SHARP RECOVERY THRESHOLDS FOR PAIRWISE SELF-REWARDED SOFTMAX UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

An evaluator can rank the optimal action above every alternative at every interior policy, yet a policy trained from sampled pairwise comparisons can still converge to a suboptimal action. We study this phenomenon for a finite softmax policy that samples two actions independently and updates its logits using their real-valued score difference. The evaluator may depend on the current policy, whereas the true rewards remain fixed. We control evaluator error by a uniform bound ε on the policy-weighted standard deviation of bias and an L-Lipschitz bound on pairwise bias contrasts. We assume that the step sizes have a divergent sum and a finite sum of squares, and that the comparison noise has bounded conditional variance. Under these conditions, we derive the sharp worst-case recovery threshold $\Psi_L(\Delta)$, where Δ is the minimum true reward gap. Below this threshold, the sampled update converges almost surely to the optimal action from any full-support initialization. At the boundary $\varepsilon = \Psi_L(\Delta)$ with $L > 0$, recovery still holds for two or three actions, but the sampled update can fail with four or more actions even when the evaluator ranks the optimal action first at every interior policy. On such failure trajectories, the optimal action is compared infinitely often, yet the cumulative conditional drift toward it remains finite. A full-information update still recovers at this boundary for every finite action set, which shows that the sampled update can fail even when the evaluator retains enough information to identify the optimum. Finally, balanced queries with an importance-corrected update recover the optimal action for $\varepsilon < \Delta/2$ without a sensitivity bound, whereas evaluator feedback alone cannot guarantee identification when $\varepsilon \geq \Delta/2$.

1 INTRODUCTION

Self-rewarding methods use evaluations produced by a model to improve its policy (Yuan et al., 2024). This setup differs from learning with fixed external rewards because the evaluator is part of the learning loop rather than a fixed source of feedback. As the policy changes, it changes which responses are compared and may also change the evaluator's score margins between those responses. The resulting feedback can therefore vary even when the external quality of the candidate responses remains fixed. Recent work has studied how such feedback can reinforce erroneous preferences (Tan et al., 2026; Yang et al., 2026).

Existing convergence theory does not fully resolve this setting. For fixed rewards, stochastic gradient bandits can converge without forced exploration (Mei et al., 2023; 2024). Comparative values can also provide unbiased policy-gradient estimators under fixed underlying values (Wu et al., 2023). These results do not cover an evaluator whose score margins change with the policy. Recent theory for self-rewarding methods gives iterative guarantees under realizability and conditions on confidence and margins (Fu et al., 2026). Other work studies binary self-confirmation and correction across agents (Yang et al., 2026). However, these analyses do not characterize the joint effect of a shrinking policy-dependent score margin and a declining probability of sampling the optimal action. This leaves open whether correct mean-score rankings alone are sufficient for recovery.

These gaps lead to two central questions. First, what quantitative control on policy-dependent evaluator bias is sufficient to guarantee recovery of the optimal action? Second, if a sampled update fails, does the evaluator itself lack enough information to identify the optimal action, or does the update fail to accumulate information that remains available?

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.

We study these questions in a finite-action model for a fixed prompt. We represent candidate responses as actions and the policy as a probability distribution over them. True rewards remain fixed and represent an unobserved external objective, such as correctness. The evaluator instead assigns mean scores that may depend on the current policy, and the learner observes only this evaluator feedback. This abstraction removes representation, generalization, and function approximation from the analysis so that we can isolate the feedback loop between sampling and evaluation. It retains the two couplings that matter here: the policy controls which responses are compared, and it may also change the evaluator margins in those comparisons.

We analyze a finite softmax policy that independently samples two actions and updates their logits using the real-valued difference between their evaluator scores. The main difficulty appears when the optimal action becomes rare. If its policy mass is p, then a comparison involving it occurs with probability of order p. Its mean score advantage may also shrink at order p, so the conditional drift toward the optimal action can fall to order p^2. Comparisons among more common actions can still retain order-one margins. The key issue is therefore not only whether the optimal action receives favorable comparisons, but whether the update accumulates enough favorable drift to recover it.

To characterize this interaction, we control two aspects of evaluator error. First, a uniform bound ε limits the policy-weighted standard deviation of evaluator bias across actions. Second, an L-Lipschitz condition limits how quickly pairwise bias contrasts can change as the policy changes. Score differences remove any additive offset shared by all actions, so the dispersion bound measures the part of evaluator bias that affects the update. We impose both controls uniformly over the closed simplex. Together with the true reward gap, these quantities provide the parameters for the recovery guarantees and failure results stated next.

Contributions. Our main contributions are as follows.

1. **Sharp recovery threshold.** We give a sharp worst-case characterization of recovery for the sampled pairwise softmax update. Let Δ denote the minimum true reward gap, ε the uniform bias-dispersion bound, and L the sensitivity of pairwise evaluator bias. If $\varepsilon < \Psi_L(\Delta)$, the policy converges almost surely to the optimal action for every finite action set and every full-support initialization. At the boundary $\varepsilon = \Psi_L(\Delta)$ with $L > 0$, the same guarantee holds for two or three actions, whereas the sampled update can fail with four or more actions even when the evaluator strictly ranks the optimal action first at every interior policy. Above the threshold, for every finite action set, we construct admissible evaluators and full-support initializations arbitrarily close to a suboptimal vertex from which convergence to that suboptimal action occurs with probability arbitrarily close to one.
2. **Why the boundary changes at four actions.** We identify the mechanism behind the dimension-dependent boundary. A conserved product prevents the optimal action from becoming rare too quickly and reduces recovery at equality to an integrability test. The correction must diverge for two or three actions but can be summable from four actions onward. Our critical evaluator realizes this latter regime: the optimal action is compared infinitely often, yet its total conditional correction remains finite and the policy can converge to a suboptimal action.
3. **Update failure versus information loss.** We show that failure of the sampled update does not necessarily mean that the evaluator has lost the information needed to identify the optimum. A full-information multiplicative-weights update still recovers the optimal action at the sampled-update boundary for every finite action set. We then introduce balanced queries, which evaluate each sampled pair at equal policy weights, together with an importance-corrected update. This method recovers the optimal action almost surely whenever $\varepsilon < \Delta/2$, without any sensitivity bound. In contrast, when $\varepsilon \geq \Delta/2$, we construct environments with different optimal actions but identical evaluator feedback under every adaptive query policy. Thus, $\Psi_L(\Delta)$ is the sharp recovery threshold for the sampled update, while $\Delta/2$ is the sharp limit for universal identification from evaluator feedback alone.

We also derive exact sample-path asymptotics on successful trajectories, including the decay of each suboptimal action probability and the frequencies of continued pairwise comparisons. Under conditionally sub-Gaussian comparison noise, we further give a finite-round identification guarantee for uniform balanced queries.

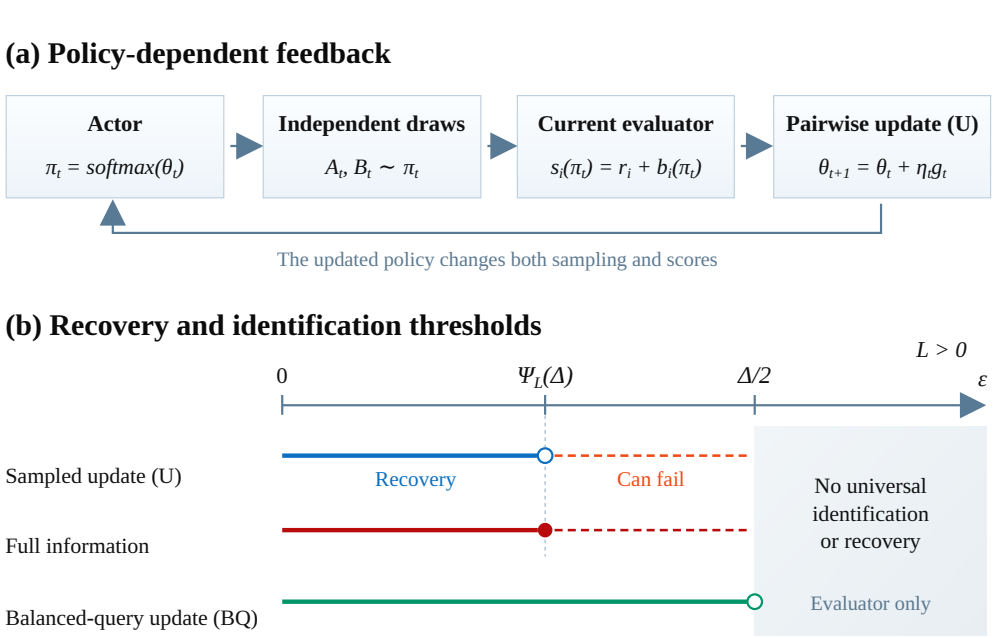

Figure 1: Problem and recovery landscape. (a) The actor samples pairs for real-valued contrast updates, and the actor policy also affects evaluator scores. (b) For $L > 0$, the sampled-update threshold $\Psi_L(\Delta)$ is below the evaluator-only limit $\Delta/2$ (schematic spacing). At $\Psi_L(\Delta)$, update (U) recovers for $K \leq 3$ but can fail for $K \geq 4$ despite correct mean-score rankings (Theorem 1); full-information updates still recover (Proposition 4). Balanced queries with update (BQ) recover for $\varepsilon < \Delta/2$, while universal evaluator-only identification fails at or above that limit (Theorem 5).

2 Policy-Dependent Evaluation and the Recovery Threshold

2.1 Problem Setup and Assumptions

Fix an integer $K \geq 2$. Let $\mathcal{P}_K = \{\pi \in [0,1]^K : \sum_i \pi_i = 1\}$, with interior $\mathcal{P}_K^\circ$ and vertices e_i. Fix $r \in \mathbb{R}^K$ with unique optimal action $i^\star := \arg\max_i r_i$, gaps $\Delta_j = r_{i^\star} - r_j > 0$, and minimum gap $\Delta = \min_{j \neq i^\star} \Delta_j$. Recovery means $\pi_t \to e_{i^\star}$.

At evaluation policy $\mu \in \mathcal{P}_K$, the mean scores are $s_i(\mu) = r_i + b_i(\mu)$. The map $b(\cdot)$ is fixed throughout learning, although $b(\mu)$ varies with the evaluation policy. At round t, let $\pi_t = \mathrm{softmax}(\theta_t)$ and draw A_t, B_t independently from π_t, conditional on the history $\mathcal{F}_t$. For a distinct-action pair, the real-valued query returns $C_t = s_{A_t}(\pi_t) - s_{B_t}(\pi_t) + \zeta_t$. If $A_t = B_t$, set $C_t = 0$ without querying. The sampled update is

$$\theta_{t+1} = \theta_t + \eta_t g_t, \qquad g_t = \tfrac{1}{2} C_t (e_{A_t} - e_{B_t}) \mathbf{1}_{\{A_t \neq B_t\}}. \tag{U}$$

Real-valued contrasts specify the feedback model: using only their signs changes the expected update (Appendix F.5).

Assumptions. We separate evaluator conditions from regularity conditions on the stochastic update. Fix $\varepsilon, L \geq 0$.

(A1) **Uniform bias dispersion.** For every $\mu \in \mathcal{P}_K$,

$$\begin{aligned}\sigma_b(\mu)^2 &:= \min_{c\in\mathbb{R}} \sum_i \mu_i (b_i(\mu) - c)^2 \\ &= \frac{1}{2}\sum_{i,j} \mu_i \mu_j (b_i(\mu) - b_j(\mu))^2 \leq \varepsilon^2.\end{aligned} \tag{1}$$

(A2) **Pairwise-bias sensitivity.** For all i, j and $\mu, \mu' \in \mathcal{P}_K$,

$$|(b_i - b_j)(\mu) - (b_i - b_j)(\mu')| \leq L\ \mathrm{TV}(\mu, \mu'), \qquad \mathrm{TV}(\mu, \mu') = \tfrac{1}{2}\|\mu - \mu'\|_1. \tag{2}$$

Both evaluator bounds hold uniformly on the closed simplex.

(A3) **Comparison noise.** On distinct pairs, $\mathbb{E}[\zeta_t \mid \mathcal{F}_t, A_t, B_t] = 0$ and $\mathbb{E}[\zeta_t^2 \mid \mathcal{F}_t, A_t, B_t] \leq \sigma^2$ for some finite σ^2; temporal independence is not required.

(A4) **Stochastic-approximation steps.** The deterministic steps satisfy $\eta_t > 0$ and

$$S_T := \sum_{t<T} \eta_t \longrightarrow \infty, \qquad \sum_{t\geq 0} \eta_t^2 < \infty. \tag{3}$$

The initial logits $\theta_0 \in \mathbb{R}^K$ are finite, so π_0 has full support. Steps need not be monotone.

Why these assumptions. Assumption (A1) ignores common offsets that comparisons cannot detect; it is weaker than absolute-bias control. Assumption (A2) applies the stability principle of performative prediction and RL to pairwise bias contrasts (Perdomo et al., 2020; Mandal et al., 2023). Appendix F.1 calibrates (ε, L) for a conformity bonus and score rescaling; the latter preserves rankings despite large dispersion relative to fixed rewards. Finitely many observations along a training trajectory do not by themselves certify these uniform bounds over the closed simplex.

Assumption (A3) needs only a conditional second moment, allowing temporal dependence; Assumption (A4) makes martingale noise converge while leaving infinite drift time. Sub-Gaussian tails enter only the finite-round bound in Proposition 12. Fixed-reward gradient bandits can converge with constant steps (Mei et al., 2023; 2024). By contrast, under policy-dependent evaluation, Appendix F.3 gives a two-action boundary example in which a constant step size prevents convergence.

2.2 Recovery Threshold and Main Result

Given the true reward gap Δ, the geometric threshold $\Psi_L(\Delta)$ is determined by the sensitivity L in Assumption (A2), and the recovery test compares it with the dispersion bound ε in Assumption (A1). Assumptions (A3)–(A4) convert this geometric margin into stochastic convergence guarantees. Under these conditions, the action count affects only the equality case.

An L-Lipschitz bias contrast cancelling a reward gap D at a segment endpoint must retain magnitude at least $(D - Lu)_+$ at distance u. Weighting by the two-group dispersion $\sqrt{u(1-u)}$ gives the scalar cancellation budget, for $D \geq 0$,

$$\Psi_L(D) := \max_{0\leq u\leq 1} \sqrt{u(1-u)}(D - Lu)_+. \tag{4}$$

Here $(x)_+ = \max\{x, 0\}$. The quantity $\Psi_L(\Delta)$ is the sampled-update recovery threshold: $\varepsilon < \Psi_L(\Delta)$ prevents cancellation of the optimal action's aggregate advantage over a mixture of alternatives. The function is strictly increasing in $D > 0$, with $\Psi_0(D) = D/2$. For $L > 0$, $\Psi_L(\Delta) < \Delta/2$, and

$$\Psi_L(\Delta) \sim \frac{2\Delta^{3/2}}{3\sqrt{3L}} \qquad \text{as } L/\Delta \to \infty.$$

Appendix A gives the exact scalar characterization and the maximizing policy weight. Figure 1(b) compares this recovery threshold with the evaluator-only limit introduced in Section 5. Theorem 1 states the recovery guarantees precisely.

Theorem 1 (Recovery threshold)**.** *Under Assumptions (A1)–(A4), update* (U) *has the following guarantees.*

1. *If $\varepsilon < \Psi_L(\Delta)$, then $\pi_t \to e_{i^\star}$ almost surely for every finite K, every admissible evaluator and noise law, and every full-support initial policy.*
2. *If $L > 0$ and $\varepsilon = \Psi_L(\Delta)$, the same universal conclusion holds for $K = 2, 3$. For every $K \geq 4$ it fails, even when the optimal action has strictly higher mean evaluator score than every alternative at every interior policy, as quantified in Theorem 2.*
3. *If $\varepsilon > \Psi_L(\Delta)$, for every finite K there are admissible evaluators and a suboptimal action w such that the probability of convergence to e_w tends to one uniformly as $\pi_{0,w} \to 1$, for any fixed schedule and variance bound.*

At equality, the zero- and positive-sensitivity cases behave differently. When $L = 0$, we have $\Psi_0(\Delta) = \Delta/2$, and an admissible noiseless evaluator can leave every initial policy unchanged. When $L > 0$, the boundary $\varepsilon = \Psi_L(\Delta)$ is the only regime in which the number of actions affects the universal recovery guarantee. Sections 3–4 explain why the transition occurs at four actions.

3 Correct Rankings with Finite Total Correction

A four-action example illustrates failure at the recovery threshold. Its design creates two competing drift scales: the optimal action's score margin shrinks with its mass $p = \pi_1$, making its sampled correction order p^2, while comparisons between the designated suboptimal action and two auxiliaries retain a fixed margin. The following bias realizes both effects while saturating the recovery boundary. Set

$$r = (1, 0, -1, -1), \quad p = \pi_1, \qquad b_1 = -(1-p)^2, \quad b_2 = b_3 = b_4 = p(1-p).$$

The pairwise bias sensitivity is $L = 1$, and

$$\sigma_b(\pi)^2 = p(1-p)^3 \leq \frac{27}{256}, \qquad \varepsilon = \Psi_1(1) = \frac{3\sqrt{3}}{16}.$$

After subtracting a common additive offset, the mean scores are $(p, 0, -1, -1)$. Hence action 1 has strictly higher mean evaluator score than every alternative whenever $p > 0$. Nevertheless, for an explicit full-support initialization near e_2 and a square-summable harmonic step schedule, the noiseless process converges to e_2 with probability greater than 0.97; Appendix C.2 gives the numerical choices. Randomness here comes entirely from action sampling.

The failure extends to every $K \geq 4$, to open sets of full-support initial policies near a suboptimal vertex, and uniformly over noise laws with a fixed variance bound.

Critical evaluator family. Fix $L, \Delta, \gamma > 0$ and $K = m + 2$ with $m \geq 2$. Let $w \neq i^\star$ be a designated suboptimal action and $\mathcal{I}_{\mathrm{aux}} = \{1, \dots, K\} \setminus \{i^\star, w\}$ the m auxiliaries. For $a \in \mathcal{I}_{\mathrm{aux}}$, set

$$r_{i^\star} = \Delta, \quad r_w = 0, \quad r_a = -\gamma, \quad h = (\Delta - L\pi_{i^\star})_+, \quad b_{i^\star} = -(1-\pi_{i^\star})h, \quad b_i = \pi_{i^\star} h \ (i \neq i^\star).$$

After subtracting a common offset, its mean scores are

$$(\min\{\Delta, L\pi_{i^\star}\},\ 0,\ -\gamma, \dots, -\gamma).$$

Theorem 2 (Failure at the recovery boundary despite correct rankings)**.** *Fix $L, \Delta, \gamma > 0$ and let $K = m + 2$ with $m \geq 2$. Choose a designated suboptimal action $w \neq i^\star$, and consider the critical evaluator family defined above. Then the following statements hold.*

1. *The evaluator satisfies Assumption (A1) with $\varepsilon = \Psi_L(\Delta)$ and Assumption (A2) with sensitivity L. Moreover,*
$$s_{i^\star}(\pi) > s_j(\pi) \qquad \text{for every } j \neq i^\star \text{ and every } \pi \in \mathcal{P}_K^\circ.$$
2. *Fix any step-size schedule satisfying Assumption (A4), any finite variance bound σ^2, and any $\rho \in (0, 1)$. For every neighborhood $\mathcal{N}$ of e_w, there exists a nonempty open set*
$$\mathcal{U} \subseteq \mathcal{N} \cap \mathcal{P}_K^\circ$$
such that, for every $\pi_0 \in \mathcal{U}$ and every noise law satisfying Assumption (A3),
$$\mathbb{P}(\pi_t \to e_w) \geq 1 - \rho.$$

3. *For every $\pi_0 \in \mathcal{U}$ and every noise law satisfying Assumption (A3), there is an event $E \subseteq \{\pi_t \to e_w\}$ with $\mathbb{P}(E) \geq 1 - \rho$ on which the optimal action is compared infinitely often. The cumulative step weight of comparisons involving $i^\star$ diverges, while the cumulative conditional drift of its logit remains finite.*

Theorem 2 rules out misranking and insufficient exposure as explanations: on failure trajectories, the optimum is compared infinitely often with divergent total step weight, yet its shrinking score margin makes cumulative conditional drift finite. Fixed-gap comparisons between w and the auxiliaries drive mass toward w while the optimal margin shrinks. At least two auxiliaries are needed: Theorem 1 guarantees recovery for $K \leq 3$, and Section 4 derives the transition from the conserved-product structure.

4 Why the Recovery Boundary Changes at Four Actions

The proof of Theorem 1 has three global ingredients, followed by local trapping. Bias geometry gives an aggregate advantage that shrinks at equality. A conserved product of noise-adjusted reciprocal coordinates forces exposure and limits how fast the optimum becomes rare. These facts yield an integrability test: correction diverges for $K \leq 3$ but can be finite for $K \geq 4$, as Theorem 2 realizes. Local trapping turns repeated approaches into convergence. Appendices A–C give full proofs.

4.1 Bias geometry gives an aggregate advantage

Key fact (Aggregate score advantage). Under Assumptions (A1)–(A2), fix an interior policy and put $p = \pi_{i^\star}$ and $q_j = \pi_j/(1-p)$ for $j \neq i^\star$. Then

$$\mathcal{A}(\pi) := s_{i^\star}(\pi) - \sum_{j \neq i^\star} q_j s_j(\pi) \geq \Delta - \Psi_L^{-1}(\varepsilon). \tag{5}$$

Thus $\mathcal{A}$ is uniformly positive if $\varepsilon < \Psi_L(\Delta)$. At $\varepsilon = \Psi_L(\Delta)$ with $L > 0$, a continuous margin a and a constant $c_0 = c_0(\Delta, L) > 0$ satisfy

$$\mathcal{A}(\pi) \geq a(p) \geq c_0 p(1-p) > 0, \qquad 0 < p < 1. \tag{6}$$

The bound concerns the sampled mixture. Below the threshold, its uniform margin gives drift toward the optimal action away from the optimal vertex. At equality the margin can vanish as $p \to 0$, so we must also control rarity. Figure 2 compares endpoint and interior cancellation costs.

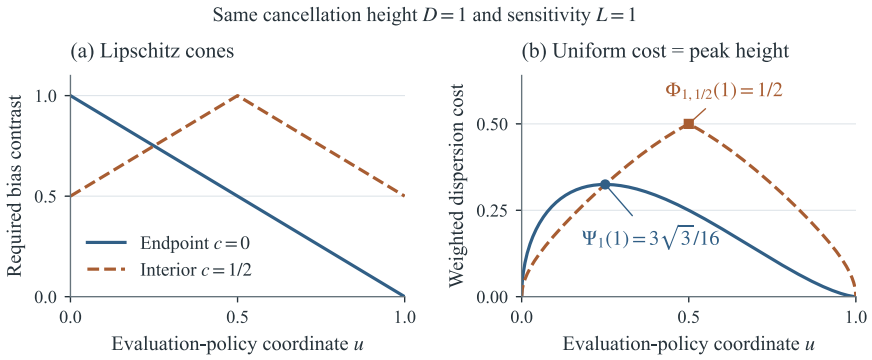

Figure 2: Bias geometry behind the recovery threshold. Cancellation is cheapest at an endpoint. Both panels use $D = L = 1$. (a) An L-Lipschitz contrast of height D must dominate a cone around its location c. (b) Weighting by $\sqrt{u(1-u)}$ gives cost $\Psi_1(1) = 3\sqrt{3}/16$ for $c = 0$, versus $1/2$ for $c = 1/2$. Appendix A proves endpoint minimality in general.

For the geometry, fix q along $\mu(u) = ue_{i^\star} + (1-u)q$ and set $h_q(u) := \sum_{j \neq i^\star} q_j(b_j(\mu(u)) - b_{i^\star}(\mu(u)))$. Dispersion bounds $\sqrt{u(1-u)}|h_q(u)|$ by ε, while sensitivity makes h_q L-Lipschitz.

Canceling a reward gap requires a bias-contrast cone, cheapest at an endpoint with cost Ψ_L. Lemma 7 in Appendix A proves the bound.

4.2 A CONSERVED PRODUCT FORCES PERSISTENT EXPOSURE

With $\bar{s}(\pi) = \sum_i \pi_i s_i(\pi)$ and $\widetilde{s}_i = s_i - \bar{s}$, averaging the ordered action samples gives

$$f_i(\pi_t) := \mathbb{E}[g_{t,i} \mid \mathcal{F}_t] = \pi_{t,i}\widetilde{s}_i(\pi_t). \tag{7}$$

The factor $\pi_{t,i}$ makes exposure necessary. By the finite-variance and step-size assumptions, $M_T = \sum_{t<T} \eta_t(g_t - f(\pi_t))$ is a convergent martingale.

Key fact (Persistent exposure). Under the assumptions of Theorem 1, define reciprocal coordinates, noise factors, and the softmax normalizer by

$$x_{t,i} = e^{-(\theta_{t,i} - M_{t,i})}, \quad \alpha_{t,i} = e^{M_{t,i}}, \quad Z_t = \sum_i e^{\theta_{t,i}}, \tag{8}$$

$$\tau_T = \sum_{t<T} \frac{\eta_t}{Z_t}. \tag{9}$$

The zero-sum update yields

$$x_{t+1,i} = x_{t,i} e^{-\eta_t f_i(\pi_t)}, \qquad \prod_i x_{t,i} = P := e^{-\sum_i \theta_{0,i}} > 0. \tag{10}$$

Almost surely, $\tau_T \to \infty$ and $x_{t,i} \le C(\omega)(1 + \tau_t)$ for every action i. Consequently,

$$\sum_t \eta_t \pi_{t,i}(1 - \pi_{t,i}) = \infty \qquad \text{for every action } i. \tag{11}$$

The realized step-weighted comparisons involving each action also have divergent total weight. The product forces persistent exposure, ruling out insufficient comparisons as the cause of Theorem 2. A vanishing score margin can still make cumulative correction finite. Appendix B proves clock divergence and the realized comparison law.

4.3 WHY THREE ACTIONS RECOVER AND FOUR CAN FAIL

At or below the threshold, $\mathcal{A} \ge 0$ makes $x_{t,i^\star}$ nonincreasing, so its total centered-score correction is finite. Below the threshold, the uniform margin in (5) and the divergent clock force repeated approaches to the optimum.

Key fact (Dimension-dependent lower bound). At $\varepsilon = \Psi_L(\Delta)$ with $L > 0$, put $p_t = \pi_{t,i^\star}$. Almost surely,

$$p_t \ge c(\omega)(1 + \tau_t)^{-(K-2)}. \tag{12}$$

If $\limsup_t p_t < 1$, the accumulated centered-score correction dominates, up to a finite initial segment and a positive path-dependent factor, the integral

$$\int_0^\infty (1 + \tau)^{-(K-2)}\, d\tau. \tag{13}$$

In the conserved product, $K - 2$ counts the factors that can grow after excluding the decreasing optimal coordinate and one comparator. The integral diverges for $K = 2, 3$, forcing an approach to the optimum, but converges for $K \ge 4$. Theorem 2 realizes finite correction despite persistent comparisons. Appendix B proves the bound and divergence argument.

Key fact (Local trapping near the optimum). Suppose $\varepsilon < \Psi_L(\Delta)$, or $L > 0$ and $\varepsilon = \Psi_L(\Delta)$. Near $e_{i^\star}$, all pairwise score gaps favoring $i^\star$ are nonnegative, and vanishing endpoint gaps have a quantitative lower bound. The optimal log-odds drift is positive on every nonempty shell away from $e_{i^\star}$. Since the martingale term converges and the nonlinear remainder is summable, $\limsup_{t\to\infty} \pi_{t,i^\star} = 1$ implies $\pi_t \to e_{i^\star}$ almost surely. The global argument gives repeated approaches to the optimal vertex; local trapping converts them into convergence. Appendix B proves the shell bound and controls the perturbations.

4.4 THE CRITICAL EVALUATOR REALIZES THE INTEGRABLE REGIME

Key fact (Finite correction in the critical family). For the critical evaluator of Theorem 2, let $K = m + 2$ with $m \geq 2$. On its high-probability failure event, $x_{t,i^\star}$ converges to a positive limit and each auxiliary reciprocal coordinate satisfies $x_{t,a} \sim \gamma \tau_t$. Consequently, as $t \to \infty$,

$$\pi_{t,i^\star} \asymp \tau_t^{-m}, \qquad \sum_t \eta_t \mathbb{E}[g_{t,i^\star} \mid \mathcal{F}_t] < \infty.$$

This realizes the integrable $m = K - 2$ tail: fixed-gap auxiliary comparisons favor w while the margin for $i^\star$ shrinks with its mass. Appendix C constructs the event; Appendix D.2 gives exact rates.

5 RATES, COMPARISONS, AND BALANCED QUERIES

5.1 EXACT RATES ON SUCCESSFUL TRAJECTORIES

On successful trajectories, both suboptimal mass and comparison frequencies depend on the evaluator's endpoint score gaps.

Theorem 3 (Sample-path equivalents)**.** *Under Assumptions (A1)–(A4), on the event $\{\pi_t \to e_{i^\star}\}$, suppose $\Gamma_j := s_{i^\star}(e_{i^\star}) - s_j(e_{i^\star}) > 0$ for every $j \neq i^\star$. Then almost surely on that event,*

$$S_T \pi_{T,j} \longrightarrow \frac{1}{K\Gamma_j}, \qquad S_T(1 - \pi_{T,i^\star}) \longrightarrow \frac{1}{K} \sum_{j \neq i^\star} \Gamma_j^{-1}. \tag{14}$$

Strictly below the boundary, the event has probability one and all $\Gamma_j > 0$.

Every pair is compared infinitely often: optimal-action pairs occur at order $\sum_{t<T} S_t^{-1}$ and alternative pairs at order $\sum_{t<T} S_t^{-2}$; both sums diverge. Appendix D gives exact constants, power-step conversions, and boundary rates when endpoint gaps vanish.

Full-information multiplicative weights removes the action-sampling factor (proof in Appendix A).

Proposition 4 (Full-information baseline)**.** *Suppose the learner observes $\widehat{s}_t = s(\pi_t) + \xi_t$ each round, where $\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0$ and $\mathbb{E}[\mathrm{osc}(\xi_t)^2 \mid \mathcal{F}_t] \leq \sigma_{\mathrm{FI}}^2$, and updates $\pi_{t+1,i} \propto \pi_{t,i} \exp(\eta_t \widehat{s}_{t,i})$. Under Assumptions (A1), (A2), and (A4), if $\varepsilon < \Psi_L(\Delta)$, put $\kappa := \Delta - \Psi_L^{-1}(\varepsilon) > 0$. Then almost surely $\liminf_T S_T^{-1} \log[\pi_{T,i^\star}/(1 - \pi_{T,i^\star})] \geq \kappa$ and $1 - \pi_{T,i^\star} \leq C(\omega) e^{-\kappa S_T}$ for a finite random $C(\omega)$. If $L > 0$ and $\varepsilon = \Psi_L(\Delta)$, recovery still holds for every finite K, without a uniform positive exponential rate. Above the boundary, a binary noiseless counterexample exists; at $L = 0$ and $\varepsilon = \Delta/2$, score cancellation can prevent recovery.*

5.2 DECOUPLING THE EVALUATION POLICY FROM THE ACTOR POLICY

The failure of update (U) leaves a second question: is the optimum hidden by the evaluator, or does actor-coupled sampling prevent the update from using information that remains available? To separate these effects, we let the learner choose the evaluation policy independently of the actor policy. The resulting balanced-query oracle serves as an information benchmark.

The actor samples A_t, B_t from π_t, while the evaluator scores them at $\mu^{A_t B_t} = (e_{A_t} + e_{B_t})/2$. Each queried action has weight $1/2$. Set $c_{ij} = s_i(\mu^{ij}) - s_j(\mu^{ij})$ for $i \neq j$ and $c_{ii} = 0$. By (1),

$$|c_{ij} - (r_i - r_j)| \leq 2\varepsilon.$$

Balanced contrasts are independent of L and actor probabilities, but rare-action sampling still attenuates updates. Set $\pi_t^{\min} := \min_i \pi_{t,i}$ and let $C_t^{\mathrm{BQ}} = c_{A_t B_t} + \zeta_t$ on distinct pairs and zero on duplicates. The corrected update is

$$\theta_{t+1} = \theta_t + \frac{\eta_t \pi_t^{\min} C_t^{\mathrm{BQ}}}{2} \left(\frac{e_{A_t}}{\pi_{t,A_t}} - \frac{e_{B_t}}{\pi_{t,B_t}} \right) \mathbf{1}_{\{A_t \neq B_t\}}. \tag{BQ}$$

Inverse probabilities cancel sampling attenuation in the mean; $\pi_t^{\min}/(2\pi_{t,i}) \leq 1/2$ bounds each multiplier. Each round uses two draws and at most one query.

Theorem 5 (Balanced-query recovery and identification limit)**.** *Under Assumptions (A1), (A3), and (A4), without requiring Assumption (A2), update* (BQ) *converges almost surely to $e_{i^\star}$ for every finite K and every full-support initialization whenever $\varepsilon < \Delta/2$.*

Conversely, under Assumption (A1) alone, if $\varepsilon \geq \Delta/2$, no method using only evaluator feedback can universally identify the optimal action or converge to it, even with arbitrary adaptive evaluation policies and unlimited queries.

Importance correction gives favorable log-odds drift below $\Delta/2$, though $\pi_t^{\min}$ can slow recovery (Appendix E). At $\Delta/2$, rewards $(\Delta, 0)$ and $(0, \Delta)$ can yield identical scores $(\Delta/2, \Delta/2)$ under every query. Thus $\Psi_L(\Delta) < \Delta/2$ for $L > 0$ separates update failure from evaluator-only indistinguishability. Classical dueling bandits assume fixed preference probabilities (Zoghi et al., 2014).

With conditionally sub-Gaussian noise, uniform balanced queries identify the optimum within $O(K^2[1 + \nu^2/(\Delta - 2\varepsilon)^2] \log(K/\beta))$ rounds with probability at least $1 - \beta$ (Proposition 12). Appendix G gives finite-horizon simulations.

6 Discussion and Conclusion

This work identifies a sharp distinction between the information contained in policy-dependent evaluator feedback and the ability of a sampled update to accumulate that information. For pairwise self-rewarded softmax updates, the worst-case recovery threshold is $\Psi_L(\Delta)$. Below this threshold, recovery is universal. At the boundary with $L > 0$, recovery still holds for two or three actions, whereas four or more actions can fail even when the evaluator ranks the optimal action first at every interior policy. On such failure trajectories, the optimal action is compared infinitely often, yet the cumulative conditional drift toward it remains finite. The failure therefore does not require incorrect rankings or a lack of comparisons; it can arise because favorable feedback becomes too weak as the optimal action becomes rare.

The comparison baselines separate this update failure from a lack of identifying information. A full-information update still recovers at the sampled-update boundary for every finite action set. Balanced queries with an importance-corrected update extend the recovery guarantee to $\varepsilon < \Delta/2$ without a sensitivity assumption. At $\varepsilon \geq \Delta/2$, evaluator feedback alone cannot universally identify the optimal action. Thus, for $L > 0$, there is a regime between $\Psi_L(\Delta)$ and $\Delta/2$ in which the evaluator retains enough information to identify the optimum, but the sampled update need not recover it.

Our model deliberately isolates this feedback mechanism. It assumes a finite action set, fixed true rewards, real-valued comparisons, and uniform dispersion and sensitivity bounds on the evaluator over the closed simplex. These uniform bounds are structural conditions on the evaluator; observations from finitely many visited policies do not by themselves certify them. Extending the analysis to learned policies with function approximation, including PPO- and GRPO-style training, is an important next step. Other open problems include finite-round recovery guarantees and a general characterization of rates when endpoint score gaps vanish.

The central lesson is that favorable evaluator feedback is not sufficient by itself: recovery also depends on whether the learning dynamics accumulate that feedback at a non-summable rate.

AI USE STATEMENT

OpenAI Codex (GPT-6) was used to polish the manuscript's language and presentation. It also helped explore the critical evaluator family used to construct the four-action counterexample at the recovery boundary.

REPRODUCIBILITY STATEMENT

Complete proofs are provided in Appendices A–F, and Appendix G specifies the simulation protocol and finite-horizon experiments. The supplementary source includes the code and inputs required to reproduce the simulations and figures.

REFERENCES

Evan Assmus, Qining Zhang, and Lei Ying. SP3O: Reinforcement learning from segment preferences without reward modeling, 2026. URL https://arxiv.org/abs/2608.02951v1. Version 1, 3 August 2026.

Debabrota Basu, Udvas Das, Brahim Driss, and Uddalak Mukherjee. Performative policy gradient: Optimality in performative reinforcement learning, 2026. URL https://arxiv.org/abs/2512.20576v2. Version 2, 2 February 2026; first submitted in 2025.

Dorian Baudry, Emmeran Johnson, Simon Vary, Ciara Pike-Burke, and Patrick Rebeschini. Does stochastic gradient really succeed for bandits? In *Advances in Neural Information Processing Systems*, volume 38, pp. 125963–126030, 2025. doi: 10.52202/085713-3787.

Peter Chen, Xi Chen, Wotao Yin, and Tianyi Lin. A zeroth-order paradigm for LLM preference alignment, 2026. URL https://arxiv.org/abs/2609.19144v1. Version 1, 16 September 2026; extends the NeurIPS 2025 conference version.

Gabriele Farina and Juan Carlos Perdomo. The stability of online algorithms in performative prediction, 2026. URL https://arxiv.org/abs/2602.24207. arXiv:2602.24207.

Shi Fu, Yingjie Wang, Shengchao Hu, Peng Wang, and Dacheng Tao. Why self-rewarding works: Theoretical guarantees for iterative alignment of language models, 2026. URL https://arxiv.org/abs/2601.22513v2. Version 2, 3 February 2026.

Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23642–23680, 2023. URL https://proceedings.mlr.press/v202/mandal23a.html.

Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19339–19351, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a12f69495f41bb3b637ba1b6238884d6-Abstract.html.

Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvári, and Dale Schuurmans. Stochastic gradient succeeds for bandits. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24325–24360, 2023. URL https://proceedings.mlr.press/v202/mei23a.html.

Jincheng Mei, Bo Dai, Alekh Agarwal, Sharan Vaswani, Anant Raj, Csaba Szepesvári, and Dale Schuurmans. Small steps no more: Global convergence of stochastic gradient bandits for arbitrary learning rates. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2371.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609, 2020. URL https://proceedings.mlr.press/v119/perdomo20a.html.

Chuyi Tan, Peiwen Yuan, Xinglin Wang, Yiwei Li, Shaoxiong Feng, Yueqi Zhang, Jiayi Shi, Ji Zhang, Boyuan Pan, Yao Hu, and Kan Li. Breaking the self-confirming loop: Diagnosing and mitigating systemic reward bias in self-rewarding RL, 2026. URL https://arxiv.org/abs/2510.08977v2. Version 2, 2 June 2026; first submitted in 2025.

Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. CREAM: Consistency regularized self-rewarding language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Vf6RDObyEF.

Tianhao Wu, Banghua Zhu, Ruoyu Zhang, Zhaojin Wen, Kannan Ramchandran, and Jiantao Jiao. Pairwise proximal policy optimization: Harnessing relative feedback for LLM alignment, 2023. URL https://arxiv.org/abs/2310.00212v3. Version 3, 10 October 2023.

Yunhao Yang, Yuexin Bian, Yunjie Tian, Di Fu, Tianjin Huang, Yuanyuan Shi, Ziang Xiao, Nuno Vasconcelos, and Yijiang Li. Co-RL: Unsupervised reasoning emerges from diverse cohort in multi-agent RL, 2026. URL https://arxiv.org/abs/2608.17253v2. Version 2, 19 August 2026.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E. Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923, 2024. URL https://proceedings.mlr.press/v235/yuan24d.html.

Chenyu Zhou. More convincing, not more correct: Self-play reward hacking of reference-free LLM judges, 2026. URL https://arxiv.org/abs/2607.05904.

Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten de Rijke. Relative upper confidence bound for the K-armed dueling bandit problem. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 10–18, 2014. URL https://proceedings.mlr.press/v32/zoghi14.html.

A Scalar Bias Geometry and the Full-information baseline

We prove the scalar bounds used throughout the paper and the full-information comparison in Figure 1. All simplex, bias, and step-size notation is as in Section 2.

Figure 2 illustrates the two cancellation locations used below; this appendix proves the corresponding scalar bounds exactly.

A.1 Exact scalar envelopes

For $\omega(u) = \sqrt{u(1-u)}$, define the scalar envelope

$$U_{\varepsilon,L}(c) = \inf_{0<u<1} \left\{ \frac{\varepsilon}{\omega(u)} + L|u-c| \right\}. \tag{15}$$

Lemma 6. *Let $L, \varepsilon \geq 0$ and $\omega(u) = \sqrt{u(1-u)}$. For an L-Lipschitz function $h : [0,1] \to \mathbb{R}$ satisfying $\omega(u)|h(u)| \leq \varepsilon$, the exact maximal value at c is $U_{\varepsilon,L}(c)$ in* (15)*. Equivalently, $h(c) \geq D \geq 0$ is feasible if and only if*

$$\Phi_{L,c}(D) := \max_u \omega(u)(D - L|u-c|)_+ \leq \varepsilon.$$

The minimum of $\Phi_{L,c}(D)$ over c is $\Psi_L(D)$, attained at the endpoints. For $D, L > 0$, $\Phi_{L,c}(D)$ is strictly increasing on $0 \leq c \leq 1/2$ and symmetric about $1/2$. In particular, $|h(c)| \leq \Psi_L^{-1}(\varepsilon)$.

For $L, \Delta > 0$ and $\varepsilon = \Psi_L(\Delta)$, the function $a(p) = \Delta - U_{\varepsilon,L}(p)$ is continuous, nonnegative on $[0,1]$, strictly positive inside, and satisfies $a(p) \geq c_0 p(1-p)$ for some $c_0 > 0$ depending only on Δ, L.

Proof. If $h(c) \geq D$, then $h(u) \geq D - L|u-c|$. Wherever this quantity is positive, it also gives a lower bound for $|h(u)|$, which proves necessity. The function $h(u) = (D - L|u-c|)_+$ proves sufficiency. The same argument applies to $-h$.

Every admissible h obeys $|h(c)| \leq \varepsilon/\omega(u) + L|u-c|$ for each interior u. Conversely, the infimum of these upper bounds, viewed as a function of c, is nonnegative and L-Lipschitz, and is at most $\varepsilon/\omega(c)$ for interior c. It is therefore admissible and attains the bound, proving that it is the exact envelope. When $\varepsilon = 0$, continuity forces $h = 0$, and the same infimum is zero including at the endpoints.

For $L, D > 0$ and $0 \leq c \leq 1/2$, the maximization defining Φ can be restricted to $u \in [c, 1/2]$. A point $u < c$ cannot improve on c, since neither its weight nor its cone height is larger. Reflecting $u > 1/2$ to $1-u$ preserves its weight and cannot increase its distance from c; a reflected point below c again cannot improve on c. Now take $c_1 < c_2 \leq 1/2$ and an optimizer $u_1 \in [c_1, 1/2]$. The optimum is positive. If $u_1 \geq c_2$, moving the cone center from c_1 to c_2 strictly increases the positive height at u_1. If $u_1 < c_2$, the point $u = c_2$ has strictly larger weight and no smaller height than the original optimizer. This proves strict monotonicity for $L, D > 0$. Reflection proves symmetry. If $L = 0$, every cost is $D/2$.

Thus the minimum cost is attained at the endpoints and equals $\Psi_L(D)$. This function is continuous, vanishes at zero, and diverges with D. To prove strict increase, evaluate a larger D at an interior maximizer for a smaller positive D. Positivity gives the comparison with $D = 0$. Its inverse is therefore well-defined, with value zero at $\varepsilon = 0$. The endpoint cone attains the global bound.

Differentiating $u(1-u)(D - Lu)^2$ on its positive part gives

$$4Lu^2 - (2D + 3L)u + D = 0.$$

The smaller root is

$$u_0(D, L) = \frac{2D}{2D + 3L + \sqrt{4D^2 - 4DL + 9L^2}}. \tag{16}$$

It lies in $(0, 1/2)$ and below D/L. The other root cannot maximize the positive part. For $L = 0$, the maximizer is $1/2$. Expanding (16) for $L/D \to \infty$ yields $u_0 \sim D/(3L)$ and the asymptotic formula stated after Theorem 1. Also $\Psi_L(D) < D/2$ for $L > 0$, since the maximal weight $1/2$ and cone height D cannot be attained at the same point.

At $\varepsilon = \Psi_L(\Delta)$, strict endpoint minimality gives $\Phi_{L,p}(\Delta) > \varepsilon$ for $0 < p < 1$, hence $U_{\varepsilon,L}(p) < \Delta$. Set $u_0 = u_0(\Delta, L)$. For $p \le u_0$, evaluating the infimum (15) at u_0 gives

$$U_{\varepsilon,L}(p) \le \frac{\varepsilon}{\omega(u_0)} + L(u_0 - p) = \Delta - Lp.$$

Symmetry gives $a(p) \ge L(1-p)$ for $p \ge 1 - u_0$. On $[u_0, 1-u_0]$, the continuous positive function $a(p)/(p(1-p))$ has a positive minimum. These bounds give the required constant c_0. □

Lemma 7 (Consequences for the evaluator)**.** *Under* (1)–(2)*, the score oscillation is at most* $G = R_r + 2\varepsilon + L$*. Every interior policy satisfies* (5)*, and*

$$s_{i^\star}(e_{i^\star}) - s_j(e_{i^\star}) \ge \Delta_j - \Psi_L^{-1}(\varepsilon).$$

At $\varepsilon = \Psi_L(\Delta)$ *with* $L > 0$*,* (6) *holds.*

Proof. At $\mu^{ij} = (e_i + e_j)/2$, the variance in (1) equals $(b_i - b_j)^2/4$, so $|(b_i - b_j)(\mu^{ij})| \le 2\varepsilon$. The sensitivity bound and $\mathrm{TV}(\mu, \mu^{ij}) \le 1$ give the oscillation bound.

For a distribution q supported on actions other than $i^\star$, the segment $\mu(u) = ue_{i^\star} + (1-u)q$ has $\mathrm{TV}(\mu(u), \mu(v)) = |u - v|$. Hence the contrast $h_q(u) = \sum_j q_j(b_j - b_{i^\star})(\mu(u))$ is L-Lipschitz. Conditioning on whether the action is $i^\star$ gives

$$\sigma_b(\mu(u))^2 = u(1-u)h_q(u)^2 + (1-u)\,\mathrm{Var}_q(b_j(\mu(u))). \tag{17}$$

Hence $\omega(u)|h_q(u)| \le \varepsilon$. Lemma 6 bounds this contrast by $\Psi_L^{-1}(\varepsilon)$, and more sharply by $U_{\varepsilon,L}(u)$. Since $r_{i^\star} - \sum_j q_j r_j \ge \Delta$, this proves both aggregate assertions. Taking $q = e_j$ and $u = 1$ gives the endpoint gap bound. □

A.2 Log-sum-exp remainders under finite second moments

Write $\mathrm{osc}(z) = \max_i z_i - \min_i z_i$. For a probability vector q and a finite vector z, define

$$\mathcal{R}_{\mathrm{lse}}(q, z) = \log \sum_j q_j e^{z_j} - \sum_j q_j z_j.$$

Then

$$0 \le \mathcal{R}_{\mathrm{lse}}(q, z) \le \frac{1}{8}\,\mathrm{osc}(z)^2. \tag{18}$$

To see the upper bound, let X take the values z_j with probabilities q_j, and put $H(\lambda) = \log \mathbb{E}e^{\lambda X} - \lambda \mathbb{E}X$. Under the exponentially tilted finite law, $H''(\lambda) = \mathrm{Var}_\lambda(X) \le \mathrm{osc}(z)^2/4$. Integrating twice from $H(0) = H'(0) = 0$ proves (18). Jensen's inequality gives nonnegativity. The inequality holds for each realized finite vector, so it requires no exponential moment of the observation noise.

Proof of Proposition 4. Let $p_t = \pi_{t,i^\star}$, $q_{t,j} = \pi_{t,j}/(1-p_t)$ and $z_t = \log(p_t/(1-p_t))$. The exact increment is

$$z_{t+1} - z_t = \eta_t\big(\mathcal{A}(\pi_t) + \nu_t\big) - \mathcal{R}_{\mathrm{lse},t}, \quad \nu_t = \xi_{t,i^\star} - \sum_j q_{t,j}\xi_{t,j}, \quad \mathcal{R}_{\mathrm{lse},t} = \mathcal{R}_{\mathrm{lse}}(q_t, \eta_t \widehat{s}_{t,-i^\star}).$$

The martingale $\sum_{t<T} \eta_t \nu_t$ converges in L^2 and almost surely, because $|\nu_t| \le \mathrm{osc}(\xi_t)$. By (18),

$$\mathbb{E}[\mathcal{R}_{\mathrm{lse},t}] \le \tfrac{1}{4}\eta_t^2(G^2 + \sigma_{\mathrm{FI}}^2), \qquad \sum_t \mathcal{R}_{\mathrm{lse},t} < \infty \quad \text{almost surely.}$$

Thus $z_T = z_0 + \sum_{t<T} \eta_t \mathcal{A}(\pi_t) + R_T$ with R_T convergent. Strictly below the recovery threshold, Lemma 7 gives $\mathcal{A} \ge \kappa$, which proves both rate statements because R_T is bounded. At $\varepsilon = \Psi_L(\Delta)$ with $L > 0$, $\mathcal{A} \ge a(p) > 0$ in the interior. The nondecreasing drift sum implies that z_T either converges to a finite value or tends to $+\infty$. A finite limit would keep p_t in a compact subinterval of the interior, where a has a positive minimum, contradicting $\sum \eta_t = \infty$.

To show that convergence can be slower at the boundary, take two actions with rewards $(\Delta, 0)$, write $p = \pi_1$, and let

$$h(p) = (\Delta - L(1-p))_+, \quad b_1 = -(1-p)h, \quad b_2 = ph.$$

Its exact uniform bias-dispersion bound is $\Psi_L(\Delta)$, and the score gap is $\min\{\Delta, L(1-p)\}$. In the noiseless update, $p_t \to 1$. Once the second branch applies, the odds $o_t = p_t/(1-p_t)$ satisfy

$$o_{t+1} = o_t \exp\left(\frac{\eta_t L}{1+o_t}\right), \qquad o_{t+1} - o_t = \eta_t L \frac{o_t}{1+o_t} + O(\eta_t^2).$$

The remainder bound is uniform for large t. Summation gives $o_T/(LS_T) \to 1$ and hence $1 - p_T \sim (LS_T)^{-1}$.

For $\varepsilon > \Psi_L(\Delta)$ and $L > 0$, choose $\Delta < c < \Delta + L$ with $\Psi_L(c) < \varepsilon$, and instead use $h(p) = (c - Lp)_+$. The score gap is $\Delta - h(p)$, negative for $0 < p < (c-\Delta)/L$ and positive above that point. The binary log odds move in the direction of this score gap. In either interval, they cannot converge to a finite value, so every initialization in the lower interval converges to the suboptimal vertex. Nonzero bounded centered contrast noise of magnitude smaller than the negative gap in a smaller interval preserves this convergence to a suboptimal vertex pathwise. For $L = 0$, a constant adverse bias $h = \Delta$ cancels scores at $\varepsilon = \Delta/2$; choosing $\Delta < h < 2\varepsilon$ makes the suboptimal action have the largest mean score when $\varepsilon > \Delta/2$. □

B Stochastic Dynamics and the Convergence Guarantees

We analyze (U) throughout this appendix. All pathwise arguments use a common probability-one event on which the finitely many coordinate martingales converge.

B.1 Moments, martingale removal, and persistent comparisons

The ordered-pair identity

$$\mathbb{E}[g_{t,i} \mid \mathcal{F}_t] = \tfrac{1}{2}\sum_j \pi_{t,i}\pi_{t,j}(s_i - s_j) + \tfrac{1}{2}\sum_j \pi_{t,j}\pi_{t,i}(s_i - s_j) = \pi_{t,i}(s_i - \bar{s})$$

uses centering conditional on the *sampled pair*. Lemma 7 gives $\mathrm{osc}(s) \le G := R_r + 2\varepsilon + L$. With $V = G^2 + \sigma^2$, the contrast C_t has second moment at most V conditional on any distinct-action pair. For such a pair, $\|g_t\|^2 = C_t^2/2$. Therefore

$$\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t] \le V/2, \qquad \mathbb{E}[g_{t,i}^2 \mid \mathcal{F}_t] \le (V/2)\pi_{t,i}(1-\pi_{t,i}). \tag{19}$$

Write $V_\eta := \sum_{t\ge 0}\eta_t^2 < \infty$. Consequently, for $M_T = \sum_{t<T}\eta_t(g_t - f(\pi_t))$,

$$\mathbb{E}\|M_\infty\|^2 \le \tfrac{1}{2}VV_\eta, \qquad \mathbb{E}\sup_T \|M_T\|^2 \le 2VV_\eta, \qquad \sum_t \eta_t^2\|g_t\|^2 < \infty \quad \text{a.s.} \tag{20}$$

Orthogonality of martingale increments and the L^2 maximal inequality give the first two bounds. Summing expectations of nonnegative variables gives the last assertion.

Since $\sum_i g_{t,i} = \sum_i f_i = 0$, both $\sum_i \theta_{t,i}$ and $\sum_i(\theta_{t,i} - M_{t,i})$ equal $\sum_i \theta_{0,i}$. Define x, α, Z by (8), τ by (9), and put $\bar{\eta}_t = \eta_t/Z_t$. The product identity (10) is exact. Also

$$Z_t \ge K \exp\left(K^{-1}\sum_i \theta_{0,i}\right) > 0,$$

so $\bar{\eta}_t \to 0$ and $\sup_t \bar{\eta}_t < \infty$. On each such path, $\alpha_{t,i}$ and $\alpha_{t,i}^{-1}$ are bounded. Since $|\widetilde{s}_i| \le G$ and $\eta_{\max} := \sup_t \eta_t < \infty$,

$$|x_{t+1,i} - x_{t,i}| \le x_{t,i}\eta_t|f_i|e^{\eta_{\max}G} \le C(\omega)\bar{\eta}_t. \tag{21}$$

Lemma 8 (Divergent clock and distinct-action comparison weight)**.** *For every admissible evaluator,*

$$\tau_T \to \infty, \quad x_{t,i} \le C(\omega)(1+\tau_t), \quad \sum_t \eta_t\pi_{t,i} = \infty, \qquad \sum_t \eta_t\pi_{t,i}(1-\pi_{t,i}) = \infty, \qquad \forall i$$

If $I_{t,i}$ indicates a distinct-action pair containing i, then

$$\frac{\sum_{t<T}\eta_t I_{t,i}}{2\sum_{t<T}\eta_t\pi_{t,i}(1-\pi_{t,i})} \to 1 \quad \textit{almost surely}.$$

Proof. If $\sum \bar{\eta}_t < \infty$, (21) makes every $x_{t,i}$ converge to a finite limit. Since the product is fixed and positive, every limit is positive. Convergence of M_t then implies that every $\theta_{t,i}$ has a finite limit and Z_t is bounded, contradicting $\sum \eta_t = \infty$. Summing (21) gives the stated growth bound.

Now $\eta_t \pi_{t,i} = \alpha_{t,i} \bar{\eta}_t / x_{t,i} \geq c(\omega) \bar{\eta}_t / (1 + \tau_t)$, whose sum diverges by comparison with $\int_0^\infty (1 + u)^{-1} \, du$. Suppose instead that $\sum_t \eta_t \pi_{t,i}(1 - \pi_{t,i}) < \infty$. Since $|\widetilde{s}_i| \leq G(1 - \pi_i)$, the drift of $\theta_{t,i}$ is absolutely summable, and its martingale converges. Hence $\theta_{t,i}$ has a finite limit. Conservation of the logit sum and the arithmetic-geometric mean inequality applied to the other logits imply

$$\sum_{j \neq i} e^{\theta_{t,j}} \geq (K - 1) \exp\left(\frac{\sum_j \theta_{0,j} - \theta_{t,i}}{K - 1} \right).$$

It follows that $1 - \pi_{t,i}$ is eventually bounded below by a positive path-dependent constant. The assumed finiteness of the distinct-action comparison weight would then contradict $\sum_t \eta_t \pi_{t,i} = \infty$, which we have just proved.

Finally, $\mathbb{E}[I_{t,i} \mid \mathcal{F}_t] = 2\pi_{t,i}(1 - \pi_{t,i})$. The martingale $\sum_{t<T} \eta_t (I_{t,i} - \mathbb{E}[I_{t,i} \mid \mathcal{F}_t])$ converges by $\sum \eta_t^2 < \infty$. Dividing by the divergent denominator proves the ratio limit. □

B.2 A LOCAL CONVERGENCE ARGUMENT FOR THE ACTUAL ITERATES

Fix an action a, write $p = \pi_a$, $q_j = \pi_j / (1 - p)$, and let $\ell_a = \theta_a - \log \sum_{j \neq a} e^{\theta_j}$. Set

$$d_j = s_a - s_j, \qquad \widetilde{s}_a = \sum_{j \neq a} \pi_j d_j.$$

Using $f_a = p\widetilde{s}_a$ and $f_j = \pi_j(\widetilde{s}_a - d_j)$ gives the exact first-order drift

$$\mathcal{D}_a := f_a - \sum_{j \neq a} q_j f_j = \left(p - \frac{\sum_{j \neq a} \pi_j^2}{1 - p} \right) \widetilde{s}_a + \frac{1}{1 - p} \sum_{j \neq a} \pi_j^2 d_j. \tag{22}$$

If all $d_j \geq 0$ and $p > 1/2$, then $\sum_{j \neq a} \pi_j^2 \leq (1 - p)^2$ and

$$\mathcal{D}_a \geq (2p - 1)\widetilde{s}_a \geq 0. \tag{23}$$

Applying (18) to each stochastic increment gives the exact decomposition

$$\ell_{a,T} = \ell_{a,0} + \sum_{t<T} \eta_t \mathcal{D}_a(\pi_t) + \mathcal{M}_{a,T} - \sum_{t<T} \mathcal{R}_{a,t}, \tag{24}$$

where $\mathcal{M}_a$ is a martingale and $\mathcal{R}_{a,t} \geq 0$. The coefficient vector in the first-order noise term is $e_a - q$, of squared norm at most 2, so

$$\mathbb{E}[\mathcal{M}_{a,\infty}^2] \leq V V_\eta, \qquad \mathbb{E} \sum_t \mathcal{R}_{a,t} \leq V V_\eta / 8.$$

For the second bound, use $\mathcal{R}_{a,t} \leq \eta_t^2 \operatorname{osc}(g_{t,-a})^2 / 8 \leq \eta_t^2 \|g_t\|^2 / 4$ and (19). Thus the perturbation in (24) converges almost surely.

Lemma 9 (Local trapping)**.** *Suppose that on $\{\pi_a \geq p_0\}$ for some $p_0 > 1/2$, the drift $\mathcal{D}_a$ is nonnegative and has a strictly positive lower bound on each nonempty shell $p_0 \leq \pi_a \leq 1 - \delta$, $\delta > 0$. Then $\limsup_t \pi_{t,a} = 1$ implies $\pi_t \to e_a$ almost surely. If, more specifically, $s_a - s_j \geq d_0 > 0$ on this neighborhood for every $j \neq a$, then, uniformly over admissible noise laws with a fixed variance bound,*

$$\lim_{u \downarrow 0} \inf_{\pi_{0,a} \geq 1 - u} \mathbb{P}(\pi_t \to e_a) = 1.$$

Proof. Let $\mathcal{Q}_T = \mathcal{M}_{a,T} - \sum_{t<T} \mathcal{R}_{a,t}$, which converges. Let $\ell_0^\circ = \log(p_0/(1 - p_0))$. For all sufficiently large t, the oscillation of $\mathcal{Q}$ over all subsequent times is smaller than 1. If $\limsup \pi_{t,a} = 1$, choose such a time with $\ell_{a,t} > \ell_0^\circ + 2$. Until a first exit below ℓ_0°, the drift sum is nonnegative, and (24) keeps the log odds above $\ell_0^\circ + 1$, including the proposed exit step. Thus no exit occurs. Thereafter, the drift sum is nondecreasing and $\mathcal{Q}_T$ converges, so the log odds have a finite limit

or tend to $+\infty$. A finite limit would lie on a shell with a positive drift lower bound, contradicting $\sum \eta_t = \infty$.

For the uniform assertion, the local gap $d_j \geq d_0$ gives $\mathcal{D}_a \geq (2p-1)d_0(1-p)$ by (23), satisfying the shell condition. For any $d > 0$, the maximal and Markov inequalities yield

$$\mathbb{P}\left(\sup_T |\mathcal{M}_{a,T}| + \sum_t \mathcal{R}_{a,t} \geq d\right) \leq \frac{16VV_\eta}{d^2} + \frac{VV_\eta}{4d}.$$

If $\ell_{a,0} > \ell_0^\circ + d$, the complement of this event prevents a first exit at any time. The same monotone-drift argument then proves convergence. Let $d \to \infty$ as $\pi_{0,a} \to 1$. These bounds depend only on the score and noise bounds and the schedule. They are uniform over the noise kernel and the distribution of initial mass among the other actions. □

B.3 Convergence below the threshold and local score gaps

Strictly below the recovery threshold, $\widetilde{s}_{i^\star} = (1-p)\mathcal{A} \geq \kappa(1-p)$ by Lemma 7, with $\kappa > 0$. From $1 - e^{-z} \geq ze^{-\eta_{\max}G}$ for $0 \leq z \leq \eta_{\max}G$ and (10),

$$x_{t,i^\star} - x_{t+1,i^\star} \geq c(\omega)\bar{\eta}_t \widetilde{s}_{i^\star}(\pi_t).$$

The left side telescopes and $x_{t,i^\star}$ is nonnegative; hence $\sum_t \bar{\eta}_t(1-p_t) < \infty$. Lemma 8 forces $\limsup_t p_t = 1$. At $e_{i^\star}$ all endpoint gaps are at least κ. By continuity, they remain positive in a neighborhood of $e_{i^\star}$. Lemma 9 proves Theorem 1(1).

Lemma 10 (Endpoint saturation at the threshold)**.** *If $L > 0$ and $\varepsilon = \Psi_L(\Delta)$, there are $c_1 > 0$ and $\delta_0 \in (0, 1/2)$ such that $s_{i^\star}(\pi) - s_j(\pi) \geq c_1\pi_j$ whenever $\pi_{i^\star} \geq 1 - \delta_0$, for all $j \neq i^\star$.*

Proof. Endpoint gaps are nonnegative by Lemma 7. A positive endpoint gap stays above half its value in a sufficiently small neighborhood, which is stronger than the desired bound since $\pi_j \leq 1$.

If an endpoint gap is zero, then $\Delta_j = \Delta$ and $(b_j - b_{i^\star})(e_{i^\star}) = \Delta$, because the same lemma bounds that contrast by Δ. Put $u_0 = u_0(\Delta, L)$ and $\nu = (1-u_0)e_{i^\star} + u_0 e_j$. Sensitivity gives $(b_j - b_{i^\star})(\nu) \geq \Delta - Lu_0 > 0$, while its two-action bias-dispersion bound gives the reverse inequality

$$(b_j - b_{i^\star})(\nu) \leq \frac{\varepsilon}{\omega(u_0)} = \Delta - Lu_0.$$

Thus equality holds. If $1 - \pi_{i^\star} \leq u_0$, a direct calculation of the coordinates gives $\mathrm{TV}(\pi, \nu) = u_0 - \pi_j$. Applying sensitivity again,

$$(b_j - b_{i^\star})(\pi) \leq \Delta - Lu_0 + L(u_0 - \pi_j) = \Delta - L\pi_j,$$

so $s_{i^\star}(\pi) - s_j(\pi) \geq L\pi_j$. Since there are finitely many actions, we can choose a common positive constant and neighborhood. □

B.4 Convergence at the threshold with two or three actions

Assume $L > 0$ and $\varepsilon = \Psi_L(\Delta)$. Then $\widetilde{s}_{i^\star} \geq c_0 p(1-p)^2 \geq 0$ by Lemma 6, so $x_{t,i^\star}$ is nonincreasing and $\sum_t \bar{\eta}_t \widetilde{s}_{i^\star}(\pi_t) < \infty$ as above. Combining the fixed product with the growth estimate of Lemma 8, we obtain

$$\frac{\pi_{t,j}}{\pi_{t,i^\star}} = \frac{\alpha_{t,j}}{\alpha_{t,i^\star}} \frac{x_{t,i^\star}}{x_{t,j}} = \frac{\alpha_{t,j}}{\alpha_{t,i^\star}} \frac{x_{t,i^\star}^2}{P} \prod_{k \neq i^\star, j} x_{t,k} \leq C(\omega)(1+\tau_t)^{K-2}.$$

Summing over j gives $p_t \geq c(\omega)(1+\tau_t)^{-(K-2)}$. If $\limsup p_t < 1$, then eventually $\widetilde{s}_{i^\star}(\pi_t) \geq c(\omega)p_t$, forcing

$$\sum_t \bar{\eta}_t \widetilde{s}_{i^\star}(\pi_t) \geq c(\omega) \sum_t \frac{\bar{\eta}_t}{(1+\tau_t)^{K-2}} = \infty \quad \text{for } K = 2, 3.$$

The sum dominates the integral of the decreasing function $(1+u)^{-(K-2)}$ over $[0, \infty)$, since $\tau_{t+1} - \tau_t = \bar{\eta}_t$. This contradicts the finite sum established above.

Finally, Lemma 10 and Cauchy–Schwarz give, near $e_{i^\star}$,

$$\widetilde{s}_{i^\star} = \sum_{j \neq i^\star} \pi_j (s_{i^\star} - s_j) \geq c_1 \sum_{j \neq i^\star} \pi_j^2 \geq \frac{c_1 (1-p)^2}{K-1}.$$

Equations (22)–(23) yield the shell condition in Lemma 9. Since $\limsup p_t = 1$, convergence follows. This proves Theorem 1(2) for $K = 2, 3$.

C Failure at and above the Recovery Threshold

C.1 The boundary-case evaluator and inductive bounds

To prove Theorem 2, we use the rewards and bias stated there, with $p = \pi_{i^\star}$, $h = (\Delta - Lp)_+$, and $v = \sum_{a \in \mathcal{I}_{\mathrm{aux}}} \pi_a$ the total auxiliary mass. Every nonzero bias contrast equals h or $-h$, so its TV sensitivity is at most L. Its weighted mean bias is zero, and

$$\sigma_b(\pi)^2 = p(1-p)(\Delta - Lp)_+^2.$$

Every $p \in [0,1]$ is attained by a policy in the simplex; hence its maximal bias dispersion is exactly $\Psi_L(\Delta)$. After a common additive offset, the scores are

$$s_{i^\star} = d := \min\{\Delta, Lp\}, \qquad s_w = 0, \qquad s_a = -\gamma.$$

Thus $i^\star$ has strictly higher mean score than every alternative throughout the interior. The score span is at most $\Delta + \gamma$, so (20) holds with $V = (\Delta + \gamma)^2 + \sigma^2$.

For $\delta > 0$ define the joint event

$$\mathcal{E}_\delta = \{\sup_{t,i} |M_{t,i}| \leq \delta\}.$$

Equation (20) gives

$$\mathbb{P}(\mathcal{E}_\delta^c) \leq \frac{2VV_\eta}{\delta^2}. \tag{25}$$

This single event controls all coordinates at all times. Given ρ, choose δ large enough to make the right side at most ρ.

We now work deterministically on $\mathcal{E}_\delta$. Normalize the initial logits by $\theta_{0,w} = 0$, so $x_{0,w} = 1$. Put

$$A_m = 2^{m+2}, \quad c_{\mathrm{aux}} = \gamma e^{-\delta}/2, \quad B_0 \geq 4A_m m e^{2\delta}, \quad H = A_m \eta_{\max} e^{\delta}.$$

$$C_1 = 2A_m L e^{3\delta} \left(H + \frac{B_0}{c_{\mathrm{aux}}(m-1)} \right),$$

$$C_2 = A_m m \gamma e^{3\delta} \left(\frac{H}{B_0} + \frac{1}{c_{\mathrm{aux}} m} \right).$$

Choose

$$R > \max\{4C_2,\, 2\sqrt{C_1},\, 1\}, \qquad R < x_{0,i^\star} < 2R, \quad B_0 < x_{0,a} < 2B_0 \quad \text{for all } a \in \mathcal{I}_{\mathrm{aux}}. \tag{26}$$

These strict inequalities define a nonempty open set of full-support policies. The logit normalization merely chooses coordinates for that set.

We prove by induction that, for all t on $\mathcal{E}_\delta$,

$$x_{t,i^\star} \geq R/2, \qquad x_{t,a} \geq B_0 + c_{\mathrm{aux}} \tau_t. \tag{27}$$

The bounds hold initially. Suppose they hold up to a given time. The conserved product is at most $2^{m+1} R B_0^m$, and hence

$$x_{t,w} \leq A_m \left(\frac{B_0}{B_0 + c_{\mathrm{aux}} \tau_t} \right)^m, \tag{28}$$

$$p_t \leq \frac{2A_m e^{2\delta}}{R} \left(\frac{B_0}{B_0 + c_{\mathrm{aux}} \tau_t} \right)^m, \tag{29}$$

$$v_t \leq \frac{A_m m e^{2\delta} B_0^m}{(B_0 + c_{\mathrm{aux}} \tau_t)^{m+1}} \leq \frac{1}{4}. \tag{30}$$

The probability bounds follow from $\pi_i/\pi_w = (\alpha_i/\alpha_w)(x_w/x_i)$ and $e^{-\delta} \le \alpha_i \le e^{\delta}$.

The centered scores satisfy

$$\widetilde{s}_{i^\star} = (1-p)d + \gamma v \ge 0, \qquad \widetilde{s}_a = -\gamma(1-v) - pd \le -\gamma/2.$$

Therefore the exact update for $x_{t,a}$ implies

$$x_{t+1,a} - x_{t,a} = x_{t,a}(e^{-\eta_t \pi_{t,a} \widetilde{s}_a} - 1) \ge \bar{\eta}_t \alpha_{t,a}(-\widetilde{s}_a) \ge c_{\mathrm{aux}} \bar{\eta}_t,$$

which preserves the lower bound for each auxiliary coordinate. For the optimal action,

$$0 \le x_{t,i^\star} - x_{t+1,i^\star} \le \bar{\eta}_t e^{\delta}(Lp_t + \gamma v_t). \tag{31}$$

Also $Z_t \ge e^{-\delta}/x_{t,w}$, so (28) gives $\bar{\eta}_t \le H$.

For any nonnegative decreasing integrable function φ and mesh increments $0 < \bar{\eta}_t \le H$,

$$\sum_t \bar{\eta}_t \varphi(\tau_t) \le H\varphi(0) + \int_0^\infty \varphi(u)\, du. \tag{32}$$

To prove this bound for a finite sum, subtract the right-endpoint sum from the left-endpoint sum. The difference is at most $H \sum_t (\varphi(\tau_t) - \varphi(\tau_{t+1})) \le H\varphi(0)$, while the right-endpoint sum is bounded by the integral. Taking limits proves the bound. We apply it to

$$\varphi_1(u) = \left(\frac{B_0}{B_0 + c_{\mathrm{aux}} u}\right)^m, \qquad \varphi_2(u) = \frac{B_0^m}{(B_0 + c_{\mathrm{aux}} u)^{m+1}},$$

whose integrals are $B_0/[c_{\mathrm{aux}}(m-1)]$ and $1/(c_{\mathrm{aux}} m)$, respectively. Summing (31) using (29)–(30) gives

$$x_{0,i^\star} - x_{T,i^\star} \le C_1/R + C_2 < R/2$$

up to and including any proposed first exit from the inductive bounds. This excludes such an exit and proves (27) inductively for all times.

By Lemma 8, $\tau_t \to \infty$. Thus all auxiliary coordinates $x_{t,a}$ diverge, $x_{t,i^\star}$ decreases to a strictly positive limit, and (28) makes $x_{t,w} \to 0$. The probability ratios therefore give $\pi_t \to e_w$ on $\mathcal{E}_\delta$. Equation (25) proves the probability bound. Choosing B_0 arbitrarily large and then R arbitrarily large makes the open sets (26) arbitrarily close to e_w. All choices depend only on Δ, L, γ, the schedule, σ^2, and ρ, and apply uniformly to every noise kernel satisfying that variance bound. The exact rate statement on this event is proved in Appendix D.

C.2 Explicit four-action constants

For the example in Section 3, use

$$m = 2, \quad \Delta = L = \gamma = 1, \quad \delta = \tfrac{1}{4}, \quad A_m = 16, \quad B_0 = 256, \quad R = 1000.$$

Here $c_{\mathrm{aux}} = e^{-1/4}/2$, $H = 0.16e^{1/4}$, and the preceding expressions give $C_1/R + C_2 < 132 < 500 = R/2$. Also $B_0 \ge 4A_m m e^{2\delta}$, and the initial coordinates $(x_{0,i^\star}, x_{0,w}, x_{0,3}, x_{0,4}) = (1500, 1, 384, 384)$ satisfy the strict box conditions. With zero evaluation noise, $V = 4$ and $V_\eta = \pi^2/60000$, so

$$\mathbb{P}(\mathcal{E}_{1/4}^c) \le \frac{8}{(1/4)^2} \frac{\pi^2}{60000} < 0.022.$$

This proves the stated failure probability greater than 0.97.

For this instance the expected update is the gradient of a scalar objective:

$$f(\mathrm{softmax}(\theta)) = \nabla_\theta \left(\tfrac{1}{2}\pi_1^2 - \pi_3 - \pi_4\right).$$

Indeed the gradient with respect to π is $(\pi_1, 0, -1, -1)$, and the softmax Jacobian converts it to $\pi_i(s_i - \bar{s})$. Thus failure can occur even when the expected update is a gradient.

C.3 Failure above the boundary and at zero sensitivity

To prove Theorem 1(3) for $L > 0$, choose

$$\Delta < c < \Delta + L, \quad \Psi_L(c) < \varepsilon, \quad a = c - \Delta, \quad \gamma = a/4.$$

Use $r_{i^\star} = \Delta$, $r_w = 0$, $r_j = -\gamma$ for the remaining $K - 2$ actions, and the same bias for the optimal action and the remaining actions as above, with $h = (c - Lp)_+$. It has uniform bias dispersion $\Psi_L(c) < \varepsilon$ and sensitivity at most L. After a common additive offset its scores are

$$s_{i^\star} = \Delta - (c - Lp)_+, \qquad s_w = 0, \qquad s_j = -\gamma.$$

For $p \le a/(4L)$, $s_{i^\star} \le -3a/4$, so w has mean score at least γ above every alternative. A sufficiently small neighborhood of e_w lies in this region. Lemma 9 gives convergence to e_w with probability tending uniformly to one as $\pi_{0,w} \to 1$. For $K = 2$, simply omit the auxiliary actions; the same local argument applies.

If $L = 0$ and $\varepsilon = \Delta/2$, take rewards $r_{i^\star} = \Delta$, $r_j = 0$ for $j \neq i^\star$, and constant biases $b_{i^\star} = -\Delta$, $b_j = 0$. Every score is zero and $\sup \sigma_b = \Delta/2$, so noiseless (U) leaves every policy fixed. If $\varepsilon > \Delta/2$, choose $\Delta < c < 2\varepsilon$, let $r_{i^\star} = \Delta$, $r_w = 0$, and $r_j = -\gamma$ for any $\gamma > 0$ on the other actions, and put $b_{i^\star} = -c$, $b_i = 0$ otherwise. The bias-dispersion bound is $c/2 < \varepsilon$. Now w has the highest evaluator score everywhere, with a uniform positive gap. The strict convergence proof in Appendix B, applied to this margin, gives $\pi_t \to e_w$ almost surely from every full-support initialization.

D Exact Asymptotic Rates and Comparison Counts

D.1 Converting coordinate growth to cumulative step size

Work on a path where the martingales converge. Renormalize the coordinates from (8) by replacing $x_{t,i}$ with $e^{-M_{\infty,i}} x_{t,i}$ and $\alpha_{t,i}$ with $e^{M_{t,i} - M_{\infty,i}}$. We retain the notation x, α for the renormalized coordinates. Their product is still P, because $\sum_i M_{\infty,i} = 0$, while now $\alpha_{t,i} \to 1$. Taylor expansion of the exact update gives

$$x_{t+1,i} - x_{t,i} = -\bar{\eta}_t \alpha_{t,i} \widetilde{s}_i(\pi_t) + r_{t,i}, \qquad |r_{t,i}| \le C(\omega) \eta_t \bar{\eta}_t \pi_{t,i} = o(\bar{\eta}_t). \tag{33}$$

The remainder bound follows from $x_{t,i} \eta_t^2 f_i^2 = \eta_t \bar{\eta}_t \alpha_{t,i} \pi_{t,i} \widetilde{s}_i^2$ and the uniform score bound.

Suppose $\pi_t \to e_{i^\star}$ and $\Gamma_j = s_{i^\star}(e_{i^\star}) - s_j(e_{i^\star}) > 0$. Continuity of score differences gives $\widetilde{s}_j(\pi_t) \to -\Gamma_j$. Since $\tau_T = \sum_{t<T} \bar{\eta}_t \to \infty$, summing (33) and taking weighted averages yields

$$x_{T,j} \sim \Gamma_j \tau_T \quad (j \neq i^\star), \qquad x_{T,i^\star} \sim \frac{P}{(\prod_{j \neq i^\star} \Gamma_j) \tau_T^{K-1}}.$$

Thus

$$Z_T = \sum_i \frac{\alpha_{T,i}}{x_{T,i}} \sim C_Z \tau_T^{K-1}, \qquad C_Z = \frac{\prod_{j \neq i^\star} \Gamma_j}{P} > 0. \tag{34}$$

Since $\bar{\eta}_t / \tau_t \to 0$, the binomial expansion gives

$$\frac{\tau_{t+1}^K - \tau_t^K}{\eta_t} = K \frac{\tau_t^{K-1}}{Z_t} (1 + o(1)) \to K / C_Z.$$

Summing these increments and using $S_T \to \infty$ gives $\tau_T^K / S_T \to K / C_Z$. Combining this limit with $\pi_{T,j} = \alpha_{T,j} / (x_{T,j} Z_T)$ proves $S_T \pi_{T,j} \to 1/(K\Gamma_j)$. Summing over $j \neq i^\star$ proves (14). Strictly below the boundary, Lemma 7 gives $\Gamma_j \ge \Delta - \Psi_L^{-1}(\varepsilon) > 0$ and Theorem 1 gives almost-sure convergence.

D.2 Exponents for failure at the threshold

Under the conditions of Theorem 2, the initial-policy sets can be chosen so that, on an event of probability at least $1 - \rho$, for a random $C > 0$,

$$\pi_t \to e_w, \qquad S_T^{m/(m+1)} \pi_{T,i^\star} \to C, \qquad S_T \pi_{T,a} \to \frac{1}{(m+1)\gamma} \quad (a \in \mathcal{I}_{\mathrm{aux}}).$$

On this same event the optimal action is compared infinitely often with divergent cumulative step weight, but has finite cumulative conditional logit drift. We derive all these conclusions on the bounded-noise event used to prove Theorem 2.

On the event $\mathcal{E}_\delta$ of Appendix C, the renormalized $x_{t,i^\star}$ converges to $x_\infty \in (0,\infty)$ and $\pi_t \to e_w$. For each auxiliary action, we have $\widetilde{s}_a \to -\gamma$, so (33) yields $x_{t,a} \sim \gamma\tau_t$. The conserved product and the softmax normalizer give

$$x_{t,w} \sim \frac{P}{x_\infty \gamma^m \tau_t^m}, \qquad Z_t \sim C_Z \tau_t^m, \qquad C_Z = \frac{x_\infty \gamma^m}{P}.$$

Applying the same conversion from τ_T to S_T, we obtain $\tau_T^{m+1}/S_T \to (m+1)/C_Z$. Therefore

$$S_T \pi_{T,a} \to \frac{1}{(m+1)\gamma}, \qquad S_T^{m/(m+1)} \pi_{T,i^\star} \to \frac{1}{x_\infty C_Z}\left(\frac{C_Z}{m+1}\right)^{m/(m+1)} > 0.$$

Since $m \geq 2$, comparison with integrals in cumulative step size gives

$$\sum_t \eta_t \pi_{t,i^\star} = \infty, \qquad \sum_t \eta_t \pi_{t,i^\star}^2 < \infty, \qquad \sum_t \eta_t \pi_{t,i^\star} \sum_{a \in \mathcal{I}_{\mathrm{aux}}} \pi_{t,a} < \infty.$$

In particular, the exponents in the last two sums are $2m/(m+1) > 1$ and $1 + m/(m+1) > 1$. The bounded step sizes justify this integral comparison after any fixed positive cumulative step size. Since $f_{i^\star} = \pi_{i^\star}[(1-\pi_{i^\star})d + \gamma v]$ with $d \leq L\pi_{i^\star}$, the cumulative conditional drift $\sum_t \eta_t f_{i^\star}$ is finite. The optimal action's logit therefore converges to a finite value, whereas $\theta_{t,w} \to +\infty$. In the explicit example, the true expected reward tends to 0 and $\sigma_b(\pi_t) \to 0$, although the optimal reward is 1.

D.3 From predictable frequencies to realized counts

Under the assumptions of Theorem 3, let $N_T = \sum_{t<T} \mathbf{1}_{\{A_t \neq B_t\}}$ count all distinct-action comparisons and $N_T^{i,j}$ those of the unordered pair $\{i,j\}$ before time T. Then, almost surely on the same convergence event,

$$N_T \sim \frac{2}{K}\left(\sum_{j \neq i^\star} \Gamma_j^{-1}\right) \sum_{1 \leq t < T} S_t^{-1}, \qquad N_T^{i^\star,j} \sim \frac{2}{K\Gamma_j} \sum_{1 \leq t < T} S_t^{-1}. \tag{35}$$

For distinct $i, j \neq i^\star$, $N_T^{i,j} \sim 2(K^2\Gamma_i\Gamma_j)^{-1} \sum_{1 \leq t < T} S_t^{-2}$. All these counts diverge.

To derive the comparison-count equivalents, let I_t be any one of the relevant comparison indicators, $q_t = \mathbb{E}[I_t \mid \mathcal{F}_t]$, and $\Lambda_T = \sum_{t<T} q_t$. On $\{\Lambda_T \to \infty\}$,

$$\frac{\sum_{t<T} I_t}{\Lambda_T} \to 1 \quad \text{almost surely.} \tag{36}$$

To prove this, observe that $\sum_t (I_t - q_t)/(1+\Lambda_{t+1})$ is a martingale with total conditional variance at most

$$\sum_t \frac{q_t}{(1+\Lambda_{t+1})^2} \leq \int_0^\infty \frac{du}{(1+u)^2} = 1.$$

Here $\Lambda_{t+1} = \Lambda_t + q_t$ is $\mathcal{F}_t$-measurable. Martingale convergence and summation by parts (Kronecker's lemma) imply $\sum_{t<T}(I_t - q_t)/(1+\Lambda_T) \to 0$, proving (36).

For the indicator of any comparison of distinct actions, $q_t = 1 - \sum_i \pi_{t,i}^2$, while for an unordered distinct pair $\{i,j\}$, $q_t = 2\pi_{t,i}\pi_{t,j}$. On a successful path with positive endpoint gaps, (14) gives

$$1 - \sum_i \pi_{t,i}^2 \sim \frac{2}{KS_t} \sum_{j \neq i^\star} \Gamma_j^{-1}, \quad 2\pi_{t,i^\star}\pi_{t,j} \sim \frac{2}{K\Gamma_j S_t}, \quad 2\pi_{t,i}\pi_{t,j} \sim \frac{2}{K^2\Gamma_i\Gamma_j S_t^2}\ (i,j \neq i^\star).$$

The deterministic bounds $S_t \leq \eta_{\max} t$ and $S_t^2 \leq tV_\eta$ imply that both $\sum_{t \geq 1} S_t^{-1}$ and $\sum_{t \geq 1} S_t^{-2}$ diverge. Summing these asymptotic equivalents and applying (36) proves all the count equivalents stated above. For power steps, $S_T \sim \eta_0 T^{1-a}/(1-a)$, giving the rates in sampling rounds and comparison counts stated in the main text. One comparison can require two scalar evaluator calls if the two scores are obtained separately, so the cost under this convention is $2N_T$ scalar calls instead of N_T contrast queries.

D.4 A BINARY BOUNDARY-CASE RATE WITH A ZERO ENDPOINT GAP

The positive-gap equivalent in Theorem 3 cannot be continued to an endpoint where the evaluator gap vanishes. The following family gives an exact alternative under the same step and noise assumptions.

Proposition 11 (Binary boundary-case rate)**.** *Let $K = 2$, $r = (\Delta, 0)$, and $\Delta, L > 0$. For $p = \pi_1$, set*

$$h(p) = (\Delta - L(1-p))_+, \qquad b_1 = -(1-p)h(p), \quad b_2 = ph(p).$$

This evaluator has sensitivity L and uniform bias dispersion $\Psi_L(\Delta)$. Under (3) *and the conditional noise assumptions of Section 2, update* (U) *satisfies, for every full-support initialization,*

$$\sqrt{S_T}(1 - p_T) \longrightarrow \frac{1}{\sqrt{4L}} \qquad \textit{almost surely}.$$

Proof. The pairwise bias is $-h$, which is L-Lipschitz. With $q = 1 - p$, its bias dispersion is $\sqrt{q(1-q)}(\Delta - Lq)_+$, whose supremum is $\Psi_L(\Delta)$. The score gap is $d(p) = \Delta - h(p) = \min\{\Delta, L(1-p)\}$. Theorem 1(2) therefore gives $p_t \to 1$ almost surely, even though $d(1) = 0$.

Write $z_t = \theta_{t,1} - \theta_{t,2} = \log(p_t/q_t)$. Its conditional mean increment is $2\eta_t p_t q_t d(p_t)$. The scalar martingale

$$M_T^\ell = \sum_{t<T} \eta_t \big(g_{t,1} - g_{t,2} - 2p_t q_t d(p_t)\big)$$

converges almost surely, by the conditional second-moment bound in (19). Fix a path on which both convergence statements hold, and define $y_t = z_t - M_t^\ell + M_\infty^\ell$ and $u_t = e^{2y_t}$. These are pathwise auxiliary variables; no conditioning on M_∞^ℓ is needed. Then $y_t - z_t \to 0$ and

$$u_{t+1} = u_t \exp\{4\eta_t p_t q_t d(p_t)\}.$$

Eventually $d(p_t) = Lq_t$. Set $a_t = 4L\eta_t p_t q_t^2$, which tends to zero since $\eta_t \to 0$. The exact increment identity gives

$$\frac{u_{t+1} - u_t}{\eta_t} = 4L u_t p_t q_t^2 \frac{e^{a_t} - 1}{a_t} \longrightarrow 4L,$$

because $u_t q_t^2 = e^{2(y_t - z_t)} p_t^2 \to 1$. Taking weighted averages of these increments yields $u_T/S_T \to 4L$. Finally, $q_T\sqrt{u_T} = e^{y_T - z_T} p_T \to 1$, which proves the asserted limit. □

E BALANCED-QUERY CONVERGENCE AND INDISTINGUISHABILITY

E.1 CONVERGENCE WITH BALANCED QUERIES AND THE CORRECTED UPDATE

We prove convergence in Theorem 5. The true rewards are fixed as before, but only the uniform bias-dispersion condition (1) is assumed for the bias. At the balanced evaluation policy $\mu^{ij} = (e_i + e_j)/2$,

$$\sigma_b(\mu^{ij}) = \tfrac{1}{2}|b_i(\mu^{ij}) - b_j(\mu^{ij})| \le \varepsilon.$$

Thus c_{ij} defined in Section 5 is an antisymmetric matrix with $|c_{ij}| \le G_{\mathrm{BQ}} := R_r + 2\varepsilon$ and $c_{i^\star,j} \ge \Delta_j - 2\varepsilon \ge \kappa_{\mathrm{BQ}} := \Delta - 2\varepsilon > 0$. Antisymmetry is the only structural property of the matrix c_{ij} used below.

Let $\widetilde{g}_t$ denote the increment in (BQ) before multiplication by η_t. Averaging over the ordered pairs and using centering conditional on the sampled pair gives

$$\mathbb{E}[\widetilde{g}_{t,i} \mid \mathcal{F}_t] = \pi_t^{\min} v_i(\pi_t), \qquad v_i(\pi) = \sum_j \pi_j c_{ij}, \qquad \sum_i \pi_i v_i = 0.$$

The last identity follows from antisymmetry. Put $V_{\mathrm{BQ}} = G_{\mathrm{BQ}}^2 + \sigma^2$. On a distinct-action pair involving i, the squared coordinate multiplier is $(\pi_t^{\min})^2/(4\pi_{t,i}^2)$, and the total probability of such pairs is $2\pi_{t,i}(1 - \pi_{t,i})$. Hence

$$\mathbb{E}[\widetilde{g}_{t,i}^2 \mid \mathcal{F}_t] \le \frac{V_{\mathrm{BQ}}(\pi_t^{\min})^2(1 - \pi_{t,i})}{2\pi_{t,i}} \le \frac{V_{\mathrm{BQ}}\pi_t^{\min}}{2}, \qquad \mathbb{E}\|\widetilde{g}_t\|^2 \le V_{\mathrm{BQ}}/2. \tag{37}$$

The final inequality uses $K\pi_t^{\min} \le 1$. Consequently, the vector martingale formed by the centered increments multiplied by η_t converges almost surely and in L^2, and $\sum_t \eta_t^2 \|\widetilde{g}_t\|^2 < \infty$ almost surely.

Define $H_T = \sum_{t<T} \eta_t \pi_t^{\min}$. If $H_\infty < \infty$ on a path where the martingale converges, the bound $|v_i| \le G_{\mathrm{BQ}}$ makes the drift of each logit absolutely summable. Every logit therefore converges to a finite value. The limiting softmax policy is interior, so $\pi_t^{\min}$ tends to a positive value, contradicting $\sum_t \eta_t = \infty$. Thus $H_T \to \infty$ almost surely. Unlike (U), update (BQ) need not conserve the logit sum; the divergence of H_T follows without this invariant.

For $p = \pi_{i^\star}$ and $q_j = \pi_j/(1-p)$, antisymmetry gives

$$v_{i^\star} - \sum_{j \neq i^\star} q_j v_j = v_{i^\star} + \frac{p}{1-p} v_{i^\star} = \frac{v_{i^\star}}{1-p} = \frac{\sum_{j \neq i^\star} \pi_j c_{i^\star,j}}{1-p} \ge \kappa_{\mathrm{BQ}}.$$

We now expand the optimal action's log odds as in (24), using $\widetilde{g}_t$. The first-order drift is at least $\eta_t \pi_t^{\min} \kappa_{\mathrm{BQ}}$. By (37) and (18), the noise martingale converges and the nonnegative remainder is summable. Thus, for a convergent R_T,

$$\log \frac{\pi_{T,i^\star}}{1-\pi_{T,i^\star}} \ge \log \frac{\pi_{0,i^\star}}{1-\pi_{0,i^\star}} + \kappa_{\mathrm{BQ}} H_T + R_T \longrightarrow +\infty.$$

The correction has bounded coordinate multipliers:

$$\frac{\pi_t^{\min}}{2\pi_{t,i}} = \tfrac{1}{2} \exp\left(\min_j \theta_{t,j} - \theta_{t,i} \right) \le \tfrac{1}{2}.$$

We can therefore compute the multipliers without taking the reciprocal of a small action probability. Reading the minimum may require $O(K)$ arithmetic in a direct implementation. The observation budget remains two actor samples and one contrast query on each round with distinct sampled actions.

E.2 A LOWER BOUND FOR ARBITRARY ADAPTIVE EVALUATION POLICIES

We prove the information lower bound in Theorem 5. Consider a method whose only information about r comes from evaluator responses. It may randomize, select actor or evaluation policies adaptively, query individual scores or arbitrary comparisons, and base its stopping decision on the transcript of all past queries and responses.

Set the noise to zero and use the binary environments

$$r^+ = (\Delta, 0), \quad b^+ = (-\Delta/2, \Delta/2), \qquad r^- = (0, \Delta), \quad b^- = (\Delta/2, -\Delta/2).$$

In each,

$$s(\mu) = (\Delta/2, \Delta/2) \quad \text{for every } \mu, \qquad \sigma_b(\mu) = \Delta\sqrt{\mu_1 \mu_2} \le \Delta/2.$$

Their biases are constant functions of the policy, so their sensitivity is zero. The environments are admissible for every bound $\varepsilon \ge \Delta/2$ (and every $L \ge 0$), yet have different unique optimal actions.

We couple the algorithm's internal random bits across the two environments. If the past transcripts agree, the algorithm chooses the same next queries, actor policies, and evaluation policies. Those queries receive equal responses. Induction therefore shows that the entire transcript, every stopping decision, and every output have the same law. For an almost surely terminating binary decision rule, its two success probabilities sum to one; at least one is at most $1/2$. A policy process with the same law in both environments also cannot converge almost surely to two different vertices. The transcript laws remain equal even for an infinite sequence of adaptive queries.

For $K > 2$, append actions with reward $-\Delta$ and zero bias in both environments. The first two rewards and biases remain as above, giving the same minimum reward gap Δ and the same full score vector in both environments. All bias values lie in $[-\Delta/2, \Delta/2]$. For any random variable in $[a, b]$, the inequality $\mathbb{E}[(X-a)(b-X)] \ge 0$ implies $\mathrm{Var}(X) \le (\mathbb{E}X - a)(b - \mathbb{E}X) \le (b-a)^2/4$. Thus the uniform bias dispersion is still at most $\Delta/2$. The two success probabilities now sum to at most one because an algorithm may choose an appended action. This proves the same impossibility for the enlarged action set.

E.3 FINITE-ROUND IDENTIFICATION UNDER SUB-GAUSSIAN NOISE

The balanced-query oracle also permits identification with a deterministic query budget when the comparison noise has stronger tail control. The strategy here keeps the actor uniform throughout data collection; it does not run either logit update.

For T rounds, draw A_t, B_t independently and uniformly from the K actions. On a distinct-action pair, query its contrast at $\mu^{A_t B_t}$ as in Section 5. For each $i < j$, average the observed contrasts, changing the sign when the sampled order is (j, i), to obtain $\widehat{c}_{ij}$. Set $\widehat{c}_{ji} = -\widehat{c}_{ij}$, and set both values to zero if that pair was never observed. Return an action i if $\widehat{c}_{ij} > 0$ for every $j \neq i$; if no such action exists, report no decision. Antisymmetry makes such an action unique whenever it exists.

Proposition 12 (Finite-round identification)**.** *Assume* (1)*, balanced-query access, and $\varepsilon < \Delta/2$. Suppose that for every $\lambda \in \mathbb{R}$ the comparison noise satisfies*

$$\mathbb{E}[\exp(\lambda \zeta_t) \mid \mathcal{F}_t, A_t, B_t] \leq \exp(\lambda^2 \nu^2/2)$$

on distinct-action pairs, for a fixed $\nu \geq 0$. Let $\kappa_{\mathrm{BQ}} = \Delta - 2\varepsilon$. For any $\beta \in (0, 1)$, the uniform-query strategy returns the optimal action with probability at least $1 - \beta$ whenever

$$T \geq K^2 \max\left\{4, \frac{8\nu^2}{\kappa_{\mathrm{BQ}}^2}\right\} \log \frac{3\binom{K}{2}}{\beta}. \tag{38}$$

It uses exactly $2T$ action draws and at most T contrast queries. No sensitivity or step-size assumption is required.

Proof. Fix an unordered pair $\{i, j\}$, with $i < j$. Let I_t^{ij} indicate its occurrence, let $N_T^{i,j} = \sum_{t<T} I_t^{ij}$, and let W_T^{ij} be the sum of the corresponding oriented noise terms. The fixed uniform sampling policy gives independent Bernoulli indicators with mean $2/K^2$. A multiplicative Chernoff bound therefore yields

$$\mathbb{P}\left(N_T^{i,j} < T/K^2\right) \leq e^{-T/(4K^2)}. \tag{39}$$

When $N_T^{i,j} > 0$, the empirical error is $\widehat{c}_{ij} - c_{ij} = W_T^{ij}/N_T^{i,j}$.

Suppose first that $\nu > 0$. Conditional sub-Gaussianity, with the sign adjusted to the sampled order, implies that

$$\exp\left(\lambda W_t^{ij} - \frac{\lambda^2 \nu^2}{2} N_t^{i,j}\right)$$

is a nonnegative supermartingale starting at one. Conditioning only on the current history and sampled pair yields this bound without a temporal-independence assumption. Taking $\lambda = \kappa_{\mathrm{BQ}}/(2\nu^2)$ and applying Markov's inequality on the event $N_T^{i,j} \geq T/K^2$ gives

$$\mathbb{P}\left(W_T^{ij} \geq \tfrac{1}{2}\kappa_{\mathrm{BQ}} N_T^{i,j},\, N_T^{i,j} \geq T/K^2\right) \leq \exp\left(-\frac{\kappa_{\mathrm{BQ}}^2 T}{8\nu^2 K^2}\right).$$

The negative tail obeys the same bound. Combining it with (39) and taking a union bound over the $\binom{K}{2}$ pairs bounds the probability of any missing pair or any error of magnitude at least $\kappa_{\mathrm{BQ}}/2$ by

$$\binom{K}{2}\left[e^{-T/(4K^2)} + 2\exp\left(-\frac{\kappa_{\mathrm{BQ}}^2 T}{8\nu^2 K^2}\right)\right] \leq \beta$$

under (38). For $\nu = 0$, the noise vanishes almost surely, and the count bound alone suffices.

On the complementary event, every comparison of the optimal action satisfies

$$\widehat{c}_{i^\star, j} > c_{i^\star, j} - \kappa_{\mathrm{BQ}}/2 \geq \kappa_{\mathrm{BQ}}/2 > 0.$$

Thus the optimal action passes all empirical comparisons. Every other action loses its empirical comparison against the optimal action, so the strategy returns exactly $i^\star$. □

Centered noise bounded in absolute value by ν satisfies the stated moment-generating-function bound, as does Gaussian noise with standard deviation ν. The fixed budget requires a lower bound on κ_{BQ} and an upper bound on ν. Uniform sampling removes the rare-action slowdown of (BQ) by sampling all pairs, so Proposition 12 is an identification bound for the balanced-query oracle rather than a recovery-time bound for (U) or (BQ).

F THE ROLE OF THE ASSUMPTIONS

We give examples that clarify the assumptions and limits of the recovery theorems.

F.1 EVALUATOR FAMILIES AND NUMERICAL CALIBRATION

The assumptions can be checked exactly in simple structured families. Fix a nonempty proper group $\mathcal{G} \subset \{1, \dots, K\}$ and $\lambda \geq 0$. A conformity bonus assigns

$$b_i(\mu) = \lambda u \mathbf{1}_{\{i \in \mathcal{G}\}}, \qquad u = \mu(\mathcal{G}).$$

The bonus grows with the policy mass of the favored group. Within either group, pairwise bias is zero; across groups it is $\pm\lambda u$. Since $|\mu(\mathcal{G}) - \mu'(\mathcal{G})| \leq \mathrm{TV}(\mu, \mu')$, the exact sensitivity is $L = \lambda$. There are only two bias values, so

$$\sigma_b(\mu)^2 = \lambda^2 u^3 (1-u), \qquad \sup_\mu \sigma_b(\mu) = \frac{3\sqrt{3}\lambda}{16} = \Psi_\lambda(\lambda).$$

The maximum occurs at $u = 3/4$. These constants do not depend on the number of actions or group sizes. By monotonicity of Ψ_λ, Theorem 1(1) certifies convergence whenever $\Delta > \lambda$.

A second family separates ranking accuracy from numerical calibration. Let $s_i(\mu) = a r_i + c(\mu)$ for a fixed $a > 0$ and any common additive offset c. Then $b_i = (a-1)r_i + c$, all pairwise biases are constant in μ, and

$$L = 0, \qquad \sigma_b(\mu) = |a-1|\sqrt{\mathrm{Var}_\mu(r)}, \qquad \sup_\mu \sigma_b(\mu) = \frac{|a-1|R_r}{2}.$$

The last equality follows by putting equal mass on a minimum- and a maximum-reward action; no distribution on an interval of length R_r has variance exceeding $R_r^2/4$. Thus the evaluator may preserve every ranking while its bias dispersion is large. The threshold is a worst-case guarantee over a class, not a necessary condition for each particular evaluator.

Scores and rewards must be expressed on a common numerical scale. Multiplying both by $a > 0$ multiplies Δ, L, ε by a and preserves the threshold inequality because $\Psi_{aL}(a\Delta) = a\Psi_L(\Delta)$. Multiplying only the evaluator changes its bias relative to the fixed reward. Removing a common additive offset does not remove this calibration requirement.

For an unrestricted learned evaluator, finitely many measurements on visited policies cannot certify either supremum over the closed simplex. Such certification needs known rewards or error certificates, together with structural control of the evaluator between queries. Increasing K or refining actions into longer sequences need not preserve ε or L; the group family above is a concrete case where the constants remain uniform.

F.2 AGGREGATE ADVANTAGE DOES NOT IMPLY PAIRWISE DOMINANCE

Let $K = 3$, $r = (1, 0, 0)$, $L = 1$, and, for a policy π, put

$$q = \pi_1 + \pi_2, \qquad h = \tfrac{4}{3} - q, \qquad b = (-h/2, h/2, 0).$$

Every pairwise bias contrast is 1-Lipschitz in TV, because $|q(\pi) - q(\pi')| \leq \mathrm{TV}(\pi, \pi')$. Direct calculation gives

$$\sigma_b(\pi)^2 = \frac{h^2}{4}\left(q - (\pi_2 - \pi_1)^2\right).$$

For fixed q, this expression is maximized at $\pi_1 = \pi_2 = q/2$. Maximizing $q(4/3 - q)^2/4$ over $q \in [0, 1]$ gives $\sup_\pi \sigma_b(\pi) = 8/27$, attained at $q = 4/9$. This is strictly smaller than $\Psi_1(1) = 3\sqrt{3}/16$. Yet at $\pi = (1/12, 1/12, 5/6)$,

$$s_1 - s_2 = 1 - h = -1/6.$$

Thus Theorem 1(1) guarantees convergence even though global pairwise dominance fails.

Even without bias, the optimal-action log odds need not have globally positive drift. For

$$r = (1, 0, -100), \qquad \pi = (1/100, 4/5, 19/100),$$

the exact drift $f_1 - \sum_{j \neq 1} \pi_j f_j/(1 - \pi_1)$ equals $-4515619/495000 < 0$. The convergence proof therefore combines global aggregate advantage with a *local* log-odds argument.

F.3 Constant step sizes can prevent convergence at the boundary

Under the stated noise assumption, a constant step size can prevent convergence even with two actions. Take two actions with rewards $(1, 0)$, $p = \pi_1$, and

$$h(p) = p, \qquad b_1 = -(1-p)p, \quad b_2 = p^2.$$

The sensitivity is 1, the uniform bias dispersion is $3\sqrt{3}/16 = \Psi_1(1)$, and the mean score gap is $1 - p$. On a comparison of distinct actions, orient the contrast in favor of action 1 and return

$$C = +1 \text{ with probability } 1 - p/2, \qquad C = -1 \text{ with probability } p/2.$$

Then $\mathbb{E}C = 1 - p$ and the centered noise has variance at most 1. Reverse the sign when the ordered pair is reversed. For a fixed $\eta > 0$, the log odds change by $\pm\eta$ at each comparison of distinct actions and remain unchanged otherwise.

Consider the embedded chain observed at comparisons of distinct actions. Index its lattice sites by $z_k = z_0 + \eta k$, $k \in \mathbb{Z}$, and write $p_k = (1 + e^{-z_k})^{-1}$. Its upward and downward probabilities are $u_k = 1 - p_k/2$ and $v_k = p_k/2$. For $k \geq 1$,

$$\rho_k = \frac{v_k}{u_k} = \frac{1}{1 + 2e^{-z_k}}, \qquad \prod_{k=1}^{n} \rho_k \longrightarrow c > 0,$$

since $\sum_{k\geq 1} e^{-z_k} < \infty$. The finite-interval gambler's-ruin recurrence gives, starting from $k = 1$,

$$\mathbb{P}_1(\text{hit } n \text{ before } 0) = \left(1 + \sum_{j=1}^{n-1} \prod_{k=1}^{j} \rho_k\right)^{-1} \to 0.$$

Thus the chain returns to site 0 from the positive side almost surely. For $k \leq 0$, $u_k \geq 1 - p_0/2 > 1/2$, so coupling with a positively drifting nearest-neighbor walk gives a return to site 0 almost surely. The Markov property then gives infinitely many visits to the starting state. Every finite log-odds state has positive probability of a distinct-action pair, so each holding time is finite almost surely. Thus the process indexed by sampling rounds also fails to converge to the optimal vertex.

F.4 Finite variance need not give a finite expected waiting time

Take the unbiased binary problem with reward gap 1 and $p_0 = 1/2$. On a first round with distinct sampled actions, use the oriented contrast $1 + X$, where X is symmetric with probabilities proportional to n^{-4} at $\pm an$, $n \geq 1$, for some $a > 0$. Choose a small enough to meet any prescribed positive variance bound; $\mathbb{E}X = 0$ and $\mathbb{E}X^2 < \infty$, but $\mathbb{E}e^{-\lambda X} = \infty$ for every $\lambda > 0$. Set all subsequent evaluation noise to zero.

Conditional on a first round with distinct sampled actions, which has probability $1/2$, the new log odds are $z_1 = \eta_0(1 + X)$. The policy remains unchanged until the next distinct-action pair. Its conditional expected waiting time in rounds is

$$\frac{1}{2p_1(1-p_1)} = \frac{e^{z_1} + 2 + e^{-z_1}}{2}.$$

Averaging over X gives infinite expectation. This example satisfies all finite-variance noise assumptions of the convergence theorem below the threshold, so almost-sure convergence is compatible with an infinite expected waiting cost. Even a noiseless process initialized with optimal-action mass p waits on average $[2p(1-p)]^{-1}$ rounds for its first binary comparison. Thus its expected sampling cost is unbounded over full-support initializations.

F.5 Real-valued comparisons, sign feedback, and inverse weights

A positive mean real-valued contrast need not imply a majority of favorable signs. For example, $C = -1$ with probability $2/3$ and $C = 5$ with probability $1/3$ has $\mathbb{E}C = 1$ but $\mathbb{E}[\operatorname{sign} C] = -1/3$. Replacing the real-valued contrast in a drift identity by its sign therefore changes the expected update. Mean score differences alone do not determine which sign is more probable.

Likewise, unscaled importance weighting can have unbounded variance near a vertex. If a pair is observed with probability $q > 0$, its conditional contrast has mean d and second moment $d^2 + \nu^2$, and I is its observation indicator, then

$$\mathrm{Var}(IC/q) = \frac{d^2 + \nu^2}{q} - d^2.$$

For fixed $d^2 + \nu^2 > 0$, this variance diverges as $q \to 0$. Update (BQ) keeps its multipliers bounded by scaling the inverse weights by the minimum action probability; Appendix E gives its conditional moment bounds.

F.6 Further related work

The estimator is the unclipped categorical specialization of the comparative policy-gradient identity in Wu et al. (2023). Segment-comparison policy gradients (Assmus et al., 2026) and zeroth-order preference alignment (Chen et al., 2026) use different feedback interfaces. For self-rewarding methods, Fu et al. (2026) establish iterative guarantees under realizability and conditions on confidence and margins, while Yang et al. (2026) study binary self-confirmation and cross-agent correction. These results do not characterize the joint effect of a shrinking policy-dependent margin and a declining probability of sampling the optimal action. Our analysis determines when the resulting stochastic update accumulates sufficient corrective drift.

Empirically, CREAM uses agreement between successive evaluators to regularize preference training (Wang et al., 2025). In reference-free judge self-play, Zhou (2026) report increased judge acceptance without corresponding accuracy gains in reasoning-suppressed Qwen3/GSM8K experiments.

Policy-dependent feedback also arises in performative prediction (Perdomo et al., 2020; Farina & Perdomo, 2026). In performative RL, Mandal et al. (2023) prove convergence of occupancy-space retraining and projected gradient updates to performatively stable policies under sensitivity conditions; Basu et al. (2026) differentiate performative return through changing rewards and transitions. Here the target is the optimum of a fixed external reward, and the update does not differentiate through the evaluator. Those guarantees therefore do not establish the recovery threshold studied here.

For bounded fixed rewards with distinct means, Mei et al. (2024) permit any fixed positive step size. Under policy-dependent evaluation, constant steps can fail at the threshold even with two actions (Appendix F.3). The present results complement expected-regret bounds (Baudry et al., 2025) and analyses of stochasticity and information access (Mei et al., 2021).

G Simulation Protocol and Finite-Horizon Checks

The simulations directly instantiate the finite categorical model with update (U). Threshold trajectories use action-sampling randomness alone, while the sensitivity sweeps additionally include Gaussian comparison noise. Their purpose is to visualize finite-horizon behavior of the analytic model.

G.1 Three- and four-action trajectories at the threshold

For Figure 3, set $\Delta = L = \gamma = 1$ in the evaluator of Theorem 2, with shifted scores $(p, 0, -1, \ldots, -1)$. The same evaluator formula is admissible for $K = 3$, where Theorem 1(2) gives convergence. Initialize the actor at $(0.1, 0.7, 0.2)$ for $K = 3$ and $(0.1, 0.7, 0.1, 0.1)$ for $K = 4$, so the optimal, designated suboptimal, and total auxiliary masses agree. We run 128 independent trajectories for each case, with seeds 5623701 and 5623702, respectively, and

$$\eta_t = 0.002(1 + t/10^6)^{-0.6}, \qquad 0 \le t < 2 \times 10^6.$$

Continued indefinitely, this schedule has divergent sum and finite sum of squares. The finite horizon corresponds to cumulative step size $S_T \approx 2759$, and to 4×10^6 action draws per trajectory. Plotted lines are pointwise medians and bands are pointwise 10th–90th percentiles, rather than confidence intervals for convergence probabilities.

The median terminal optimal-action probability is 0.99972 for $K = 3$ and 0.00222 for $K = 4$. The initializations expose contrasting finite-horizon transients on an accessible time scale. Figure 3 reports terminal policy mass with an analytic failure-event reference; asymptotic success or failure is determined by the theorems rather than by thresholding finite trajectories.

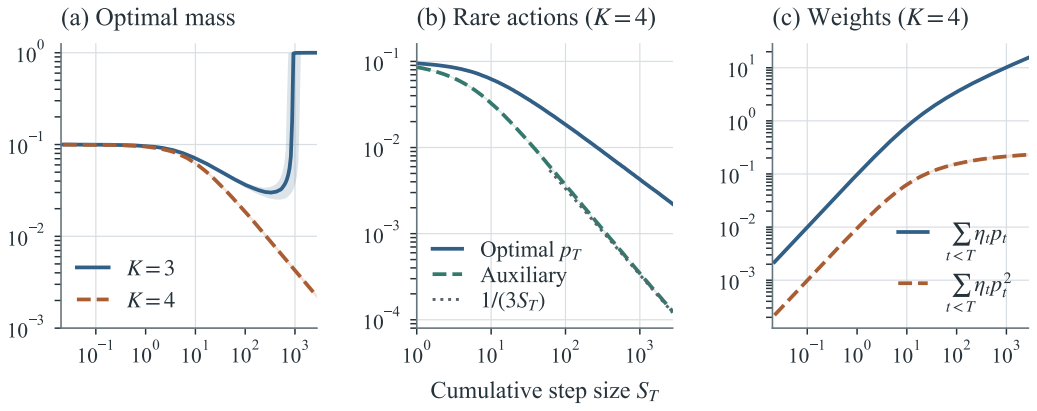

Figure 3: Sampled trajectories at $\Delta = L = \gamma = 1$: medians and 10th–90th percentiles across 128 runs. (a) Optimal-action mass for $K = 3, 4$. (b) Optimal and mean per-action auxiliary mass for $K = 4$; the dotted line is the asymptotic equivalent on the failure event. (c) Cumulative step-weighted sampling and correction terms for $K = 4$. The horizontal axis is cumulative step size S_T.

G.2 Exact simulation with duplicate rounds retained

Duplicate pairs leave the logits unchanged. At a fixed policy π, the probability of a distinct-action pair is

$$q_{\text{dist}} = 2 \sum_{i<j} \pi_i \pi_j.$$

We sample the number $W \in \{1, 2, \ldots\}$ of rounds until the next distinct-action pair from a geometric distribution with parameter q_{dist}. Conditional on that event, choose the unordered pair $\{i, j\}$ with probability $2\pi_i\pi_j/q_{\text{dist}}$. Reversing the order of the pair reverses both the contrast and the coordinate difference, so it leaves the update unchanged when evaluation noise is zero. With independent symmetric Gaussian contrast noise, reversing the orientation also preserves the update law. This procedure therefore samples the same policy process as two independent ordered draws at every round.

Each event retains the step size associated with its original sampling-round index. Because the policy is constant across skipped duplicate rounds, prefix sums of the deterministic schedule account for their contributions to S_T, $\sum_{t<T} \eta_t p_t$, and $\sum_{t<T} \eta_t p_t^2$.

G.3 Finite-Horizon Check of the Proof Event

We numerically evaluate the sufficient martingale event used in the analytic counterexample. At $T = 10^4$, the harmonic schedule gives only $S_T \approx 0.09788$, while the suboptimal-action probability already starts at 0.994159; the martingale event is therefore the informative finite-horizon quantity.

For this experiment we use direct ordered sampling in all rounds, the initialization in Appendix C.2, and 1,024 independent runs with seed 5623703. Recall $M_T = \sum_{t<T} \eta_t (g_t - f(\pi_t))$ and let

$$A_T = \left\{ \max_{0 \le t \le T} \|M_t\|_\infty \le \frac{1}{8} \right\}.$$

On A_T, a future martingale displacement of at most $1/8$ implies the event $\mathcal{E}_{1/4}$ in (25). The explicit constants in Appendix C then imply convergence to the suboptimal vertex. The analytic vector

martingale bound, with $V = 4$ and $\eta_t = 1/[100(t+1)]$, gives

$$\mathbb{P}\left(\sup_{u \geq T} \|M_u - M_T\|_\infty > \frac{1}{8}\right) \leq \frac{2V}{(1/8)^2} \sum_{t \geq T} \eta_t^2 \leq \frac{0.0512}{T}.$$

Thus $\mathbb{P}(\pi_t \to e_2) \geq \mathbb{P}(A_T) - 0.0512/T$, without any independence assumption between the prefix and tail events. All 1,024 simulated prefixes satisfy A_T, with largest observed coordinate fluctuation 0.007423, providing a finite-horizon check of the sufficient event. The > 0.97 convergence guarantee is the analytic result of Appendix C.

G.4 Finite-Horizon Illustration of the Zero-Gap Rate

Figure 4 illustrates the rate proved in Proposition 11. Use its binary evaluator with $\Delta = L = 1$, so the mean gap is $1 - p$. Start at $(1/2, 1/2)$ and use the same schedule and horizon as Figure 3, with 128 runs and seed 5623704. The median terminal normalized error $\sqrt{S_T}(1 - p_T)$ is about 0.51, near the proved limit $1/2$.

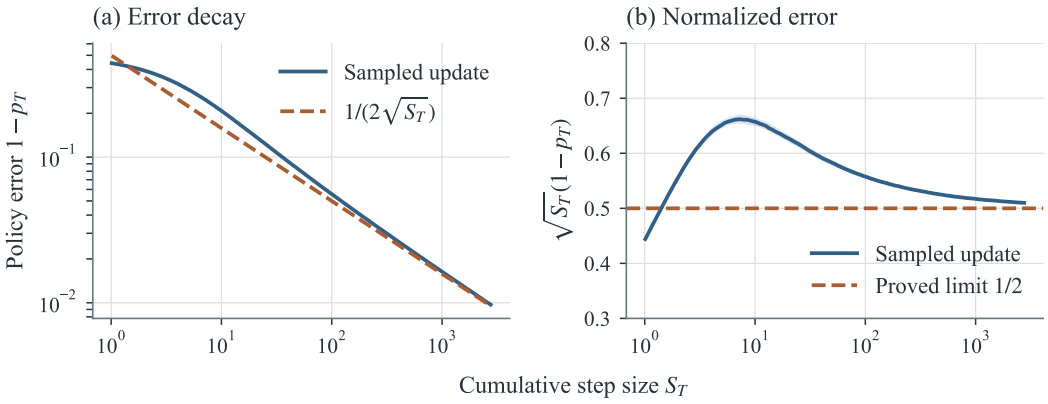

Figure 4: Finite-horizon illustration of the binary boundary rate in Proposition 11. Medians and 10th–90th percentiles across 128 runs show (a) policy error and (b) its $\sqrt{S_T}$ normalization. Dashed lines give the asymptotic equivalent and limit. The finite-time normalized error can overshoot its limit.

These rate experiments use zero evaluation noise, while the following sweeps add Gaussian noise. The heavy-tailed waiting-time phenomenon of Appendix F.4 is analytic and is not simulated here.

G.5 Within-Family Finite-Horizon Sensitivity Sweeps

All sweeps use the critical evaluator family of Theorem 2. To vary bias dispersion and sensitivity independently, take $r = (1, 0, -1, \ldots, -1)$ and, for a bias-amplitude parameter $D_{\text{bias}} > 0$ and sensitivity $L > 0$, set

$$h_{D_{\text{bias}}}(p) = (D_{\text{bias}} - Lp)_+, \qquad b_{i^\star} = -(1-p)h_{D_{\text{bias}}}(p), \qquad b_i = p h_{D_{\text{bias}}}(p) \quad (i \neq i^\star).$$

Every nonzero pairwise bias contrast is $\pm h_{D_{\text{bias}}}$, so its exact TV-Lipschitz constant is L. Also

$$\sigma_b(\pi) = \sqrt{p(1-p)}(D_{\text{bias}} - Lp)_+, \qquad \sup_{\pi \in \mathcal{P}_K} \sigma_b(\pi) = \Psi_L(D_{\text{bias}}).$$

The shifted scores are $(1 - (D_{\text{bias}} - Lp)_+, 0, -1, \ldots, -1)$. For each target ratio $\varepsilon/\Psi_L(1)$, solve the strictly monotone equation $\Psi_L(D_{\text{bias}}) = \varepsilon$ numerically using (16). Thus the horizontal axis records the evaluator's exact bias-dispersion ratio, with L fixed within each curve. We vary the bias-amplitude parameter rather than rescaling the entire bias, which preserves L within a sweep.

We use $\varepsilon/\Psi_L(1) \in \{0.8, 0.9, 0.95, 1, 1.05, 1.1, 1.2\}$ and three sweeps: $L \in \{0.25, 1, 4\}$ with $K = 4$ and $\sigma = 0$; $K \in \{3, 4, 6\}$ with $L = 1$ and $\sigma = 0$; and $\sigma \in \{0, 0.5, 2\}$ with $K = 4$ and $L = 1$. Each of the 49 distinct cells has 64 independent runs with 2×10^6 sampling rounds and the schedule above, so $S_T \approx 2759$. All start with optimal-action mass 0.1, suboptimal-action mass 0.7, and total auxiliary mass 0.2, divided equally over the $K - 2$ auxiliary actions. On rounds with distinct sampled actions, add independent $N(0, \sigma^2)$ noise to the real-valued contrast.

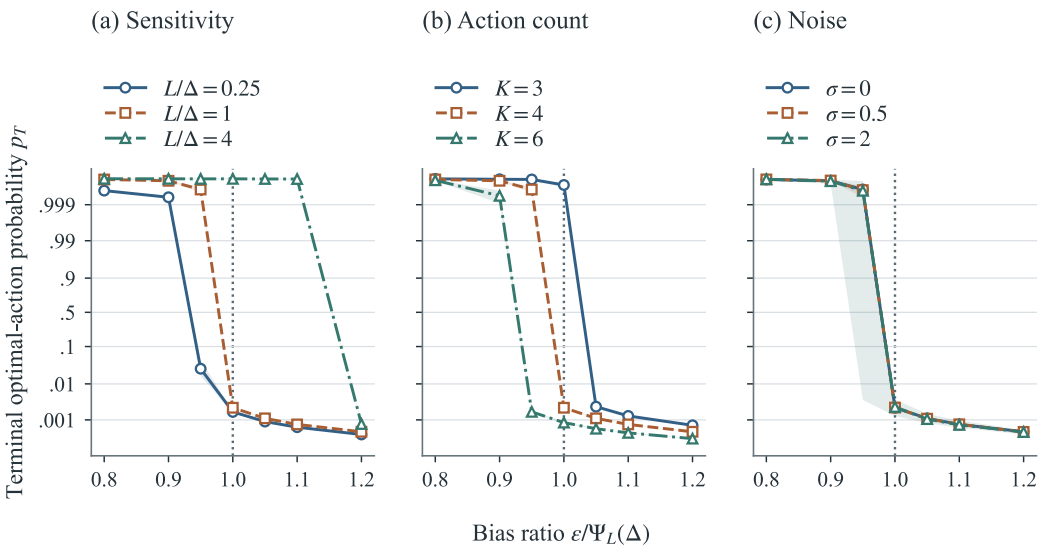

Figure 5: Within-family finite-horizon sensitivity sweeps: medians and 10th–90th percentiles across 64 runs per setting, varying (a) sensitivity, (b) action count, or (c) comparison noise from $K = 4$, $L = \Delta = 1$, $\sigma = 0$. Dotted vertical lines mark the asymptotic worst-case threshold; the horizontal axis uses the evaluator's exact bias-dispersion ratio. The probability axis has a logit scale. Connecting lines guide the eye; terminal mass reports finite-horizon progress.

Figure 5 also shows substantial variation in finite-horizon transients. At $L = 1$ and $\varepsilon/\Psi_L(1) = 1$, the median terminal optimal-action probabilities are 0.99972, 0.00218, and 0.00084 for $K = 3, 4, 6$. Finite-time behavior need not mirror the asymptotic classification: at $\varepsilon/\Psi_L(1) = 0.95$, the $K = 6$ median remains 0.00168 despite eventual recovery, whereas for $K = 4$, $L = 4$, and $\varepsilon/\Psi_L(1) = 1.1$, the median is 0.99981 although failure is possible above the boundary. Gaussian noise preserves this qualitative pattern. Terminal mass therefore serves as a finite-horizon progress measure within this evaluator family.

G.6 Reproduction

Code for reproducing all simulations and figures is included in the supplementary material. The accompanying README records the software environment, fixed seeds, and execution details; no external dataset or online service is required.