b(\mathcal{I}_0) \leq 96LC_s - \tfrac{\beta}{2}\mathbb{E}|\mathcal{I}_0| + 3T\delta. \tag{C.10}$$

C.2.3 THE TRUE SADDLE ROW IS ELIGIBLE

For $i \neq i_*$, consider the records indexed by $(i, i_*, 1)$, on which the saddle row passes the exploration test. Let K_i^* count entries with $J_t = j_*$, let O_i count its other entries, put $K_i = K_i^* + O_i$, and denote its total true deficit by r_i. Pathwise, including outside $\mathcal{G}$,

$$r_i \geq \alpha_i K_i^* - 2O_i, \qquad K_i \leq \frac{(r_i)_+}{\alpha_i} + \frac{4O_i}{\alpha_i}. \tag{C.11}$$

Set $c_i = 4\sqrt{2L/\alpha_i}$. Equations (C.7), (C.11) give $r_i \leq c_i\sqrt{(r_i)_+} + 2c_i\sqrt{O_i} + 2$ on $\mathcal{G}$. For $r_i > 0$, absorb $c_i\sqrt{r_i} \leq r_i/2 + c_i^2/2$, obtaining $r_i \leq 32L/\alpha_i + 16\sqrt{2LO_i/\alpha_i} + 4$; for $r_i \leq 0$ this bound is immediate. Young's inequality gives, for $\epsilon > 0$,

$$r_i \leq \frac{32L}{\alpha_i} + \frac{128L}{\alpha_i \epsilon} + 4 + \epsilon O_i. \tag{C.12}$$

With $\epsilon = \beta/8$, summing the first three terms gives at most $(64+1024+16)LC_s$, using $m-1 \leq 4C_s$. Leader i_* has nonpositive deficit, so $R_b(\mathcal{I}_1) \leq 1104LC_s + \beta T_{\text{off}}(\mathcal{I}_1)/8$ on $\mathcal{G}$. The nonnegativity of the count and the bound $R_b \leq 2T$ on the complement give

$$\mathbb{E}R_b(\mathcal{I}_1) \leq 1104LC_s + \tfrac{\beta}{8}\mathbb{E}T_{\text{off}}(\mathcal{I}_1) + 2T\delta. \tag{C.13}$$

Both (C.10) and (C.13) hold for every q, including zero.

C.3 Centered learning and actual action counts

Lemma C.2 (sparse centered learner). For the learner defined by (B.11)–(B.15), with arbitrary predictable leaders and eligible sets, every fixed recommendation a satisfies

$$\mathbb{E}\sum_{t=1}^{T}\big(\langle z_t, u_t\rangle - u_{ta}\big) \le K_0 \sum_i \sqrt{\mathbb{E}V_i} + D_1, \tag{C.14}$$

where

$$V_i = \sum_t \mathbf{1}\{i \in \mathcal{A}_t\}\min\{z_{ti}, 1 - z_{ti}\}, \quad K_0 = 80\sqrt{\log D_0}, \quad D_1 = 100m\log D_0.$$

Section C.4 gives the proof.

Write $\bar{\ell}_{ij} = (1 - A_{ij})/2$. Conditional on the completed history and J_t, before the row is drawn, an active coordinate satisfies

$$\mathbb{E}_t u_{ti} = \bar{\ell}_{iJ_t} - \bar{\ell}_{b_t J_t}, \qquad |u_{ti}| \le 2/z_{ti}, \qquad \mathbb{E}_t u_{ti}^2 \le 4/z_{ti}; \tag{C.15}$$

an inactive coordinate is zero. Thus each expected update equals the loss difference between the mapped recommendation and the leader, using every round's reward.

We next express the right-hand side of (C.14) through actual action counts. Fix any recommendation index a, and let $N_i^{\mathrm{play}} = \sum_t \mathbf{1}\{I_t = i\}$, including plays of row i when it is the leader. The comparator follows $f_t(a)$, which may be row a or the current leader. For $i \neq a$, $\mathbb{E}V_i \le 2\mathbb{E}N_i^{\mathrm{play}}$, since an active coordinate is played with probability $z_{ti}/2$. When a is active, the leader is a different row and is played with probability at least one half. The contribution of that round to V_a is at most one, so it is bounded by twice the conditional probability of playing a row other than a. Consequently,

$$\mathbb{E}V_a \le 2\sum_{i\neq a}\mathbb{E}N_i^{\mathrm{play}}, \qquad \sum_i \sqrt{\mathbb{E}V_i} \le 2\sqrt{2}\sum_{i\neq a}\sqrt{\mathbb{E}N_i^{\mathrm{play}}}. \tag{C.16}$$

This comparison holds on every matrix. For the remainder of the section, assume a strict pure saddle and take $a = i_*$.

Define $S = \sum_{i\neq i_*}\alpha_i \mathbb{E}N_i^{\mathrm{play}}$. Apply Lemma C.2 to recommendation i_*, use (C.15), and multiply by two to convert normalized losses to rewards. Weighted Cauchy–Schwarz and (C.16) give the unconditional comparison

$$\mathbb{E}\sum_t \left(A_{f_t(i_*)J_t} - \sum_a Q_t(a)A_{aJ_t}\right) \le \mathcal{B} := 4\sqrt{2}K_0\sqrt{A_\alpha S} + 2D_1. \tag{C.17}$$

C.4 Complete proof of Lemma C.2

We compare the learner with a fixed recommendation a. Each round changes both the cumulative estimated loss and a regularizer coefficient. We first bound the cumulative loss difference using the contributions from these two changes, then estimate each contribution.

Allow arbitrary leaders and eligible sets determined before the current row draw. Write $\Lambda = \log D_0$, $\gamma = 64$, and

$$\Psi_t(x) = \sum_i \lambda_{ti}F(x_i) - \gamma\sum_i \log x_i.$$

Use the conditional expectation $\mathbb{E}_t$ from Section C.1. Minimize over Δ_m. Define the minimum objective value by $\Phi_t(G) = \min_x\{\langle G, x\rangle + \Psi_t(x)\}$. For the final telescoping sum, define x_{T+1} using G_T, λ_{T+1}. We compare with $\bar{e}_a = (1 - m\eta)e_a + \eta\mathbf{1}$, whose positive coordinates keep the logarithmic barrier finite.

For analysis, split the update into two steps: first change G_{t-1} to G_t while fixing λ_t, then change λ_t to λ_{t+1}. Write

$$\mathcal{S}_t = \langle x_t, u_t\rangle + \Phi_t(G_{t-1}) - \Phi_t(G_t)$$

for the difference between the loss at x_t and the change in the minimum objective in the first step. In the second step, $\Phi_t(G_t) - \Phi_{t+1}(G_t) \le (\Psi_t - \Psi_{t+1})(x_{t+1})$. Summing these differences gives

$$\sum_t \langle x_t - \bar{e}_a, u_t \rangle \le \Psi_{T+1}(\bar{e}_a) - \min_x \Psi_1(x) + \sum_t \mathcal{S}_t + \sum_t (\Psi_t - \Psi_{t+1})(x_{t+1}).$$

The endpoint terms are bounded directly below. We bound the two sums by one and three times $\sum_{t,i} a_{ti}/\lambda_{ti}$, respectively. The adaptive coefficient rule and the second moments of u_t then control the expectation of this common sum.

Moments of the loss estimate For active i, (B.13)–(B.14) contribute $\bar{\ell}_{iJ_t}$ and $-\bar{\ell}_{b_t J_t}$ to the mean, and their second moments sum to at most

$$\frac{1}{p_t(i)} + \frac{1}{p_t(b_t)\nu_i} = \frac{2}{z_{ti}} + \frac{s_t}{p_t(b_t) z_{ti}} \le \frac{4}{z_{ti}}.$$

Both possible magnitudes are at most $2/z_{ti}$. This proves (C.15). The mean update vector belongs to $[-1, 1]^m$.

Fixed-coefficient stability Abbreviate $x = x_t, z = z_t, \lambda = \lambda_t$, and let the update be ue_j. Let y be the minimizer after adding this update while keeping λ fixed. Smoothing gives

$$x_j \le 2z_j, \qquad 1 - x_j \le 2(1 - z_j). \tag{C.18}$$

Monotonicity of the coordinate gradients and the Karush–Kuhn–Tucker (KKT) equations give $y_j \le x_j$ for $u \ge 0$. For $u < 0$, $y_j \ge x_j$, all other coordinates decrease, and multiplier comparison gives $|u| \ge \gamma(1/x_j - 1/y_j)$. Since $|u| \le 2/z_j \le 4/x_j$, both cases yield

$$y_j \le 2x_j \le 4z_j. \tag{C.19}$$

Write $D_{\Psi_t}(y, x) = \Psi_t(y) - \Psi_t(x) - \langle \nabla \Psi_t(x), y - x \rangle$ for the Bregman divergence. With G denoting the pre-update cumulative vector, the one-step stability is

$$\mathcal{S}_t = \langle x, ue_j \rangle + \Phi_t(G) - \Phi_t(G + ue_j) = u(x_j - y_j) - D_{\Psi_t}(y, x). \tag{C.20}$$

The equality follows from the KKT condition at x.

The square-root, entropy, and logarithmic-barrier terms provide the first, second, and third expressions in the minimum in (B.15). We derive their bounds separately. Coordinate divergences are nonnegative. By (C.19), the square-root curvature on the segment is at least $\lambda_j/(16 z_j^{3/2})$; maximizing the resulting quadratic gives

$$\mathcal{S}_t \le 8u^2 z_j^{3/2}/\lambda_j. \tag{C.21}$$

The barrier curvature is at least $\gamma/(16 z_j^2)$, giving

$$\mathcal{S}_t \le 8u^2 z_j^2/\gamma. \tag{C.22}$$

For $F_2(x) = (1 - x)\log(1 - x) + x$, set $\xi = 1 - x_j$, $\zeta = 1 - y_j$. Maximizing $u(\zeta - \xi) - \lambda_j[\zeta \log(\zeta/\xi) - \zeta + \xi]$ over $\zeta \ge 0$ gives

$$\mathcal{S}_t \le \lambda_j(1 - x_j)(e^{u/\lambda_j} - 1 - u/\lambda_j).$$

If $u/\lambda_j \le 1$, use $e^r - 1 - r \le r^2$ for $r \le 1$, and (C.18), to bound this by $2u^2(1 - z_j)/\lambda_j$. Otherwise, $u > \lambda_j \ge 16$ and $|u| \le 2/z_j$ imply $z_j < 1/8$; then $z_j^{3/2} < 1 - z_j$, so the middle branch in (B.15) cannot be the smallest one. Equations (C.21)–(C.22) therefore prove in all cases

$$\mathcal{S}_t \le a_{tj}/\lambda_{tj}. \tag{C.23}$$

The penalty from changing a coefficient Hold the cumulative loss fixed and increase only λ_j by $d \ge 0$. Let x, x' be the old and new minimizers. We first show

$$d \le \gamma/(2\sqrt{x_j}) \quad \Longrightarrow \quad x'_j \le 2x_j. \tag{C.24}$$

For $x'_j > x_j$, the other coordinates decrease. With the decreasing function $g(x) = x^{-1/2} + \log(1 - x)$, the KKT comparison gives

$$Z = \lambda_j(g(x_j) - g(x'_j)) - dg(x'_j) + \gamma(1/x_j - 1/x'_j) \le 0.$$

The inequality forces $g(x'_j) > 0$, so $\gamma(1/x_j - 1/x'_j) \le dg(x'_j) \le d/\sqrt{x_j} \le \gamma/(2x_j)$, proving (C.24).

For the actual update (B.15), the barrier branch and $|u| \le 2/z_j$ give

$$d = \frac{a_{tj}}{\lambda_j h_{tj}} \le \frac{16}{\gamma\sqrt{z_j}}. \tag{C.25}$$

Since the post-loss minimizer satisfies $y_j \le 4z_j$ and $\gamma = 64$, the right side of (C.25) is at most $\gamma/(2\sqrt{y_j})$. Applying (C.24) after the loss update gives $x_{t+1,j} \le 2y_j \le 8z_j$.

As $F_2(x) \ge 0$, we have $-F(x) \le 2\sqrt{x}$. Consequently the time-variation penalty satisfies

$$(\Psi_t - \Psi_{t+1})(x_{t+1}) = -dF(x_{t+1,j}) \le 2d\sqrt{8z_j} \le 3dh_{tj} = 3a_{tj}/\lambda_{tj}. \tag{C.26}$$

Zero updates have zero stability and zero penalty.

Summing the adaptive terms Fix a coordinate, suppress its index, and set $a_t = 0$ when not updated. The barrier branch gives $0 \le a_t/\lambda_t \le 1/2$ and $h_t \ge 2\sqrt{\eta}$, so each coefficient increment is at most $\sqrt{D_0}/4$ and

$$\lambda_{T+1}/16 \le 1 + T\sqrt{D_0}/64 \le D_0^2. \tag{C.27}$$

Define the coefficient growth ratio $\rho_t = a_t/(\lambda_t^2 h_t)$, so $\lambda_{t+1} = \lambda_t(1 + \rho_t)$. On rounds with $\rho_t \le 1$, Cauchy–Schwarz and $\rho_t \le 2\log(1 + \rho_t)$ give

$$\sum_{\rho_t \le 1} \frac{a_t}{\lambda_t} \le \sqrt{\left(\sum_t h_t a_t\right)\left(\sum_{\rho_t \le 1} \rho_t\right)} \le 2\sqrt{\Lambda \sum_t h_t a_t}.$$

Every remaining round more than doubles its coefficient and contributes at most one half. Equation (C.27) bounds their total contribution by $\Lambda/\log 2$. Therefore, deterministically,

$$\sum_t a_t/\lambda_t \le 2\sqrt{\Lambda \sum_t h_t a_t} + \Lambda/\log 2. \tag{C.28}$$

By (C.15) and (B.15),

$$\begin{aligned} \mathbb{E}_t[h_{ti}a_{ti}] &\le 64\mathbf{1}\{i \in \mathcal{A}_t\} \min\{z_{ti}, z_{ti}^{-1/2}(1 - z_{ti})\} \\ &\le 96\mathbf{1}\{i \in \mathcal{A}_t\} \min\{z_{ti}, 1 - z_{ti}\}. \end{aligned} \tag{C.29}$$

For the last inequality use the first branch for $z_{ti} \le 1/2$ and the second otherwise. Sum (C.28) over coordinates, then use (C.29) and Jensen's inequality:

$$\mathbb{E}\sum_{t,i} a_{ti}/\lambda_{ti} \le 2\sqrt{96\Lambda}\sum_i \sqrt{\mathbb{E}V_i} + m\Lambda/\log 2. \tag{C.30}$$

Summing the potential differences We now bound the endpoint terms in the decomposition above. Since $F \le 0$, $\Psi_{T+1}(\bar{e}_a) \le 64m\Lambda$. Also $-\min_x \Psi_1(x) \le 32\sqrt{m}$, since $F(x) \ge -2\sqrt{x}$, the barrier is nonnegative, and $\sum_i \sqrt{x_i} \le \sqrt{m}$.

Insert these endpoint bounds, (C.23), and (C.26) into that decomposition, including the final coefficient update. The two per-round bounds contribute $1 + 3 = 4$ times the adaptive sum, giving

$$\sum_t \langle x_t - \bar{e}_a, u_t\rangle \le 64m\Lambda + 32\sqrt{m} + 4\sum_{t,i} a_{ti}/\lambda_{ti}. \tag{C.31}$$

Both $\|z_t - x_t\|_1$ and $\|\bar{e}_a - e_a\|_1$ are at most $2m\eta$. The mean update lies in $[-1, 1]^m$, so replacing $x_t, \bar{e}_a$ by z_t, e_a costs at most $4m\eta T \le 4$ in expectation. Substituting (C.30) into (C.31) proves (C.14): $8\sqrt{96} < 80$, and $(64 + 4/\log 2)m\Lambda + 32\sqrt{m} + 4 \le 100m\Lambda$ for $m \ge 2$, $D_0 \ge 4$.

All quantities are integrable: $z_{ti} \ge \eta$, $|u_{ti}| \le 2D_0$, and (C.27) bounds every adaptive coefficient. The argument takes unconditional expectations and holds for every $q \in [0, 1]$. □

C.5 A REVELATION-INDEPENDENT GUARANTEE ON EVERY MATRIX

Fix a pure maximin row a, so $A_{aj} \geq v$. As above, assume $C_c > 0$. Let I_e be the rounds on which a's eligibility state is e. The internal learner compares the payoff of the mapped recommendation $f_t(a)$ with the mean payoff under Q_t; write their cumulative difference as

$$C_T = \sum_t \left(A_{f_t(a)J_t} - \sum_i Q_t(i) A_{iJ_t} \right).$$

On I_0, the map sends recommendation a to the leader. On I_1, it recommends a, whose payoff is at least v. In the half-and-half mixture, the leader therefore contributes a full coefficient on I_0 and one half on I_1:

$$\sum_t \left(v - \sum_i p_t(i) A_{iJ_t} \right) \leq R_b(I_0) + \tfrac{1}{2} R_b(I_1) + \tfrac{1}{2} C_T. \tag{C.32}$$

We bound the leader terms first, then the comparison term.

Put $T_e = |I_e|$, and let $K_{i,e}$ count assignments of leader i on I_e. Their totals satisfy $\sum_i K_{i,e} = T_e$ and $T_0 + T_1 = T$. Applying (C.7) to the groups with comparison row a and using Cauchy–Schwarz gives, on $\mathcal{G}$,

$$\begin{aligned} R_b(I_0) + \tfrac{1}{2} R_b(I_1) &\leq 4\sqrt{2L} \sum_i \left(\sqrt{K_{i,0}} + \tfrac{1}{2} \sqrt{K_{i,1}} \right) + 3m \\ &\leq 4\sqrt{2mL} \left(\sqrt{T_0} + \tfrac{1}{2} \sqrt{T_1} \right) + 3m \\ &\leq \sqrt{40mTL} + 3m. \end{aligned} \tag{C.33}$$

On every path the same weighted leader deficit is at most $2T_0 + T_1 \leq 2T$. Removing the confidence event therefore costs at most $2T\delta$, and we conservatively allow $3T\delta$ below.

The sum of the quantities V_i in Lemma C.2 is at most the total recommendation probability assigned to active alternatives over all rounds:

$$\sum_i V_i = \sum_t \sum_{i \in \mathcal{A}_t} \min\{z_{ti}, 1 - z_{ti}\} \leq \sum_t \sum_{i \in \mathcal{A}_t} z_{ti} \leq T. \tag{C.34}$$

Consequently $\sum_i \sqrt{\mathbb{E}V_i} \leq \sqrt{mT}$. The conditional update identity (C.15), which holds for every mapped recommendation, converts Lemma C.2 to

$$\mathbb{E}C_T \leq 2K_0 \sqrt{mT} + 2D_1. \tag{C.35}$$

This comparison holds unconditionally. Take expectations in (C.32), use (C.33)–(C.35), and use the unconditional tower property for the actual row draw. Since $\log D_0 \leq L$ and $L \geq 1$,

$$\begin{aligned} R_T &\leq \sqrt{40mTL} + 3m + K_0 \sqrt{mT} + D_1 + 3T\delta \\ &\leq (\sqrt{40} + 80) \sqrt{mTL} + 103mL + 3T\delta. \end{aligned} \tag{C.36}$$

If $T \geq mL$, then $mL \leq \sqrt{mTL}$, and the leading constant becomes less than 190. If $T < mL$, use the direct bound $R_T \leq 2T \leq 2\sqrt{mTL}$. These two cases prove (A.3) with constant 200.

C.6 REGRET FROM REVEALED MATRIX ENTRIES

For a cell with mean payoff below v, enough revealed observations place its upper bound below v on the confidence event. We first bound its contribution before this observation threshold. After that point, the cell can still be explored while the lower bound for its column remains low. We bound the number of further revealed observations by comparing the rewards in a completed block with that column lower bound. The following fact ensures that a block covers a constant fraction of any interval of revealed visits.

Any N consecutive visit indices contain a canonical dyadic block of length at least $N/4$. For $N = 1$, take its singleton. Otherwise choose the largest power $h \leq N/2$. The first aligned block of length h starts within $h - 1$ positions of the left endpoint and fits inside the interval; also $h > N/4$. Lemma C.1 covers this block even when the interval is selected later.

Write $\Delta_{ij} = v - A_{ij}$. A cell with mean payoff below v has $\Delta_{ij} > 0$; set $k_{ij} = \lceil 512L/\Delta_{ij}^2 \rceil$ there and $k_{ij} = +\infty$ otherwise. It is unsaturated when its revealed count $N_{ij}(t-1) < k_{ij}$. At round t, sum the positive deficits from these cells, weighted by their play probabilities:

$$u_t^{\mathrm{c}} = \sum_{i:\Delta_{iJ_t}>0} p_t(i)\Delta_{iJ_t}\mathbf{1}\{N_{iJ_t}(t-1) < k_{iJ_t}\} \geq 0. \tag{C.37}$$

The tower property, applied without conditioning on $\mathcal{G}$, gives

$$\begin{aligned}\mathbb{E}\sum_t B_t u_t^{\mathrm{c}} &= \mathbb{E}\sum_t B_t \Delta_{I_tJ_t}\mathbf{1}\{\Delta_{I_tJ_t} > 0,\ N_{I_tJ_t}(t-1) < k_{I_tJ_t}\} \\ &\leq \sum_{\Delta_{ij}>0} \Delta_{ij}k_{ij} \leq (512L+4)C_c.\end{aligned} \tag{C.38}$$

Each cell contributes at most k_{ij} revealed visits before saturation.

Work pathwise on $\mathcal{G}$. For a saturated cell with mean payoff below v,

$$U_{ij} \leq v - 7\Delta_{ij}/8 < v \leq w_t. \tag{C.39}$$

By (C.3), this cell cannot lie in the leader row. If its row has positive probability as an alternative, eligibility requires $H_j \leq U_{ij}$. Writing $\rho_j = (v - H_j)_+$ for the amount by which the column lower bound lies below v, this implies $\Delta_{ij} \leq 8\rho_j/7$. The total probability of rows other than the leader is at most one half. Omitting nonpositive deficits gives the following upper bound on every round:

$$d_t \leq u_t^{\mathrm{c}} + \tfrac{4}{7}\rho_{J_t}(t-1). \tag{C.40}$$

For a column containing a cell with mean payoff below v, put $\Delta_j = \min_{i:A_{ij}<v}\Delta_{ij}$ and $\varepsilon_j = 7\Delta_j/8$. If $\rho_j < \varepsilon_j$ before the round, (C.39) excludes every saturated cell with mean payoff below v, so $d_t \leq u_t^{\mathrm{c}}$ when this column is chosen. The process ρ_j is nonincreasing, in $[0,2]$, and changes only upon revelation of this column.

While $H_j < v$, we count observations over intervals in which the positive difference $v - H_j$ stays within a factor of $3/2$. Partition the positive range of ρ_j into

$$(a_k, 3a_k/2], \qquad a_k = \tfrac{4}{3}(2/3)^k, \quad k = 0, 1, \ldots. \tag{C.41}$$

Analyze all bands with $3a_k/2 \geq \varepsilon_j$. In the revealed-column sequence, observations with ρ_j in one such band before the round form an interval. Its last observation may move ρ_j out of the band, so set it aside before comparing completed blocks and add its deficit separately below. Let N be the remaining length and U the sum of u_t^{c} on that interior.

If $N > 0$, we still have $\rho_j > a_k$ before the removed observation. Since unrevealed rounds leave H_j unchanged, at the interior endpoint $H_j \leq v - a_k$. Choose a contained canonical block D of length $h \geq N/4$. It is complete, so (C.5) gives

$$\sum_{t\in D}\mu_t \leq h(v - a_k) + 2\sqrt{2Lh}.$$

Every entry in D lies in the same band. Equation (C.40) gives the reverse estimate $\sum_{t\in D}\mu_t \geq h(v - 6a_k/7) - \sum_{t\in D}u_t^{\mathrm{c}}$. These estimates constrain the same block average. The low column bound gives an upper bound of $v - a_k + 2\sqrt{2L/h}$; the exploration test gives a lower bound of $v - 6a_k/7$ minus the average of u_t^{c}. The difference $a_k/7$ must therefore be accounted for by these deficits or by the confidence radius. Since $u_t^{\mathrm{c}} \geq 0$, its sum over the block is at most U, the sum over the entire interior. Consequently,

$$\frac{a_k}{7}h \leq U + 2\sqrt{2Lh}. \tag{C.42}$$

Using $2\sqrt{2Lh} \leq a_k h/14 + 28L/a_k$ shows

$$h \leq 14U/a_k + 392L/a_k^2, \qquad N \leq 4h \leq 56U/a_k + 1568L/a_k^2.$$

Summing (C.40) over all interior visits gives a total deficit of at most

$$U + \tfrac{6}{7}a_k N \leq 49U + 1344L/a_k. \tag{C.43}$$

The term $49U$ includes both the direct deficits from cells below their observation thresholds and the additional deficits while those cells keep the column average low. Equation (C.38) already bounds the expected sum of these U terms. This bound also holds when $N = 0$. The single removed observation contributes at most two to the deficit. The argument also covers a band ending at the horizon.

The smallest lower endpoint among the analyzed bands is at least $2\varepsilon_j/3$, and the reciprocal endpoints form a geometric progression. Therefore

$$\sum_{k:\, 3a_k/2 \geq \varepsilon_j} a_k^{-1} \leq \frac{9}{2\varepsilon_j} = \frac{36}{7\Delta_j}. \tag{C.44}$$

Their number is at most $(4/3)\sum_k a_k^{-1}$. Summing (C.43), the removed observations, and the observations below the analyzed bands gives

$$\sum_{t:B_t=1,J_t=j} d_t \leq 49 \sum_{t:B_t=1,J_t=j} u_t^{\mathrm{c}} + \frac{6912L + 96/7}{\Delta_j} \leq 49 \sum_{t:B_t=1,J_t=j} u_t^{\mathrm{c}} + 6928L/\Delta_j.$$

For columns whose cells all have mean payoffs at least v, set $\Delta_j = +\infty$; these columns have $d_t \leq 0$. Since $\sum_j \Delta_j^{-1} \leq C_c$, on $\mathcal{G}$ we have the pathwise bound

$$\sum_t B_t d_t \leq 49 \sum_t B_t u_t^{\mathrm{c}} + 6928 L C_c. \tag{C.45}$$

On $\mathcal{G}^c$, the left side is at most $2T$. Since $u_t^{\mathrm{c}} \geq 0$, taking expectations in (C.45) and using (C.38) gives

$$\mathbb{E}\sum_t B_t d_t \leq (32016L + 196)C_c + 2T\delta \leq 33000 L C_c + 2T\delta. \tag{C.46}$$

Independence of the current revelation indicator gives the **unconditional** identity

$$\mathbb{E}\sum_t B_t d_t = q\,\mathbb{E}\sum_t d_t = qR_T. \tag{C.47}$$

This identity remains valid when the conditional deficit depends on previous revelation indicators. Together, (C.46)–(C.47) prove (A.5).

C.7 Closing the saddle-complexity bound

Return to a strict pure saddle, with $\mathcal{I}_0, \mathcal{I}_1, S$, and $\mathcal{B}$ as defined in Sections C.2–C.3. On $\mathcal{I}_0$, recommendation i_* is mapped to the leader. On $\mathcal{I}_1$, it is mapped to i_*, whose mean payoff exceeds the safety value by at least β whenever $J_t \neq j_*$. Decomposing the half-and-half mixture (7), the leader contributes a full coefficient on $\mathcal{I}_0$ and one half on $\mathcal{I}_1$. Equations (C.10), (C.13), and (C.17) imply

$$R_T \leq 648 L C_s + \frac{\mathcal{B}}{2} - \frac{\beta}{2}\mathbb{E}|\mathcal{I}_0| - \frac{7\beta}{16}\mathbb{E}T_{\mathrm{off}}(\mathcal{I}_1) + 4T\delta. \tag{C.48}$$

The constant is $96 + 1104/2 = 648$.

It remains to control the weighted action count S in $\mathcal{B}$. Separate the rounds on which i_* is eligible and $J_t = j_*$ from all other rounds. On the former, the probability-weighted row gaps equal the conditional regret. Let Z be their expected sum and let U count the other rounds in expectation:

$$U = \mathbb{E}[|\mathcal{I}_0| + T_{\mathrm{off}}(\mathcal{I}_1)], \qquad Z = \mathbb{E}\sum_{t\in\mathcal{I}_1: J_t = j_*} \sum_{i \neq i_*} p_t(i)\alpha_i.$$

On the rounds counted by U, the conditional deficit is at least -2 and the weighted row gaps are at most two, since every $\alpha_i \leq 2$. Hence $Z \leq R_T + 2U$ and $S \leq Z + 2U$. Set $Y = Z + 2U$: it bounds S from above and lets us express the negative count terms in (C.48) using the same quantity. Specifically,

$$S \leq Y, \qquad U \geq (Y - R_T)/4, \qquad Y \geq 0. \tag{C.49}$$

Both negative coefficients in (C.48) are at most $-7\beta/16$. Put $b = 4\sqrt{2}K_0\sqrt{A_\alpha}$; substituting (C.17) and (C.49) gives

$$(1 - 7\beta/64)R_T \le 648LC_s + D_1 + 4T\delta + \frac{b}{2}\sqrt{Y} - \frac{7\beta}{64}Y. \tag{C.50}$$

The maximum of the last two terms over $Y \ge 0$ is $4b^2/(7\beta)$, and $b^2 = 32K_0^2A_\alpha$. Since $\beta \le 2$,

$$\begin{aligned} R_T &\le \frac{32}{25}\left[\left(648L + \frac{128K_0^2}{7}\right)C_s + D_1 + 4T\delta\right] \\ &\le (830L + 24K_0^2)C_s + 2D_1 + 6T\delta \\ &\le 156030LC_s + 6T\delta \le 160000LC_s + 6T\delta. \end{aligned} \tag{C.51}$$

For the last line, use $\log D_0 \le L$, $K_0^2 = 6400 \log D_0$, and $m \le 8C_s$. The denominator in (C.50) can be replaced by $25/32$ since the maximized bracket is nonnegative. This proves (A.4), including $q = 0$.

C.8 Final constants and endpoints

Set $\delta = 1/[6(T+1)^2]$, so $L = L_T$. Equations (A.3) and (C.51) give the general and strict-saddle bounds in (A.1). For $q > 0$ and $C_c > 0$, at least one positive gap lies in $(0, 2]$, hence $C_c \ge 1/2$. Since $2T\delta \le 1/12$, (A.5) gives

$$R_T \le 33001 L_T C_c/q. \tag{C.52}$$

For $C_c = 0$, all mean-payoff deficits are nonpositive, so $R_T \le 0$. At $q = 0$, (A.3) and (C.51) remain valid in their respective domains. Together with $R_T \le 2T$, these observations prove Theorem A.1.

Incomparable gap complexities. Both matrices in Section 2 have strict saddle $(1, 1)$ and row gap $1/4$. Their displayed complexities satisfy $C_s/C_c \to \infty$ in the first example and $C_c/C_s \to \infty$ in the second as $\varepsilon \downarrow 0$.

D The enhanced policy and fixed-instance coefficients

D.1 Concentration on selected observation subsequences

For the analysis, expose the current column before the row draw. The conditioning time depends on the statistic. For actual-leader samples, first expose I_t, so the selection indicator $\mathbf{1}\{b_t = I_t = i\}$ is fixed before the reward and revelation indicator; the selected reward has mean A_{iJ_t} and the revelation indicator has mean q. For pair reward sums, condition before drawing I_t: the increment $\mathbf{1}\{I_t = a\}r_t$ then has mean $p_t(a)A_{aJ_t}$. Write $\mathbb{E}_t$ for conditional expectation at the corresponding time, before the increment being analyzed. Set that increment to zero on excluded rounds.

Lemma D.1 (concentration bounds holding simultaneously at all times). Let $X_t \in [0, 1]$ be adapted, $A_s = \sum_{t \le s} \mathbb{E}_t X_t$, and $S_s = \sum_{t \le s} X_t$. For $L \ge 1$ and $0 < \epsilon \le 1$, with probability at least $1 - 2e^{-L}$, every $s \le T$ satisfies

$$|S_s - A_s| \le \epsilon A_s + 4L/\epsilon. \tag{D.1}$$

In particular $S_s \ge A_s/2 - 8L$. Also, $S_s \le (e-1)A_s + L$ simultaneously for all $s \le T$, with failure probability at most e^{-L}. If D_t are martingale differences with $|D_t| \le b$ and $V_s = \sum_{t \le s} \mathbb{E}_t D_t^2 \le cW_s$, where $0 \le W_s \le T$, then, for $d_T = 2 + \lceil \log_2(T+1) \rceil$,

$$\left|\sum_{t \le s} D_t\right| \le 2\sqrt{cL(W_s + 1)} + \frac{2bL}{3} \quad \text{for all } s \le T \tag{D.2}$$

outside an event of probability at most $2Td_Te^{-L}$.

Proof. Convexity gives $\mathbb{E}_t e^{\lambda X_t} \le \exp((e^\lambda - 1)\mathbb{E}_t X_t)$. The exponential supermartingale and Ville's inequality, with $\lambda = \pm\epsilon/2$, yield (D.1), since $e^\lambda - 1 - \lambda \le \lambda^2$ for $|\lambda| \le 1/2$. Taking $\lambda = 1$ gives the upper bound. For (D.2), Freedman's inequality gives a two-sided error $\sqrt{2vL} + 2bL/3$ on $\{V_s \le v\}$, with failure probability at most $2e^{-L}$. Partition the range of W_s into $[0, 1]$ and dyadic intervals. Their upper endpoints are at most $2W_s$ when $W_s > 1$; a union bound over at most d_T intervals and T possible times proves (D.2). It also covers adaptively interrupted subsequences. □

D.2 AN ENHANCED POLICY FOR THE TWO-ROW FAMILY

Theorem D.2 (adaptation on the two-row family). A single policy receiving only m, n, T satisfies, for $L_+ = \log(2^{30}m^2(n+1)(T+1)^{10})$ and a universal C,

$$R_T \le C\sqrt{mTL_+} + C, \qquad R_T \le CL_+C_c/q \ (q>0), \qquad R_T \le CL_+C_s + C \text{ at strict saddles.} \tag{D.3}$$

On the two-row family (17), with $0 < \alpha \le 1/16$ and $0 < \beta \le 1/96$, it also satisfies

$$R_T \le CL_+ \min\left\{\alpha T, \sqrt{T}, \frac{1}{\alpha} + \min\left\{\frac{1}{\alpha\beta}, \frac{1}{\sqrt{\alpha q}}\right\}\right\}. \tag{D.4}$$

At the boundary $\beta = 0$, its guarantee is

$$R_T \le CL_+ \min\left\{\alpha T, \sqrt{T}, \frac{1}{\alpha} + \frac{1}{\sqrt{\alpha q}}\right\}. \tag{D.5}$$

Reciprocal revelation terms are $+\infty$ at $q = 0$.

Use the enhancement for $m = n = 2$ and the base policy otherwise. Write $L = L_+$, let b be the current leader, $w = \max_i M_i$, $\mathcal{E}$ the refined eligible set, and $\mathcal{A} = \mathcal{E} \setminus \{b\}$. Decisions use the observed history from completed rounds.

D.2.1 CONFIDENCE INTERVALS AND ADDITIONAL STATISTICS

We keep two kinds of records. For each row, actual-leader records use rounds when that row both leads and is played, with weight one. For each ordered pair (b, a), pair records use rounds when b leads and $a \in \mathcal{A}$, with weight p_a, the probability of playing a. Each group has its own accumulated weights, revealed counts, and column-frequency interval. The raw intervals for individual matrix entries use all revealed observations of that entry.

Put $\epsilon = 2^{-12}$, $h_0 = 4L/\epsilon$, and $\mathcal{B}(w) = 8\sqrt{(w+1)L} + 8L$. In either kind of record, the weights $\omega_t \in [0,1]$ are fixed before the new revelation indicator is drawn. After each new record, maintain

$$W = \sum \omega_t, \quad R = \sum \omega_t B_t, \quad R_2 = \sum \omega_t B_t \mathbf{1}\{J_t = 2\}; \qquad O = \sum \omega_t \mathbf{1}\{J_t = 2\}.$$

The learner can compute W, R, R_2; O is used in the analysis. Lemma D.1 gives

$$|R - qW| \le \epsilon qW + h_0, \qquad |R_2 - qO| \le \epsilon qO + h_0. \tag{D.6}$$

Each weight is fixed before its independent revelation indicator is drawn, giving (D.6) for that group. For $W > 0$, define the group's column-2 fraction as $\theta = O/W$. We first use R/W to form an interval for the unknown q, then use that interval with R_2 to bound θ. Write $[x]_0^1 = \min\{1, \max\{0, x\}\}$ and set

$$\begin{aligned} q_- &= \left[\frac{R - h_0}{(1+\epsilon)W}\right]_0^1, \quad q_+ = \left[\frac{R + h_0}{(1-\epsilon)W}\right]_0^1, \\ \theta_- &= \left[\frac{R_2 - h_0}{(1+\epsilon)q_+W}\right]_0^1, \quad \theta_+ = \left[\frac{R_2 + h_0}{(1-\epsilon)q_-W}\right]_0^1. \end{aligned} \tag{D.7}$$

For a zero denominator, use zero for the corresponding lower endpoint or one for the upper endpoint. Use $[0,1]$ at $W = 0$ or for contradictory intervals, and average only for $W > 0$. On (D.6), the intervals contain q, θ and satisfy

$$qO \ge 2^{40}L \quad \Longrightarrow \quad |\theta' - \theta| \le 8\epsilon\theta \quad (\theta' \in [\theta_-, \theta_+]). \tag{D.8}$$

Indeed $h_0/(qO) \le 2^{-26}$ and $h_0/(qW) \le 2^{-26}$; substitution first gives $q_\pm/q \in [1-3\epsilon, 1+3\epsilon]$, then (D.8). Raw intervals $[l_{ij}^0, u_{ij}^0]$ use actual revealed-cell rewards. At each positive sample count k, use radius $\sqrt{2L/k}$ and intersect with all previous intervals and $[-1,1]$. A cell with no revealed samples has interval $[-1,1]$.

Actual leader samples. For row i, we seek an upper bound on its minimum payoff $v_i = \min_j A_{ij}$. Select rounds with $b_t = I_t = i$. With count $k > 0$ and empirical mean of observed rewards $\widehat{\mu}$, let $u = \widehat{\mu} + \mathcal{B}(k)/k$. With unit weights for revelation counts and $g = \min\{1/2, \theta_-\}$ from this group's frequency interval, use the ordinary upper bound u and, when $l^0_{i2} > 0$,

$$P_i^{\text{extra}} = \frac{u - g l^0_{i2}}{1 - g}. \tag{D.9}$$

Set each upper bound to the minimum of one and all its values computed so far. Let $\bar{\mu}_i$ be the subsequence average of A_{iJ_t}. Reward confidence and $g \le \theta$ give $u \ge \bar{\mu}_i \ge v_i + \theta(A_{i2} - v_i) \ge (1-g)v_i + g l^0_{i2}$. Thus both quantities upper-bound v_i and enter the stored row bound used to choose the leader.

Comparing row rewards with the same weights. For each ordered pair (b, a), select rounds with leader b and $a \in \mathcal{A}$. Sum rewards from plays of a and weight each reward from a play of b by the selection-probability ratio p_a/p_b. Both sums then have conditional-mean increments weighted by p_a. Record

$$\Delta W = p_a, \qquad \Delta S_a = \mathbf{1}\{I = a\} r, \qquad \Delta S_b = (p_a/p_b)\mathbf{1}\{I = b\} r, \tag{D.10}$$

with weights for revelation counts $\omega_t = p_a$. Each reward-sum increment has mean $p_a A_{iJ}$, magnitude at most one, and second moment at most p_a, so Lemma D.1 gives

$$\left| S_i - \sum p_a A_{iJ} \right| \le \mathcal{B}(W), \qquad i \in \{a, b\}. \tag{D.11}$$

For $W > 0$, let $l_i, u_i = S_i/W \mp \mathcal{B}(W)/W$. They bound the weighted mean $\bar{\mu}_i = (1-\theta)A_{i1} + \theta A_{i2}$ from (19). Solving for A_{i1} gives $(\bar{\mu}_i - \theta A_{i2})/(1-\theta)$. When $\theta_+ < 1$, take the largest and smallest values allowed by the input intervals:

$$\widetilde{U}_{i1} = \max_{\theta' \in [\theta_-, \theta_+]} \frac{u_i - \theta' l^0_{i2}}{1 - \theta'}, \qquad \widetilde{L}_{i1} = \min_{\theta' \in [\theta_-, \theta_+]} \frac{l_i - \theta' u^0_{i2}}{1 - \theta'}. \tag{D.12}$$

Substituting the true means $(1-\theta)A_{i1} + \theta A_{i2}$ proves containment. Extrema occur at endpoints. Skip inference at $W = 0$, $\theta_+ = 1$, or contradictory raw inputs. Intersect each new inferred interval with $[-1, 1]$ and discard it if empty. Combine the accepted endpoints with all earlier endpoints and the raw bounds: L_{ij} is the largest lower endpoint and U_{ij} the smallest upper endpoint. Tests use these endpoints even if they cross when confidence bounds fail.

Dominance filter. Refine the original eligible set by removing row i if some j satisfies

$$U_{ij} \le w \quad \text{and} \quad U_{ij} < \max_k L_{kj}. \tag{D.13}$$

When the entry intervals contain their true means, a strict-saddle row i_* can be excluded by this test only through $j \ne j_*$, because $U_{i_* j_*} \ge v \ge L_{k j_*}$ for every k. Its exclusion therefore implies $w \ge U_{i_* j} \ge v + \beta$.

Joint confidence event. For each pair (b, a), let K_i, $i \in \{a, b\}$, count column-2 revelations on its selected rounds, with increment $B\mathbf{1}\{I = i, J = 2\}$. This is an actual sample count for row i; the same pair's R_2 instead adds $p_a B\mathbf{1}\{J = 2\}$. The conditional mean of the K_i increment is $q p_i \mathbf{1}\{J = 2\} \ge q p_a \mathbf{1}\{J = 2\}$, so Lemma D.1 gives $K_i \ge qO/2 - 8L$.

Define $\mathcal{G}_+$ by intersecting these bounds, the base event $\mathcal{G}$ of Lemma C.1, (D.6), and all new reward and raw-cell bounds. On the two-row family, including $\beta = 0$, also require the simultaneous comparison event of Lemma D.3, proved below. All statistics use the conditioning times specified in Lemma D.1. With $d_T = 2 + \lceil \log_2(T+1) \rceil$, total failure probability is at most $100(T+1)(d_T+1)e^{-L}$, including the original $32Te^{-L}$, hence $\Pr(\mathcal{G}_+^c) \le \delta_+ := 1/[10(T+1)^3]$.

D.2.2 Implementation and inherited bounds

Update order. Let E_i^{cap} be the running minimum bound from actual-leader samples, initially 1, and use

$$M_i = \max\{-1, \min(1, \min_j U_{ij}, \min_{\iota, e} P_{i,\iota,e}, E_i^{\text{cap}})\}.$$

Algorithm D.1 Enhanced policy

Require: Dimensions m, n and horizon T.

if $(m, n) \neq (2, 2)$ **then**
  Run the policy of Section B.2 with L_+; **return** .
end if
Initialize the original and additional statistics, and states $(G^{(b)}, \lambda^{(b)}) = (0, 16\mathbf{1})$, $b = 1, 2$.
for $t = 1, \ldots, T$ **do**
  Compute (D.7), update the bounds in (D.9) and (D.12), and form M_i. Choose $b \in \arg\max_i M_i$ by the fixed tie rule and set $w = M_b$.
  Recompute the base eligible set using the current refined U, H, and w; remove rows satisfying (D.13) to obtain $\mathcal{E}$.
  Obtain z from state b. Set $f(i) = i$ for eligible i and $f(i) = b$ otherwise, $Q_j = \sum_{i:f(i)=j} z_i$, $p = (e_b + Q)/2$, and $\mathcal{A} = \mathcal{E} \setminus \{b\}$.
  Save $b, \mathcal{E}, \mathcal{A}, x, z, p$ and the coefficients; announce p, draw $I \sim p$ after J_t is chosen, and observe feedback.
  Form $X = \mathbf{1}\{I = b\}r/p_b$ and u by (B.13)–(B.14), with saved probabilities and a fresh auxiliary draw. Append X, including zeros, to each $(b, \iota, \mathbf{1}\{\iota \in \mathcal{E}\})$ subsequence using saved $\mathcal{E}$, and update its P by (B.10).
  Update actual-leader samples when $I = b$ and each selected pair once by (D.10), using saved conditions. Update their totals R, R_2 for revealed observations. Only selected pairs increment K_i by $B\mathbf{1}\{I = i, J = 2\}$; every revelation updates its raw cell and dyadic column blocks.
  Update only state b: $G^{(b)} \leftarrow G^{(b)} + u$ and apply (B.15) with saved coefficients and z.
end for

Before each decision, re-evaluate (D.9) and (D.12) for all additional subsequences, including inactive ones, from completed totals and current raw bounds; keep the largest lower and smallest upper endpoints so far. For an inactive pair, W, S_a, S_b stay unchanged, but revelations on other rounds can tighten its raw column-2 intervals. Substituting those intervals into its saved reward sums can then tighten its column-1 intervals. Both leader states use (B.11)–(B.15) with global horizon T.

Counts and sums start at zero, raw intervals at $[-1, 1]$, upper bounds at 1, and $H_j = -1$. Each leader resumes its state on return. Bounds $p_b \geq 1/2$ and $p_a = z_a/2 \geq 1/[4(T + 1)]$ for active a ensure positive denominators; zero updates preserve coefficients.

Storage. Two actual-leader and two ordered-pair subsequences, totals R, R_2, stored endpoints, and the additional learner state use $O(1)$ extra arithmetic words. Updating and re-evaluating these sums and counts costs $O(1)$ per round, totaling $O(T)$ under the convention of Section B.3; exact convex optimization is counted separately.

Learner comparison. For each state define $V_{bi} = \sum_t \mathbf{1}\{b_t = b, i \in \mathcal{A}_t\} \min\{z_{ti}, 1 - z_{ti}\}$. Apply Lemma C.2 at its predictably selected updates, sum over both states, and use $\sum_b \sqrt{\mathbb{E}V_{bi}} \leq \sqrt{2\mathbb{E}\sum_b V_{bi}}$; only universal constants change.

Optimism and cell bounds. Equations (D.9) and (D.12) preserve optimism, so $M_i \geq v$ for a maximin row and $w \geq v$. Refined U_{ij} are nonincreasing and bounded by raw upper bounds. For $\Delta_{ij} = v - A_{ij} > 0$, raw saturation gives $U_{ij} \leq v - 7\Delta_{ij}/8 < w$. Thus its row is played only as an alternative, requiring $H_j \leq U_{ij}$. Alternative rows still receive total probability at most one half, and the H_j block bounds hold, so Appendix C's cell argument applies.

Saddle bounds and conclusion. The original subsequence bounds give the same reward comparisons. Excluding a strict-saddle row by (D.13) forces $w \geq v + \beta$; Appendix C's saddle argument then proves (D.3). In the cell bound, $C_c \geq 1/2$ for $C_c > 0$ absorbs exceptional events; at $C_c = 0$ every deficit is nonpositive.

D.2.3 Regret when the unsafe row is the leader

Fix a matrix from (17), allowing $\beta = 0$, with safe row h and unsafe row a. Write $c = 1/2 - \alpha \geq 7/16$. We split the regret according to which row leads. When a leads, plays of h have nonpositive mean-payoff deficits, so we use the stored row upper bound to control actual plays of a. When h

leads, the pair records compare the rows with weights equal to the probability of choosing a. This comparison bounds the conditional mean-payoff deficit on those rounds; taking expectations will combine the two parts.

On $\mathcal{G}_+$, $M_h \geq 0$. Let K count actual plays $b_t = I_t = a$, and O their column-2 plays. Their cumulative mean-payoff deficit is

$$r_a = \alpha K - O/2. \tag{D.14}$$

For $q > 0$, set $H = 2^{40}L/q$. Before the last such sample, $K^- = K-1$ and $O^- \in \{O-1, O\}$. Since a still leads, $M_a \geq M_h \geq 0$, so every stored upper bound entering M_a is nonnegative. If $O^- > H$, (D.8) gives $g \geq O^-/(2K^-)$ and the revealed count is at least $(1-\epsilon)qO^- - h_0 \geq qO^-/2 - 8L$, so $l^0_{a2} \geq 1/4$. Nonnegativity of (D.9) then requires $u \geq g l^0_{a2}$. Put $r_a^- = \alpha K^- - O^-/2$. Its negative is the sum of the true means on these samples, so reward confidence gives

$$\frac{O^-}{8K^-} \leq g l^0_{a2} \leq u \leq -\frac{r_a^-}{K^-} + \frac{2\mathcal{B}(K^-)}{K^-},$$
$$r_a^- \leq 2\mathcal{B}(K^-) - O^-/8.$$

The final sample increases the deficit by at most one. Using $O^- \geq O - 1$ and weakening the subtraction to cover the threshold H gives

$$r_a \leq 2\mathcal{B}(K) + 1 - (O - H - 1)_+/8. \tag{D.15}$$

If $O^- \leq H$, the ordinary upper bound gives the same inequality with a zero subtraction term; empty and single-sample sequences also satisfy it. For $r_a > 0$, substitute $K = (r_a + O/2)/\alpha$, split the square root in $\mathcal{B}(K)$, and use $16\sqrt{Lr_a/\alpha} \leq r_a/2 + 128L/\alpha$. The remaining square root in O and its negative linear term satisfy $u\sqrt{O} - v_0(O - H')_+ \leq u\sqrt{H'} + u^2/(4v_0)$ for $u, H' \geq 0, v_0 > 0$. Applying this with $H' = H + 1$ gives

$$r_a \leq CL\left(\frac{1}{\alpha} + \frac{1}{\sqrt{\alpha q}}\right). \tag{D.16}$$

For $r_a \leq 0$ the bound is immediate. When the unsafe row is the leader, plays of the safe row contribute nonpositive deficits.

D.2.4 Cumulative comparison on rounds led by the safe row

The following comparison will be used at the last round when (h, a) is active. That round depends on the observed history, so we need the bound to hold simultaneously at all times.

Lemma D.3. For the two-row family (17), including $\beta = 0$, put $\chi_t = \mathbf{1}\{b_t = h\}$, $W_s = \sum_{t \leq s} \chi_t p_t(a)$, and $D_s = \sum_{t \leq s} \chi_t p_t(a)(A_{hJ_t} - A_{aJ_t})$. With probability at least $1 - [1 + 2T(1 + d_T)]e^{-L}$, every $0 \leq s \leq T$ satisfies

$$D_s \leq 1000(\sqrt{LW_s} + L). \tag{D.17}$$

On $\mathcal{G}_+$, this bound holds for every s.

Proof. We first control the adaptive coefficient increments and apply the deterministic learner comparison. Two martingale corrections then account for probability smoothing and replace the update estimates by their conditional means. For the analysis, expose the past and current column before the row and auxiliary draws, and insert zero increments outside leader-h rounds. Write $\mathbb{E}_t$ for this conditional expectation. State h has $u_h = 0$; when a is active, it is the sole updated coordinate and $z_a = 2p_a$. Otherwise $u = 0$ and $p_a = 0$. Set $\mathfrak{h}_t = 2\sqrt{z_{ta}}$ and let $\mathfrak{a}_t = a_{ta}$ be the update quantity in (B.15), zero when inactive. Then

$$0 \leq \mathfrak{h}_t \mathfrak{a}_t \leq 16 u_{ta}^2 z_{ta}^2 \leq 64, \qquad \mathbb{E}_t[\mathfrak{h}_t \mathfrak{a}_t] \leq 192 \chi_t p_t(a). \tag{D.18}$$

The exponential bound of Lemma D.1 gives, except with probability e^{-L}, simultaneously for all $s \leq T$,

$$\sum_{t \leq s} \mathfrak{h}_t \mathfrak{a}_t \leq 192(e-1)W_s + 64L \leq 330 W_s + 64L. \tag{D.19}$$

By any round $s \leq T$, each learner state has received at most T updates. Equation (C.28) controls the accumulated coefficient terms, and (C.31) bounds the learner's comparison with a smoothed pure

action by these terms. Write $\eta = 1/[2(T+1)]$ and $\bar{e}_h = (1-2\eta)e_h + \eta\mathbf{1}$. With $\Lambda = \log(2(T+1)) \le L$, these deterministic bounds yield

$$\sum_{t\le s}\langle x_t - \bar{e}_h, u_t\rangle \le 128\Lambda + 32\sqrt{2} + 8\sqrt{\Lambda(330W_s + 64L)} + 4\Lambda/\log 2$$
$$< 146\sqrt{LW_s} + 244L,$$

The sums use only state-h updates.

Two martingale corrections recover the conditional payoff comparison. First, put $C_t = \chi_t 2\eta(1 - x_{ta})u_{ta}$. Since $z_a - (x_a - \eta) = 2\eta(1 - x_a)$, $\sum \chi_t z_{ta}u_{ta} = \sum\langle x_t - \bar{e}_h, u_t\rangle + \sum C_t$. The bounds on the update estimates give $|\mathbb{E}_t C_t| \le 2\eta\chi_t$, $\mathbb{E}_t C_t^2 \le 16\eta^2\chi_t/z_{ta} \le 16\eta\chi_t$, and $|C_t| \le 4$. Thus the centered increments have magnitude at most 8, total variance at most 8, and total mean at most 1. Freedman's inequality and a union bound over all times $s \le T$ give $\sum_{t\le s} C_t \le 1 + 4\sqrt{L} + (16/3)L \le 11L$, with failure probability at most $2Te^{-L}$.

Second, $Z_t = \chi_t z_{ta}u_{ta}$ has $|Z_t| \le 2$, second moment at most $8\chi_t p_t(a)$, and mean $\chi_t p_t(a)(A_{hJ_t} - A_{aJ_t})$. Apply (D.2) with $b = 4, c = 8$ to $\mathbb{E}_t Z_t - Z_t$ to obtain $\sum_{t\le s}(\mathbb{E}_t Z_t - Z_t) \le 8\sqrt{LW_s} + 16L$ for all $s \le T$, with failure probability at most $2Td_T e^{-L}$. Summing gives $D_s < 154\sqrt{LW_s} + 271L$, which proves (D.17) and its failure bound. Both leader choices enter the joint failure bound. □

For the pair (h, a), define $O_s = \sum_{t\le s}\chi_t p_t(a)\mathbf{1}\{J_t = 2\}$ and $r_s = \alpha W_s - O_s/2$. These weighted quantities satisfy $r_s = D_s - \beta O_s \le D_s$, so (D.17) also bounds r_s.

D.2.5 Exclusion and final expectation

We now give sufficient conditions for the observable exclusion test (D.13) to succeed. These conditions use the true parameters and the unobserved total O in the analysis; the policy applies (D.13) to its computed interval endpoints. For $q > 0$ put $A_0 = 2^{30}L/\alpha^2$, $H = 2^{40}L/q$. If the totals for (h, a) after a completed round satisfy

$$W \ge A_0, \qquad O \ge H, \qquad \theta = O/W \le 8\alpha \tag{D.20}$$

then, on $\mathcal{G}_+$,

$$U_{a1} \le -7\alpha/8 < -\alpha/8 \le L_{h1}. \tag{D.21}$$

Indeed each raw column-2 count is at least $qO/2 - 8L$, giving endpoint error at most ϵ. Also $|\theta' - \theta| \le 8\epsilon\theta$, $\theta_+ < 0.51$, and $2\mathcal{B}(W)/W \le \alpha/256$. After subtracting A_{i1} from the upper fraction in (D.12), its numerator error is $(\theta - \theta')(A_{i2} - A_{i1}) + e_i + \theta'(A_{i2} - l^0_{i2})$, where $|e_i| \le \alpha/256$. Using $|A_{i2} - A_{i1}| \le 1/2$ and $\theta \le 8\alpha$, its magnitude is less than $\alpha/32$. Division by $1 - \theta' > 0.49$ gives error below $\alpha/8$; the lower fraction obeys the same bound. This proves (D.21).

Since $w \ge 0$, dominance excludes a; historical endpoints preserve this exclusion. If (h, a) is never active, $W = O = r = 0$. Otherwise, before its last active round, at least one condition in (D.20) fails. Adding that round changes W, O by at most $1/2$, so their final totals W, O, with $r = \alpha W - O/2$, satisfy

$$W \le A_0 + 1, \qquad \text{or } O \le H + 1, \qquad \text{or } O \ge 8\alpha(W - 1). \tag{D.22}$$

The first case gives $r \le \alpha(A_0 + 1) = O(L/\alpha)$; the third gives $r \le 4\alpha$. In the second, for $r > 0$ substitute $W = (r + O/2)/\alpha$ in (D.17) and absorb the square-root term in r. Including $r \le 0$, this gives in all cases

$$r \le C\left(\frac{L}{\alpha} + \sqrt{\frac{L(H+1)}{\alpha}} + L\right) \le CL\left(\frac{1}{\alpha} + \frac{1}{\sqrt{\alpha q}}\right). \tag{D.23}$$

On leader-h rounds, regret and the sum of conditional deficits have equal expectations, and this sum is at most r. Its positive part outside $\mathcal{G}_+$ is at most αT. On leader-a rounds, (D.16) bounds the cumulative mean-payoff deficit on actual plays of the leader, with the same bound on the failure-event contribution. Consequently

$$R_T \le CL\left(\frac{1}{\alpha} + \frac{1}{\sqrt{\alpha q}}\right) + 2\alpha T\delta_+. \tag{D.24}$$

Absorb the last term and take the minimum with (D.3) and the pointwise bound $R_T \le \alpha T$. For $\beta > 0$, $C_s = 1/(\alpha\beta)$ supplies (D.4). The same proof gives (D.24) at $\beta = 0$; the revealed, worst-case, and pointwise bounds then give (D.5). At $q = 0$ use the finite branches. □

D.3 Fixed-instance logarithmic coefficients

At the boundary $\beta = 0$ of the enhanced family, the safe row pays zero in every column, so the opponent's best responses are tied. For the parameters below, we show that the enhanced policy's $q^{-1/2}$ dependence matches a fixed-instance lower bound. Each matrix and revelation probability is fixed as the horizon grows.

We use the common policy class Π of Appendix A.3. Its policies satisfy the general worst-case bound, have logarithmic regret on every fixed matrix at every fixed $q > 0$, and also have logarithmic regret at $q = 0$ for every strict pure saddle. Recall from (A.6)–(A.7) that Φ_π is the positive part of worst-case regret and

$$c_\pi(A, q) = \limsup_{T\to\infty} \frac{\Phi_\pi(A, q, T)}{\log T}, \qquad c_*(A, q) = \inf_{\pi\in\Pi} c_\pi(A, q).$$

The arguments of $\mathrm{d}(u, v) = \mathrm{kl}((1 + u)/2, (1 + v)/2)$ are means of $\{-1, +1\}$ rewards; kl is binary relative entropy. The proof uses change of measure for the joint law of rewards, revelation indicators, and observed column labels, together with signed regret (Lai & Robbins, 1985; Kaufmann et al., 2016; Combes et al., 2017).

D.3.1 Coefficients with tied opponent best responses

Theorem D.4 (coefficients with tied opponent best responses). Fix either row order of

$$A_\alpha = \begin{pmatrix} 0 & 0 \\ -\alpha & c \end{pmatrix}, \qquad c = \tfrac{1}{2} - \alpha, \quad 0 < \alpha \le 1/16, \quad 0 < q \le \frac{\alpha}{4c}. \tag{D.25}$$

The single enhanced policy π^+ belongs to Π and satisfies

$$\frac{1}{2\sqrt{\alpha c q}} \le c_*(A_\alpha, q) \le c_{\pi^+}(A_\alpha, q) \le \frac{20C}{\sqrt{\alpha c q}}, \tag{D.26}$$

where C is the universal constant in (D.5), and α, q are fixed before the horizon limit. The lower bound also covers policies receiving q and satisfying the guarantees (i)–(iii) in Appendix A.3.

D.3.2 Proof

Fix $\pi \in \Pi$ and the displayed row order. We choose a column distribution under which row 2 has mean payoff $-z < 0$, then compare it with an alternative where row 2 is better than row 1. Only observations from row 2 distinguish the two environments, and each play of this row costs z in expected regret. As the alternative mean tends to zero, the divergence per such play will be bounded by $z^2 + q\alpha c$: the first term comes from feedback with hidden columns, the second from revealed columns. We choose $z^2 = q\alpha c$ to balance these terms in the ratio $z/(z^2 + q\alpha c)$. The following column probabilities give the desired mean $-z$:

$$z = \sqrt{\alpha c q}, \qquad \mu_2 = \frac{\alpha - z}{\alpha + c}, \qquad \mu_1 = \frac{c + z}{\alpha + c}. \tag{D.27}$$

Since $0 < z \le \alpha/2$, both coordinates are positive. Choose independent columns with law μ, rewards in $\{-1, +1\}$ with matrix entries as means, and the prescribed revelation indicator. Let $N_i(T) = \sum_{t\le T} \mathbf{1}\{I_t = i\}$. Under this column law, the mean rewards of the two rows are $0, -z$, giving

$$R_T(A_\alpha; \mu) = z\mathbb{E}_{A_\alpha} N_2(T), \qquad \mathbb{E}_{A_\alpha} N_2(T) = O(\log T). \tag{D.28}$$

Keep row 1 unchanged and give row 2 a positive mean in both columns. For each fixed $\varepsilon \in (0, 1/2)$, use the alternative

$$A^\varepsilon = \begin{pmatrix} 0 & 0 \\ \varepsilon & \varepsilon \end{pmatrix}, \qquad R_T(A^\varepsilon; \mu) = \varepsilon\mathbb{E}_{A^\varepsilon} N_1(T), \qquad \mathbb{E}_{A^\varepsilon} N_1(T) = O(\log T), \tag{D.29}$$

with the same column law, binary reward construction, and revelation rate. Property (ii) gives both count bounds: the regrets are nonnegative and the alternative has safety value ε.

After a row-2 play, the divergence between the observed feedback distributions is

$$D_\varepsilon = (1 - q)\mathrm{d}(-z, \varepsilon) + q\big[\mu_1 \mathrm{d}(-\alpha, \varepsilon) + \mu_2 \mathrm{d}(c, \varepsilon)\big]. \tag{D.30}$$

When the column label is hidden, binary rewards have means $-z$ and ε; when it is revealed, the divergences are weighted by the common column law. Row 1 contributes zero, and action kernels coincide at the same observed history. Let P_T and Q_T be the laws of the learner's observed history and private randomization under A_α and A^ε, respectively. Then

$$\mathrm{KL}(P_T \| Q_T) = D_\varepsilon \mathbb{E}_{A_\alpha} N_2(T). \tag{D.31}$$

Positive probabilities for both rewards, common supports for the revelation indicators, and identical private randomness laws ensure mutual absolute continuity.

For $E_T = \{N_1(T) \geq T/2\}$, let $p_T = P_T(E_T)$ and $s_T = Q_T(E_T)$. Markov's inequality gives $p_T \geq 1 - O(\log T/T)$ and $s_T = O(\log T/T)$. Data processing and the binary entropy bound imply

$$D_\varepsilon \mathbb{E}_{A_\alpha} N_2(T) \geq \mathrm{kl}(p_T, s_T) \geq p_T \log(1/s_T) - \log 2 \geq (1 - o(1)) \log T. \tag{D.32}$$

Take the lower limit in T at fixed ε, then let $\varepsilon \downarrow 0$. Continuity gives

$$\begin{gathered} \liminf_{T\to\infty} \frac{R_T(A_\alpha; \mu)}{\log T} \geq \frac{z}{K}, \\ K = (1-q)\mathrm{d}(-z, 0) + q[\mu_1 \mathrm{d}(-\alpha, 0) + \mu_2 \mathrm{d}(c, 0)] > 0. \end{gathered} \tag{D.33}$$

For $|u| \leq 1/2$, the value and first derivative of $\mathrm{d}(u, 0)$ vanish at zero, and its second derivative is $1/(1-u^2) \leq 4/3$. Thus $\mathrm{d}(u, 0) \leq u^2$ and

$$\mu_1 \alpha^2 + \mu_2 c^2 = \alpha c - z(c - \alpha) \leq \alpha c, \qquad K \leq z^2 + q\alpha c = 2\alpha c q. \tag{D.34}$$

Since Φ_π dominates the nonnegative regret in (D.28), $c_\pi(A_\alpha, q) \geq 1/(2\sqrt{\alpha c q})$ for every $\pi \in \Pi$, including policies receiving the same q in both experiments. Taking the infimum proves the lower bound.

For the upper bound, (D.3) gives $\pi^+ \in \Pi$. Since $L_+/\log T \to 10$, (D.5) yields $c_{\pi^+}(A_\alpha, q) \leq 10C(1/\alpha + 1/\sqrt{\alpha q})$. Put $S = 1/\sqrt{\alpha c q}$. The restrictions give $1/\alpha \leq S/2$ and $1/\sqrt{\alpha q} = \sqrt{c}\, S \leq S/\sqrt{2}$, hence $c_{\pi^+} \leq 20CS$. Taking $c_* \leq c_{\pi^+}$ and exchanging rows proves (D.26). □

The positive entry c in (D.25) supplies both information and favorable payoffs. On this family, Theorem D.4 gives a smaller dependence on q than the q^{-1} coefficient for the different family in (A.8).

D.4 Identical gaps and different learning difficulty

We now ask whether matching gap parameters imply comparable learning difficulty when the candidate matrices are known. For each choice of parameters below, we compare two candidate families at the same finite horizon and revelation probability. We then vary these parameters together to measure how their minimax regrets separate.

Fix $0 < \varepsilon \leq 1/32$, $q = \varepsilon$, and $T = \lceil \varepsilon^{-4} \rceil$. For $c \in \{0, 7/16\}$, let $\mathcal{C}_c$ consist of the matrix

$$A_c = \begin{pmatrix} 0 & \varepsilon^2 \\ -1/16 & c \end{pmatrix} \tag{D.35}$$

and its row swap. Every matrix has safety value zero, a strict pure saddle in column 1, row gap $\alpha = 1/16$, column gap $\beta = \varepsilon^2$, $C_s = 16/\varepsilon^2$, and $C_c = 16$.

For each known candidate family, define the minimax regret $V_c = \inf_\pi \sup_{A \in \mathcal{C}_c} \sup_\nu R_T^{\pi,\nu}(A, q)$, where the learner knows $\mathcal{C}_c$, its numerical parameters, q, and T, and ν ranges over the opponents and reward processes allowed in Section 2. The infimum ranges over all policies allowed to use this information. The actual row order is unknown.

Corollary D.5. For a universal constant $K > 0$,

$$\begin{gathered} V_0 \geq \frac{2^{-12}}{\varepsilon}, \qquad 0 < V_{7/16} \leq \frac{K \log(1/\varepsilon)}{\sqrt{\varepsilon}}, \\ \frac{V_0}{V_{7/16}} \geq \frac{2^{-12}}{K\sqrt{\varepsilon}\log(1/\varepsilon)} \longrightarrow \infty \qquad (\varepsilon \downarrow 0). \end{gathered} \tag{D.36}$$

Thus the same dimensions, horizon, revelation rate, and gap parameters can correspond to different learning difficulty.

Proof. For $c = 7/16$, Theorem 2 gives an upper bound $CL_+(16+4/\sqrt{\varepsilon})$, where $L_+ = O(\log(1/\varepsilon))$. Always choosing column 1 gives average regret $1/32$ on the first round across the two row orders, and nonnegative expected regret thereafter; hence $V_{7/16} > 0$.

For $c = 0$, make column 1 rare at first. With probability ε of that column, the two row means differ by $O(\varepsilon)$ when labels are hidden, while a revealed column-1 sample occurs with probability $q\varepsilon = \varepsilon^2$. The calculation below bounds the divergence per round by ε^2. A phase of order ε^{-2} rounds therefore keeps the row orders hard to distinguish. The testing bound below then gives an unsafe-row probability of at least $3/8$, averaged over the two row orders, enough to incur a net loss of order ε per round.

Put $\ell = \lfloor 1/(8\varepsilon^2) \rfloor \le T$. During the first ℓ rounds, choose column 1 independently with probability ε; afterwards, always choose column 1. Use independent rewards in $\{-1, +1\}$ with the corresponding matrix entries as means. During the first ℓ rounds, the safe row has mean $g = \varepsilon^2(1-\varepsilon) \le \varepsilon/32$, and the other row has mean $-\varepsilon/16$.

For binary rewards with means $u, v \in [-1/2, 1/2]$, the inequality $\log x \le x - 1$ gives $\mathrm{d}(u, v) \le (u-v)^2/(1-v^2) \le 2(u-v)^2$. Conditional on either selected row, the divergence between the observed feedback distributions under the two row orders is therefore at most

$$2(\varepsilon/16 + \varepsilon^2)^2 + 2\varepsilon(\varepsilon/256 + \varepsilon^4) \le \varepsilon^2. \tag{D.37}$$

The first term bounds hidden-column feedback; the second accounts for revealed columns. The revelation indicators have the same law, and the learner's action kernels coincide at each observed history. The chain rule therefore bounds the divergence of the histories, including actions and revelation indicators, before round $t \le \ell$ by $(t-1)\varepsilon^2 \le 1/8$. Let $\Pr_+$ and $\Pr_-$ denote probabilities under the displayed row order and its swap. Pinsker's inequality gives $\Pr_+(I_t = 2) + \Pr_-(I_t = 1) \ge 3/4$. The probability of choosing the unsafe row, averaged over the two environments, is at least $3/8$. Thus their average regret per round is at least $(\varepsilon/16)(3/8) - (\varepsilon/32)(5/8) = \varepsilon/256$. Since $\ell \ge 1/(16\varepsilon^2)$ and all subsequent rounds have nonnegative expected regret, $V_0 \ge 2^{-12}/\varepsilon$. Dividing the two bounds proves the claim. □